2015

Cyber Blackbox for collecting network evidence

Jooyoung Lee
Electronics and Telecommunications Research Institute, Chungnam University, Korea

Sunoh Choi
Electronics and Telecommunications Research Institute, Chungnam University, Korea

Yangseo Choi
Electronics and Telecommunications Research Institute, Chungnam University, Korea

Jonghyun Kim
Electronics and Telecommunications Research Institute, Chungnam University, Korea

Ikkyun Kim
Electronics and Telecommunications Research Institute, Chungnam University, Korea

See next page for additional authors

DOI: 10.4225/75/57b40079fb893
13th Australian Digital Forensics Conference, held from the 30 November – 2 December, 2015 (pp. 114-147), Edith Cowan University Joondalup Campus, Perth, Western Australia.

A CYBER BLACKBOX FOR COLLECTING NETWORK EVIDENCE

Jooyoung Lee, Sunoh Choi, Yangseo Choi, Jonghyun Kim, Ikkyun Kim and Youngseok Lee
Electronics and Telecommunications Research Institute, Chungnam University, Korea
{joolee, suno, yschoi92, jhk, ikkim21}@etri.re.kr, lee@cnu.ac.kr*

Abstract

In recent years, the hottest topics in the security field are related to the advanced and persistent attacks. As an approach to solve this problem, we propose a cyber blackbox which collects and preserves network traffic on a virtual volume based WORM device, called EvidenceLock to ensure data integrity for security and forensic analysis. As a strategy to retain traffic for long enough periods, we introduce a deduplication method. Also this paper includes a study on the network evidence which is collected and preserved for analyzing the cause of cyber incident. Then, a method is proposed to suggest a starting point for incident analysis to a forensic practitioner who has to investigate on the vast amount of network traffic collected using the cyber blackbox. Experimental results show this approach is effectively able to reduce the amount of data to search by dividing doubtful flows from normal traffic. Finally, we discuss the results with the forensically meaningful point of view and present further works.

Keywords
Cyber incident forensics; Digital forensics; Cyber attacks; Network forensics, Incident investigations

INTRODUCTION

In recent years, the hottest topics in the security field are related to the advanced and persistent attacks. Attackers have been threatening public institutes and social infrastructures as well as giving deep anxiety social members by revealing their secret information or hindering their activities. However, it is not easy to prevent these threats with existing security devices on the market although a lot of security venders have been trying to address these issues by providing intelligent security systems. Furthermore, it could say that it is rarely possible to perfectly defend them. So that detecting and analysing these threatening activities has become one of the main challenges of the security researches.

Digital forensics, meanwhile, has become important for the role as a forward lean security solution. It usually means a process to find legal evidence from computers and digital storage media. Nowadays its scope has continuously expanded to refer to investigating cyber incidents or crimes. A general approach for obtaining evidence for network forensics after a cyber-attack occurred is to gather the network logs remained in the systems because network traffic is volatile. In this case, however, it is difficult to have enough data to be required for finding the cause of the attack because an attacker usually deletes relevant data or destroys the systems when he has accomplished his task.

In order to solve this problem, we take another approach getting the network traffic before a cyber-incident happens using a cyber blackbox that we have been developing. It has a similar concept and goal with a black box adhere to a car or with a CCTV device in front of an entrance of a building. Additionally, in order to ensure integrity of collected network data for security and forensic analysis, the cyber blackbox preserves collected traffic on a virtual volume based WORM device, called EvidenceLock. In this paper, we describe a cyber blackbox system and study on the traffic data gathered from the cyber blackbox. Then a method is proposed to suggest a starting point for incident analysis to a forensic practitioner who has to investigate on the vast amount of network traffic collected using the cyber blackbox. The experimental results by the method are discussed with the forensically meaningful point of view. Finally, we present conclusion and further works.

DESIGN OF CYBER BLACKBOX

A cyber blackbox is a system that collects and preserves network traffic at the front of a network entry point or of important IT facilities in an organization. Its main goal is to retain enough network traffic given to a forensic investigator so that the investigator finds evidence and understands the cause of cyber incident. Now that network data give rich information on the activities through network, by analysing them, the investigator is able to obtain not only network connection information among parties, but also contents transmitted through network communications. Figure 1 shows components of cyber blackbox and data flows between them.

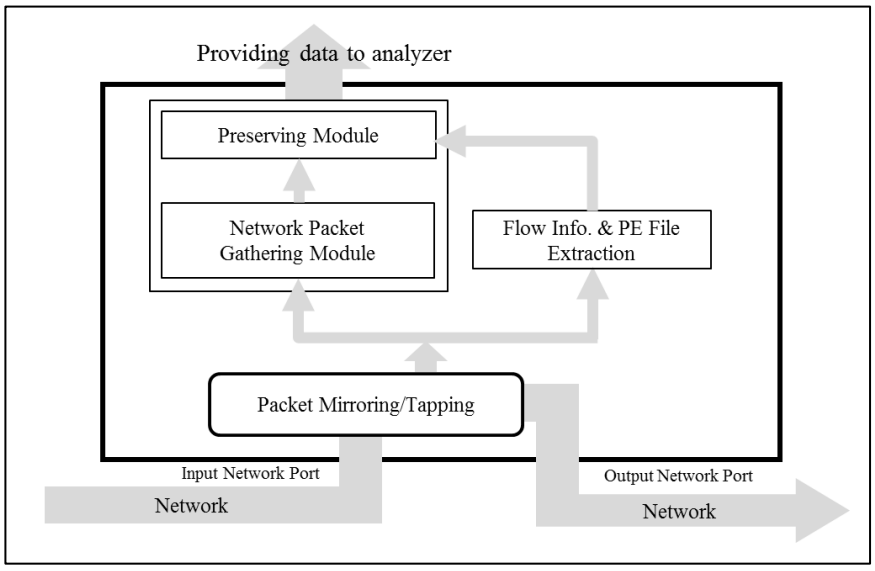


Figure 1. Components of cyber blackbox and the data flows between them

Cyber Blackbox Gathering Module

Our cyber blackbox includes a gathering module and a preserving module. The gathering module taps network traffic and stacks collected packets during a period of time in a buffer. After time unit, buffered packets are combined to make a file in a PCAP file compatible format.

One of main issues in cyber blackbox is related to data volume. Theoretically, the gathering module could collect the amount of 2 TBs traffic from 1 Gbps link with an about 40% network utilization rate for 12 hours. According to (Woo et al, 2013) which had studied on network traffic at a core network switch on GGSN serving as a gateway to IP networks, logging all IP traffic at a 10 Gbps link during a week amounts to 370 TBs in volume. A large amount of data volume causes a big data problem and it is very difficult to find evidence in vast data.

In order to solve the problem, we conduct a de-duplication of traffic by chunking and fingerprinting on a packet basis. Applying de-duplication technique in the gathering process gives us two benefits by making a de-duped file smaller than an original one in size. One is that we can reduce I/O time to take for storing a large file and the other is able to save the storage required to retain a lot of traffic.

Cyber Blackbox Preserving Module

The preserving module gets to save a collected file into WORM(Write Once Read Manage) storage device in which information, once written, cannot be modified. This write protection affords the assurance that the data cannot be tampered with once it is written to the device.

There are many types of media supporting WORM functionality, such as CD-ROM, Magnetic Tape and so on. But it is not good solutions for our system because they are too slow to write network traffic to be collected at line rate into them. In order to provide suitable device instead of them, we have developed EvidenceLock to provide virtual volume based WORM storage.

EvidenceLock creates a virtual volume to be mapped to an image file which consists of actual files to be stored into a virtual volume. It is also designed to manage a storage usage rate by automatically deleting aged virtual volumes in chronological order when the storage is almost full of data. Figure 2 shows the metadata file structure for an image and an image file structure. EvidenceLock stores files into an image file in sequence arrived. It doesn't allow users to modify any data in a file to ensure that data in this storage are not damaged and keep their integrity. In addition, during preserving process, it calculates a hash value of the file with a cryptographic algorithm to prove integrity of data.

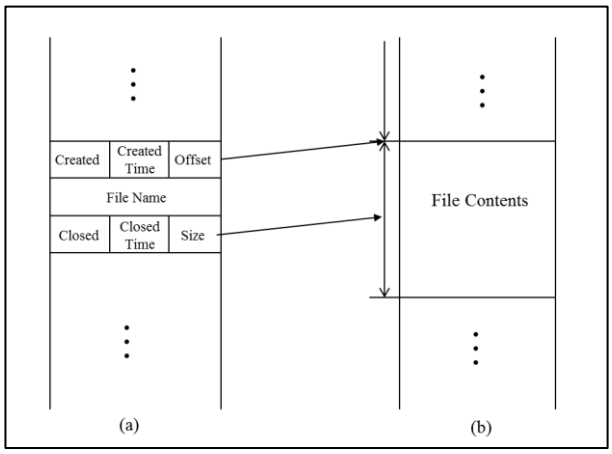


Figure 2. Structures of (a) a metadata file for an image and (b) the image file for data used in EvidenceLock

CYBER BLACKBOX DATA

In this section, we describe data collected using our cyber blackbox. The gathering module collects network traffic and produces some kinds of cyber blackbox data using the traffic. The cyber blackbox data consists of preservation data and management data as depicted in Figure 3. We define the preservation data as what are used for forensic analysis and evidence and it needs to freeze not to be modified. Therefore, we save it into the EvidenceLock to guarantee that data can't be tampered with when it is written to the device.

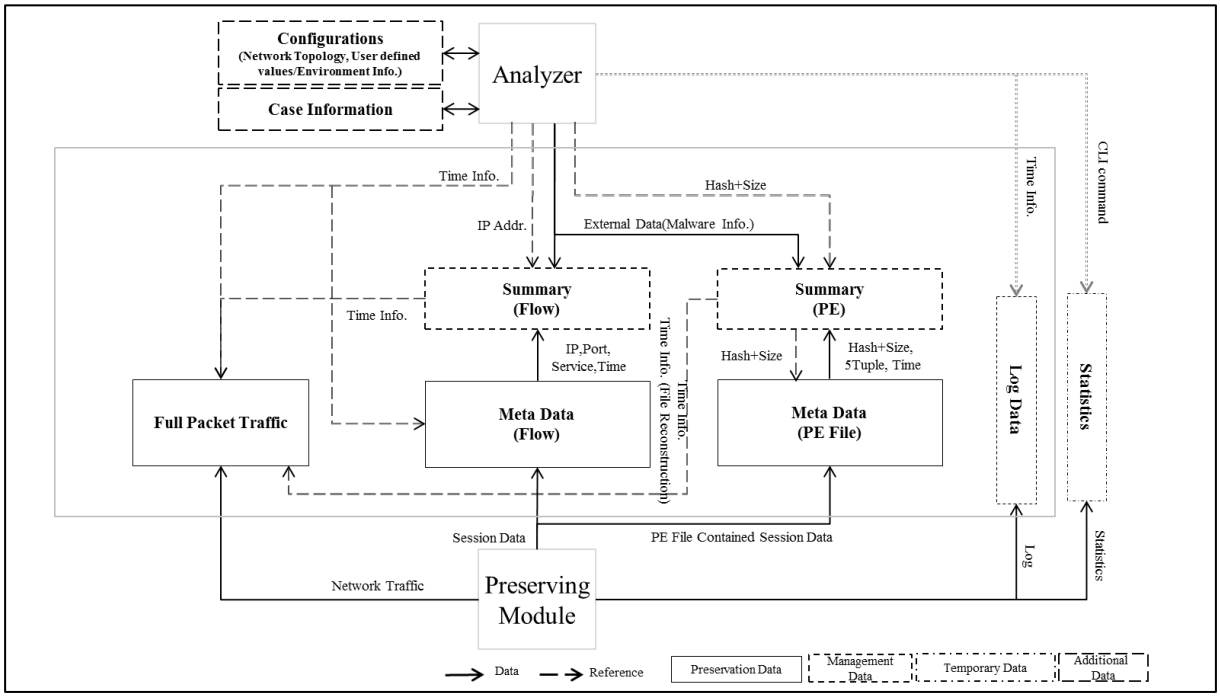


Figure 3. Data types to be collected in cyber blackbox and their relation

Preservation data includes full packets and their metadata. Capturing full packets means to have all data through the cyber blackbox without sampling. Investigating tremendous full packets is important but difficult because it causes a big data problem. So, in order to give helpful information to a forensic practitioner, we additionally extract metadata from the full traffic.

The metadata encompasses a wide range of information including flow information and file information. The flow information provides source IP address, destination IP address, source port number, destination port number, time, service, amounts of transferred data in byte, the number of packets, kinds of application, and so on. Also it reconstructs files in various formats using packets transferred on the network and then produces their additional information comprising a file hash value, time to be sent, file size in byte, and file name. Besides, we collect management data including statistical data, summary data, log and so on.

Meanwhile, what kinds of information we can obtain from the network depends on where the cyber blackbox is deployed. Figure 4 shows an example of typical positions the cyber blackbox can be located. We could consider three positions to place it. The first is an entry point of network traffic. By placing it in the position, an investigator is able to show data flows from internal to external network of an organization and vice versa.

Second position is in front of a mission critical system, such as a patch management system, financial database system, asset management system and so on. Acquiring data at this point helps the investigator analyze the cause of cyber incident happened in the important system. The collected data includes which IP connected the system and what contents delivered from and to the system. The last considerable place is a subnet switch where the cyber blackbox is able to tap the data transferred among systems in the subnet. The investigator can analyze lateral movement of malwares or information in the internal network using this data.

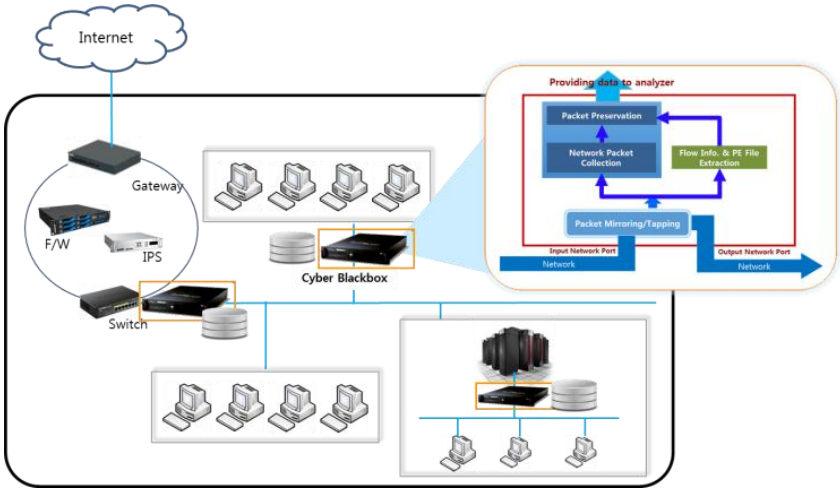


Figure 4. An example of possible places where cyber blackbox system is deployed

STUDY ON THE COLLECTED DATA AND THEIR FEATURES

In this section of the paper reports a case study on data collection through our test-bed network for about a month using our cyber blackbox. We have configured that collected traffic is saved once every minute and the produced flow information is stored in a SQLite database file once every ten minutes

The cyber blackbox have tapped traffic on a switch system of test-bed. On average, it collected 1.75GB network traffic in volume and 1,187,852 flows each day. Although this amount of traffic is not large volume since our test-bed is not an environment used in daily work, we found that 12.14GB traffic happened on a specific day and it means a small scale of DDoS attack occurred.

Generally, it is difficult to investigate a large volume of data rapidly and accurately and sometimes, a forensic practitioner doesn't know what he examines at first. Therefore it is important decreasing the scope and volume of data that he has to focus on. To address this problem, we have applied a statistical analysis method to 129,000 flows collected for a specific period to identify unusual data flows.

We have employed principal component analysis (PCA) which uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. There are some studies applying PCA to network traffic data.

A research in (Lakhina el al, 2004) has proposed a method to determine anomalies using PCA which separates high dimensional space into disjoint subspaces, which corresponds to normal and anomalous network conditions. This research has worked with the sampled link traffic data, of the kind obtained by SNMP collected from two backbone networks. They have applied PCA on their link data matrix that represents a set of links measurement over time. From the examining the amount of variance captured by each principal component, they have observed that the vast majority of the variance in each link time-series can be well captured by 3 or 4 principal components and it allows to treat diagnosis of network-wide traffic anomalies as a spatial problem.

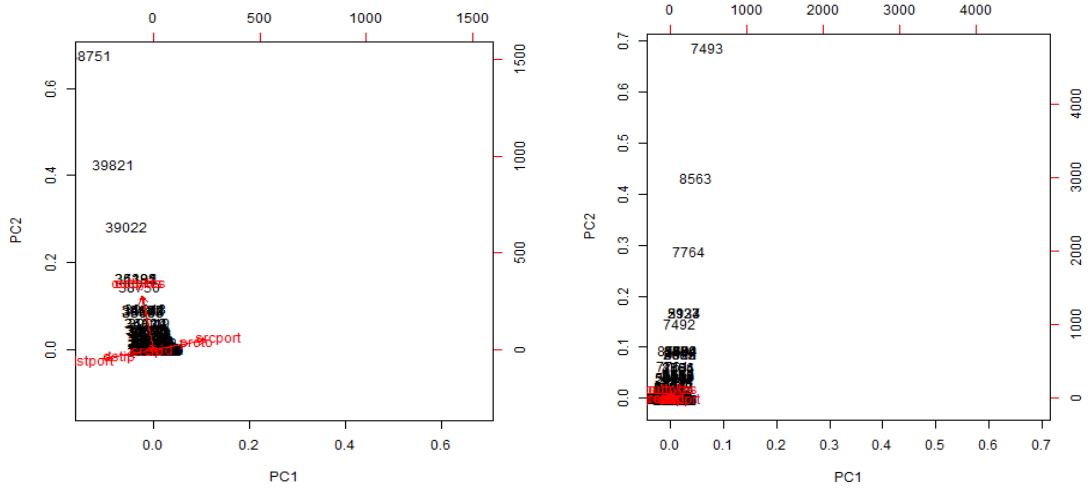
Another study in (Kim el al, 2008) has proposed an approach to select useful network elements which are able to efficiently model network data. In order to do, they have quantified packets captured from a volume of network traffic and have used a statistical analysis method to identify the most effective elements for efficiently classifying the modelled data.

Table 1 shows flow information to be used for constructing a matrix. We have used 8 features which are source IP address, destination IP address, source port address, destination port address, network protocol, service id, number of packets, and number of bytes. We have used a PCA module provided by the R package which is a free software environment for statistical computing and graphics.

Table 1. Flow information used in PCA

Features	Description
SrcIP	IP address of a party sending packets
DstIP	IP address of a party receiving packets
SrcPort	Port number of a party sending packets
DstPort	Port number of a party receiving packets
Protocol	Communication protocol to transmit data
ServiceID	Applications in OSI layer 7
Packets	Number of packets to transport in a flow
Bytes	Number of bytes to transport in a flow

The results by the PCA show that it is possible to describe 80~85% of total amount data with four principal components. Considering eight features used, we can say it is hard to find major principal components to represent traffic data clearly. However, when a result by the single value decomposition has been projected on 2-dimensional space, we have obtained outcome as shown in Figure 5.



(a) *(b)*

(c)

Figure 5. Two dimensional projection results by single value decomposition analysis on the flows collected for a period of time.

The figure has three graphs presenting transformed flows by PCA. The number depicted in each graph is a Flow ID representing each flow data. On the graph, it is plotted on a position which is determined by its first and second principal component values, PC1 and PC2, respectively. As seen in all graphs, most of the flows are clustered in a part of position, whereas a few of them are separated from the cluster. It means that the isolated flows have conspicuous component values comparing to the clustered flows. In other words, isolated flows have unusual information comparing with the other flows.

For example, the graph (a) which is depicted with a PCA result of 129,361 flows shows that only three flows, ID 8751, 39821, 39022 are apart from others. From the data collection, we have found that 183,910 packets to be sent through a flow having ID 39821. In our data, it is an unusual case comparing to that most flows usually transport a few of packets in our test-bed. As another case, the graph (b) shows 16,721 flows used for the DDoS attack happened in our test. The flows are collected for 20 minutes and flow ID 7493, 8563, and 7764 are top three flows to send a large amount of traffic. The graph (c) presents the flows which contain the injected malware files. The flows are approximatively divided into two groups in this graph.

From the results, we have found that it is possible to distinguish a few of abnormal flows from many other flows by applying PCA to the flow data. It could be helpful for a network forensic practitioner to distinguish suspicious flows from a large amount of traffic when he conducts investigation without preliminary information for it.

RELATED WORKS

Applications in some domain, such as network trouble shooting, security analysis, performance debugging and so on need to hold network traffic for a long time. Therefore, researches have been done to capture packets to disk at line rate without expensive dedicated network adapters.

In the paper (Fusco et al, 2010; Deri et al, 2013), Luca Deri et al proposed the design and implementation of n2disk, which dumps 10 Gbit traffic to disk using commodity hardware. In order to achieve it, they have implemented PR_RING introducing the concept of virtual capture device. It provides a zero-copy mechanism and it moves each incoming packet in a temporary memory called a socket buffer in kernel space to a PF-RING ring buffer which is memory mapped to the user space. Also, they have exploited the parallelism of multi-core architecture to overcome the performance limitations and have shown that it is possible to capture packets and store them to disk efficiently at 10 Gbit.

FloSIS (Lee et al, 2015) is another network packet capture system proposed by J. Lee et al. It supports flow level indexing for responding query quickly. As like n2disk, it exploits multi-cores and disks with full parallelism and minimizes expensive disk access for querying by writing the packets of same flow at a contiguous disk location. In this paper, they have presented that the zero drop performance at 30 Gbps line rate has been achieved.

Although these papers have presented good approaches for collecting entire traffic without loss at high speed rate network, there is more requirements used for forensic investigation. Firstly, integrity of collected data is guaranteed. It means that not only data must be gathered without loss, but also it should not be modified after stored. Secondly, it is required to provide some mechanism to efficiently analyse collected data since it is a kind of big data so that it is difficult to find something like evidence in case of that a user doesn't know what he really wants to find. Finally, a lot of investigators are not expert in network analysis so that they need useful analysis tools to help them.

CONCLUSION AND FURTHER WORKS

Analysing cyber blackbox data gives an investigator benefits in a point at which he can use a lot of data obtained during a long period of time. Although it seems like to be as not different as existing researches on the network data analysis in some context, it could be a very important and effective approach because recent cyber-crimes tent to persistently be carried out for a long time.

However, it has not been enough tools to investigate a volume of network data. Existing network protocol analysers are having difficulty to handle data at big scale. We have been research on the issue and developed a tool to support a user who investigates cyber incident with enormous amount of network data.

This paper has described the study on the traffic data gathered from the cyber blackbox. Especially, we show that the distinctive flows could be effectively divided from a large amount of data using principal component analysis. Using this result, it is expected to improve our further research to analyse cause of cyber incident and to provide a useful network forensic tool.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (B0101-15-0300, The Development of Cyber Blackbox and Integrated Security Analysis Technology for Proactive and Reactive Cyber Incident Response)

REFERENCES

Write Once Read Manay(WORM), available online at https://en.wikipedia.org/wiki/Write_once_read_many

Woo, S., Jeong, E., Park, S., Lee, J., Ihm, S., and Park, K. (2013), "Comparison of caching strategies in modern cellular backhaul networks," MobiSys'13, June 25-28, 2013, Taipei, Taiwan

Principal Component Aanalysis(PCA), available online at https://en.wikipedia.org/wiki/Principal_component_analysis

Lakhina, A., Corvella, M., and Diot, C. (2004), "Diagnosing Network-Wide Traffic Anomalies," SIGCOMM'04, Aug. 30–Sept. 3, 2004, Portland, Oregon, USA.

Kim, H., Cho, J., Lee, I., and Moon, J. (2008), "Effective Feature Selection Model for Network Data Modeling," Journal of Broadcasting Engineering, Vol 13, No 1, pp.92-98, 2008 (In Korean)

Fusco, F. and Deri, L. (2010), High Speed Network Traffic Analysis with Commodity Multi-Core Systems, Proc. of IMC 2010, 2010

Deri, L., Cardigliano, A., and Fusco, F. (2013), "10 Gbit Line Rate Packet-to-Disk Using n2disk," The 5th IEEE International Traffic Monitoring and Analysis Workshop (TMA), pp 441-446, 2013

Lee, J., Lee, S., Lee, J., Yi, Y., and Park, K. (2015), "FloSIS: A Highly Scalable Network Flow Capture System for Fast Retrieval and Storage Efficiency," the Proceedings of the 2015 USENIX Annual Technical Conference (USENIC ATC '15) July 8–10, 2015 • Santa Clara, CA, USA