Visual Investigations of Botnet Command and Control Behavior

Lancope

Tom Cross and Andrea Fletcher
Lancope, Atlanta, GA

SLIC™

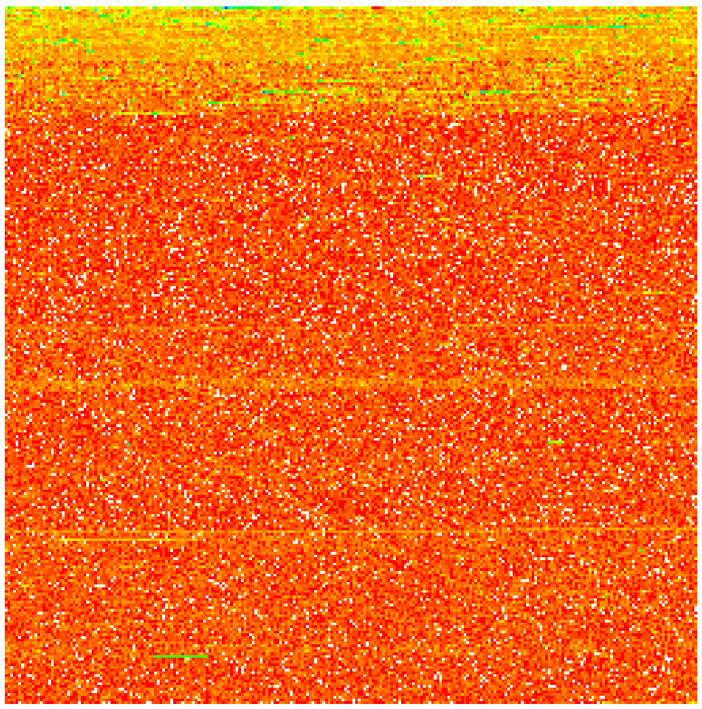
Abstract

Malicious botnets are a problem that continues to plague the Internet. Every year, attackers infect millions of computer systems with botnet software that is designed to steal information or launch other attacks. Attackers control these networks of infected computer systems using command and control channels that operate over the Internet Protocol. Once botnet software has infected a computer system, it reaches out to a predesignated command and control system for further instructions, typically on a predetermined TCP or UDP port.

Malware data available to Lancope suggests that 85% to 95% of malware samples use TCP port 80 to communicate with command and control servers. The alternate ports chosen by the remaining samples are worth investigating to determine if there are patterns of port selection behavior that can be useful for detection. Our research explores the heuristically identified command and control behaviors of a collection of nearly two million unique botnet malware samples that were active between 2010 and 2012. These samples reached out to nearly 150,000 different command and control servers on over 100,000 different TCP and UDP ports. This data set is complex and heterogeneous, and thus it is difficult to analyze. However, when the data is represented visually, patterns emerge that lead to interesting insights.

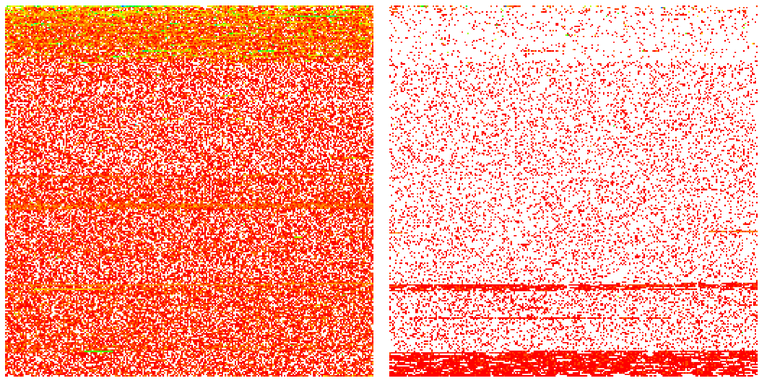
We created color heat maps representing TCP and UDP port popularity on 255x257 pixel charts. Each pixel represents a single port number, starting from port 0, and the hue of each pixel represents the number of command and control hosts in our sample set utilizing that port (on a log scale). These charts are compared with similar charts representing port utilization in a small office computer network over the course of a single month. Several interesting features are identified, associated with particular malware campaigns, and observations are made about distinctions between the malware data and the control set.

Combined Malware Command and Control TCP and UDP Port Utilization



Our heat map of combined TCP and UDP port utilization is presented along with a hue scale for reference. Unused ports are white. Ports with low utilization are on the left of the hue scale moving right as utilization increases. The data representation is scaled so that hues from the violet part of the spectrum are not reached.

TCP Port Utilization – Malware Command and Control vs. Small Office LAN



Port Utilization

Ranges of TCP and UDP ports are used by different software applications for different purposes. IETF RFC 6335 explains the different port ranges used in the Internet. Ports between 0 and 1023 are known as “well known ports” and are assigned for use by particular applications. On many multiuser computer systems only the super-user has access to bind an application to listen for connections on these ports. Ports between 1024 and 49,151 are known as “registered ports” that are available for user applications to bind to. Ports greater than 49,151 are known as dynamic or ephemeral ports that are available for random assignment, however, in practice a lot of Internet software utilizes the entire port range above 1023 for dynamic assignment. Internet standards also indicate that TCP and UDP port ranges are supposed to be utilized in similar ways, but many differences exist in practice, as our control data illustrates.

TCP Ports Used by Malware

To a certain extent, malware port utilization reflects, in aggregate, the personal preferences of malware authors. Our data shows that TCP ports below 10,000 are particularly popular. 866 of the 1024 “well known ports” below 1024 were used by malware in our dataset. There are clearly visible bands of popular ports near port 30,000, between port 35,000 and 36,000, near port 49,000, and near port 60,000.

A band of popular ports exists between port 61,000 and port 61,020. These ports are used by a family of over one thousand password stealing trojans controlled by two servers on the same network in Russia. The malware samples refer to these two servers by more than 300 different domain names.

Another popular band exists between port 41,000 and 41,005. These ports are popular with a variety of malware samples, but most notably a collection of over 18,000 samples that all communicate with a single command and control server in France.

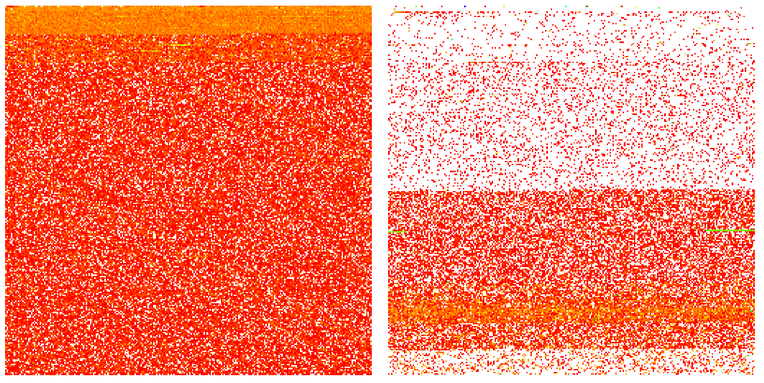
TCP Ports Used in a Small Office LAN

For comparison we created a control data set by monitoring the ports of Internet hosts that were accessed from a small office computer network over the course of a month of normal business activity. In contrast to the malware samples, there is a far lower density of port utilization below port 10,000 in the control set. Only 166 of the “well known ports” below 1024 were used. There is a stark increase in density above port 49,151, corresponding with the ephemeral port range. Most ports above port 61,200 were used.

UDP Ports Used by Malware

Similar to the TCP Malware ports, UDP ports below 10,000 are popular, with a particular emphasis on ports below 5,000. Almost all of the “well known ports” below 1024 are used by malware in the collection (1018). A band of popular ports used by thousands of unrelated malware samples begins at port 7000 and another at port 8000.

UDP Port Utilization – Malware Command and Control vs. Small Office LAN



The most striking feature of the UDP malware port image is the set of three diagonal lines of popular ports that stretch through the image. These lines start at port 0, port 36, and port 45, and in all three cases represent sequences of every 257th port from the starting point. We isolated the exclusive use of UDP ports fitting this sequence down to 14 specific malware samples. Due to the unique nature of the pattern of port utilization by these samples, it seems likely that they are all related to each other, in spite of the fact that they communicate with 6 different domain names that have been hosted in 8 different countries, all over the world. It is possible that the same botnet operator is responsible for propagating all of these samples.

UDP Ports Used in a Small Office LAN

Similar to the TCP control set, the UDP control set has a low density below port 10,000. Only 19 of the “well known ports” below 1023 were used. The density increases significantly above the half way point (32,768) and particularly in the ephemeral port range between 49,152 and about 57,000. UDP port utilization drops off above port 61,000, which is different from the behavior observed in the TCP control set.

Conclusions

Our heat map visualizations made it easy to observe malware port utilization in aggregate and identify ranges of popular ports as well as specific regions of interest. Comparison with similar visualizations of legitimate traffic are also easy to make using these images. The port heat maps significantly reduced the difficulty associated with extracting meaningful observations from a complex dataset.

Malware authors seem to prefer to use low port numbers, whereas legitimate software often uses higher ports. The difference is particularly clear for “well known ports” below 1024. Our malware samples used 866 “well known” TCP ports, but the legitimate traffic only used 166. On the UDP side, 1018 “well know ports” were used by malware, but only 19 were used on the legitimate network. This suggests that use of unusual ports below 1024 is a behavioral anomaly that could be indicative of a malware infection.

A similar observation can be made about the use of ephemeral ports. TCP and UDP ports above 49,151 are supposed to be dynamically assigned for use by legitimate software applications. This would suggest that they are used transiently. However, many of these ports were used for command and control communications by malware in our sample set. Command and control communications tend to involve consistent communication over the same port. Consistent use of a port above 49,151 is another indicator that could be indicative of a malware infection.

Further investigation of consistent versus transient use of ports by legitimate software applications may provide larger port ranges for which consistent communications can be considered suspicious behavior.

Lancope would like to thank the Georgia Tech Information Security Center for their assistance with this research.