The Contact Surface: A Technique for Exploring Internet Scale Emergent Behaviors

Carrie Gates[1] and John McHugh[2]

[1] CA Labs
Islandia, NY, USA
carrie.gates@ca.com
[2] Dalhousie University
Halifax, NS, Canada
mchugh@cs.dal.ca

Abstract. Large scale internet data analysis often concentrates on statistical measures for volume properties or is focused on the epidemiology of specific malcodes. We have developed a high level abstraction that we call the contact surface that allows us to visualize internet scale connection behaviours across the border of a monitored network. The contact surface is a time series of contact lines, each line plotting the number of outside sources that contact a specific number of inside hosts in a given time interval (typically an hour). In general, the lines follow a power law in the mid range with distinct outliers at the one destination per source and the hundreds to thousands of destinations per source ends. During some periods, however, the lines are perturbed with what appears to be a persistent bump or waterfall. We have studied two such episodes, one that persisted from at least January 2003 until August 2003 and another that appeared on February 11, 2004 and lasted until May 31, 2004. The exact cause of the former is unknown, however the later appears to have been caused by the Welchia.B worm. Similar activities are currently being reported by other observers. We hypothesize that the cause of the perturbation is low frequency periodic scanning by a small population of hosts scanning at the same rate. We have created simulations to explore the range of activities that might be observable and find reasonable agreement with the observed phenomena.

1 Introduction

In 2003 and 2004, we had access to NetFlow data from the border of a composite network (multiple, disjoint network blocks) that covered address space that was equivalent to multiple /8s in aggregate. The network is sparsely populated but contains several million active hosts. As part of the preliminary efforts to investigate schemes for detecting coordinated distributed scans, we became interested in "typical" host to host connection behavior. Many of the papers that examine the characteristics of network traffic focus on packet properties such

D. Zamboni (Ed.): DIMVA 2008, LNCS 5137, pp. 228–246, 2008.

as counts of hosts, protocols, port usage, payload size and characterization, etc. Other papers, inspired by work in social networks attempt to map communications patterns on the physical structure of the internet, overlooking the fact that from the standpoint of an IP layer user, the internet is flat and fully connected.

Our fundamental question was essentially "Is there any regularity in cross border connection behaviour?" We were interested in determining how many of the outside hosts generating traffic into the network connect to (or attempt connections to) one service, two services, three services, etc. In the beginning, we were concerned only with the quantitative nature of the connecting populations and not with the identity of the participants. In this case, the initial investigation was restricted to TCP and a service was defined as the combination of a host IP address and a service port.

One of the tools that we used for studying this traffic was a visual representation that we call a *contact surface*. This is a three dimensional time series of lines in which each line shows the number of outside sources, Y vertical, that contact a specific number of internal addresses (or address-service combinations[1]) on the inside of the network, X horizontal. Time is into the page, Z. Because of the large dynamic range ov values represented, we present the contact surface as a log-log plot. The lines represent hourly flows and the shading separates days of the week.

The first period for which we developed a contact surface was a week of data from January, 2003. This data manifested a "bump" or standing wave (perturbation) in the surface as seen in Figure 1. This phenomenon was observed from January 2003, the earliest data available to us, until mid August 2003 when it abruptly disappeared. A similar phenomena appeared in mid February 2004 and persisted through the end of May, 2004. We have been told that similar phenomena are present in recent data, but we lack current access to the data source.

To a first approximation in the absence of a disturbance, the data for each hour can be represented as a straight line of the form $\log(y) = A + B\log(x)$ and we plot the x and y values on logarithmic scales. Figure 2a shows an example of the undisturbed contact surface while Figure 2b shows the regression line superimposed on several aggregated lines. These are discussed in more detail is Section 2.

This paper discusses our present views of this phenomenon. Section 2 discusses the initial observations and analysis with separate discussions of the 2003 and 2004 perturbations. We hypothesize that this phenomenon was related to the appearance of scanning worms that exhibited particular timing characteristics and consider the minimum amount of address space that must be monitored in order to observe this phenomenon in Section 3. Recent observations of similar perturbations resurrected our interest in the wave feature and in the visualizations.

[1] An address-service combination is a unique combination of an IP address and a service. For TCP and UDP a service is the combination of the protocol and the destination port. For other protocols, the notion may vary with the protocol, but these are sufficiently rare so we can assume a single service per protocol.

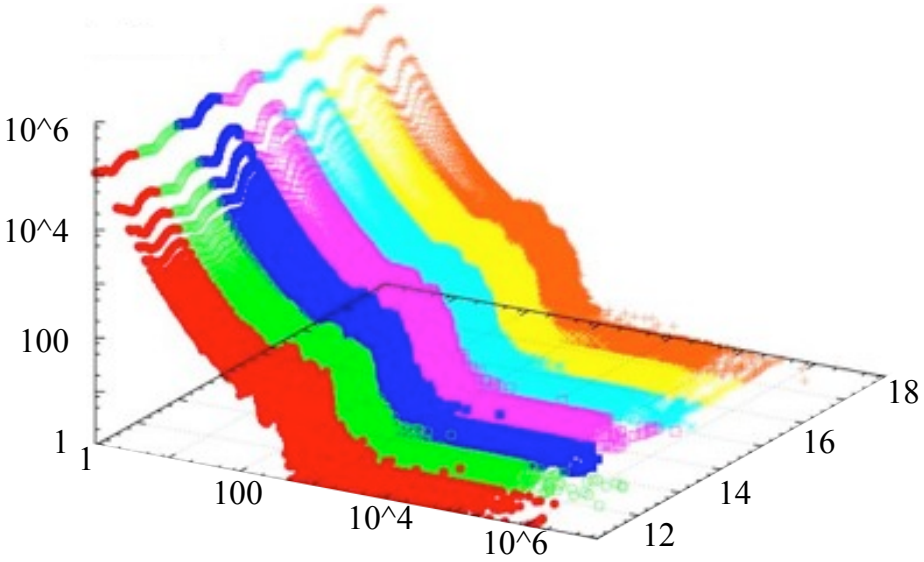

Fig. 1. Contact surface for January 11-18, 2003. Vertical axis is count of outside sources. Horizontal axis is count of inside services targeted by each outside source. Time is into the page.

Although we do not have access currently to either the historical or current data from this source, we have developed a simulation that provides a plausible explanation for the observed phenomena. Described in Section 4, this allows us to vary the background and perturbing parameters to examine the effects of altering the perturbing population, its probe rate, the size of the monitored network and the percentage of the probes that are observed. Related work is presented in Section 5 and our conclusions and future plans in Section 6.

2 Observed Phenomenon

NetFlow traffic from the border routers of a large ISP was collected using the SiLK [1] collection system. The network was heterogeneous and globally distributed, with routers at multiple locations within the United States and in several other countries. The majority of traffic is generated within the United States. The network consists of a number of discontinuous address blocks assigned to subcomponents of the ISP. Asymmetric routing policies were commonly used, so that traffic from host A to host B does not cross the border via the same router used for traffic from host B to host A. In any event, NetFlow is unidirectional and the two sides of a bidirectional connection are collected and stored separately. Both incoming and outgoing flow traffic was collected, but matching of forward and reverse flows is not usually done. Traffic that was sent to the null interface on the router as specified by an access control list (ACL) was also collected,

however it was not analyzed for this paper. Null routed traffic consists primarily of traffic destined for TCP or UDP ports known to contain vulnerabilities.

The SiLK tools[2] [1] were used for the collection system. SiLK stores a subset of the information contained in the NetFlow records: source IP address, destination IP address, source port, destination port, protocol, number of packets, number of bytes, start time and duration of each flow. The records are compacted to use the minimum number of bits necessary to represent the recorded information and are organized into hourly files with each hour being partitioned in ways that make many searches more efficient. The archive data is unsorted and serial search is required to extract all records matching a given search criteria within a given hour.

Network traffic from January 2003 through early June 2004 was analyzed. We developed a variety of visualizations to help us understand this data. One of these is the *contact line*, which shows the number of external hosts that contacted a specific number of internal hosts during a specified time interval. Figure 1 demonstrates the *contact surface* that was generated by processing only incoming, routed[3], TCP flows over one week in September 2003. Each hour of the data results in an hourly contact line and the 168 hourly lines are plotted as a surface. The X and Y axes are plotted on log scales, while the Z axis, time, is linear.

The initial analysis was computationally intensive, involving creating a text file for each hour that contains only the source and destination IP addresses for the incoming routed TCP data collected at the border router (so that source IP addresses are always external to the monitored network while destination IP addresses are always internal to the monitored network, regardless of who initiated the session). The result was sorted by source IP and then destination IP, passing the sorted data through `uniq` to remove duplicates, using `cut` to remove the destination IPs leaving only the source, and using `uniq -c` to count the number of destinations associated with each source. Using `cut` again to remove the sources, sorting the counts and, again, using `uniq -c`, counting the number of occurrences of each value. Applied to data from one hour, this gives a single line, which we call a *contact line*. Plotting a time series of contact lines as a pseudo three dimensional plot gives a *contact surface*. These lines and surfaces form the underlying basis for our analyses.

This contact surface (see Figure 2a) has several interesting properties. The first is a persistent linear relationship between the $\log(x)$ and $\log(y)$ values. This is examined in more depth later. The second is a distinct diurnal pattern, particularly observable at $x = 2$. This demonstrates that traffic fluctuates in a predictable pattern with the time of day. (We note that this pattern is much more distinctive for $x = 2$ than for $x = 1$ because the graph was plotted on a log

[2] SiLK stands for System for Internet Level Knowledge, and is named in honour of its creator, the late Suresh L. Konda.

[3] Our data was divided into two partitions: packets that were routed and packets that were dropped due to access control restrictions (e.g., packets destined for particular ports were dropped).

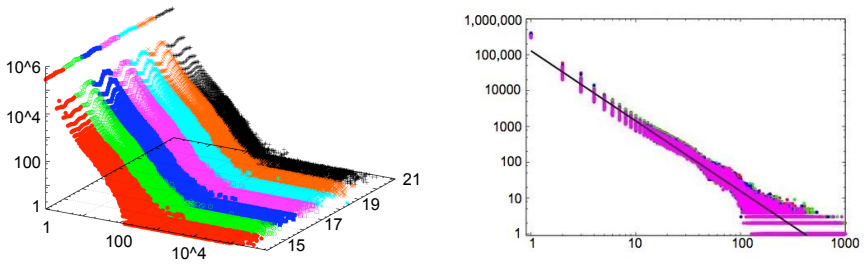

Fig. 2. The contact surface for a week in September of 2003 and its regression line. The y-axis is the number of outside hosts that contacted x inside hosts. The z-axis is time.

scale and so the fluctuations at $x = 1$ were not great enough to be observable at that resolution.) The third is a substantial jump in the number of external hosts that contact only one host per hour when compared to the numbers that contact two or more hosts per hour. While the number of hosts making 2 contacts per hour is about twice the number making 3 contacts per hour, the number making 1 contact per hour is ten times the number making 2 contacts per hour. The large amount of singleton traffic is also seen in traffic from other sources (e.g.Figure 11). The nature and sources of this traffic are the subject of a current investigation. The fourth is the spreading or flat area that occurs for a small number of sources that contact hundreds to hundreds of thousands of destinations per hour. Sources for this area are mostly high volume scanners.

The straight line in the log / log scale exhibits a power-law relationship. We fit a regression line to the week's data shown in Figure 2a. The line has the form $y = e^{11.763367} \times x^{-1.957496}$. in the x, y reference framework or approximately $\log(y) = 5.11 - 1.96 \log(x)$ in the log / log framework. Figure 2b, shows the regression line superimposed on the aggregated contact lines from the five weekdays from Figure 2a. The outliers at count 1 and the spreading at high counts are easily seen. We were somewhat surprised to find a power law relationship here, but note that they describe many other internet traffic characteristics[4]. Faloutsos *et al.* [2] have observed this relationship with regards to out-degree and hop-count.

2.1 The 2003 Disturbance

Our first experience with this traffic was from an earlier time period, the week of January 12 – 18, 2003. The contact surface for this week is shown in Figure 1. This traffic exhibits all of the general characteristics discussed above, except that the one contact outlier is less extreme. In addition, it has an additional property, a perturbation that appears as a"bump" or "waterfall." This perturbation was present across the entire week of data. We examined additional data to determine

[4] It has been observed that everything follows a power law if you graph it with a fat enough marker.

the duration of the perturbation. It was present in the earliest data we had for January 2003. The perturbation was continuously present is all the samples that we examined through early August 2003, however the shape of the wave changed slightly and the 1 contact outlier grew to upwards of a million hosts per hour in July. The perturbation disappeared on August 11, 2003. This is shown in Figure 3a, where one week of data is provided, providing context surrounding August 11.

Figure 3b focuses on the area of the perturbation, with only the Y axis shown on a log scale. The graph shows that the perturbation (two different bumps) is present for the four days previous to August 11, and that it has disappeared for the four days subsequent to August 11, 2003. The largest deviation (or bump) is the second one, which occurs at roughly 20 to 35 external source IP addresses each contacting approximately 150 to 350 internal destination IP addresses per hour.

This traffic was examined further by extracting the source IP addresses that contacted between 150 and 350 destination IP addresses per hour. We discovered that the bulk of the traffic in this region came form three /8 networks, two in the Asian registry, one in the Latin American registry. This distribution was present in each of the weekly samples that we analyzed with roughly constant proportions. The traffic was largely untargeted TCP SYN packets (SYN packets directed at hosts or services that did not exist) destined for port 80. We examined the target distribution for a week in July and found that the targets were not randomly distributed throughout the monitored network. 49% of the flows for this week went to one of the 60 /16s that were being monitored within a single /8 (the remaining 196 /16s are not part of the monitored network), with 14% going to a single /16.

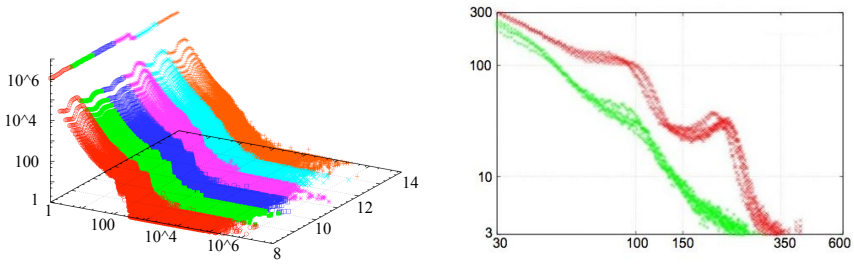

Fig. 3. Wave disappearance details, contact surface and selected lines The y-axis is the number of outside hosts that contacted x internal hosts. The z-axis represents August 9–14.

Given the consistency of this behavior over time, we speculated that the hump was caused by coordinated activity that would exhibit regular temporal behaviour. We attempted to analyze the distribution of the interarrival times for the flows, expecting to see a clustering around the interval (24 seconds for 150

hosts) that would account for the observations if only the monitored network was being targeted.

Our analysis did not support our original hypothesis. The largest cluster (with approximately 5,000,000 observations) shows less than 1 second between successive flows. The detailed analysis did reveal another interesting structure, a scalloping that occurred at regular intervals from about 15 to 70 seconds, after which the decrease becomes more linear (on a log-linear scale). This scalloping indicates a periodic behaviour from some of the sources that we were unable to explain at the time.

2.2 The 2004 Disturbance

In February of 2004, a perturbation in the contact surface was again observed. It started to reappear on February 11, 2004. Figure 4a shows the average traffic patterns for each day from Sunday, February 8, to Saturday, February 14, 2004. The bump first increases in amplitude and then slides to the right during its developmental phase. The number of sources detected peaks at around 50 at the beginning of the period and at about 75 by the end, but the effects of the process can be seen in terms of displacement of Monday / Tuesday baseline. The total number of disturbing hosts involved would be the integral between the baseline and the disturbed line and contains up to 1500 source IP addresses contacting up to 100 or so destinations each. The disturbed behavior continues until June 1, 2004, when it abruptly disappears (see 4b). An examination of source addresses during the disappearance, indicates that the activity ceased as the May 31 / June 1 dividing line progressed around the world.

It is also interesting to note that Figures 4a and 4b display another disturbance in the contact line, a small spike at 150 destinations per hour. We did not examine its sources during our investigation, but, as a result of our simulations, believe that we now know its cause. This will be discussed further in Section 4.1.

There are some differences in the behavior between this occurrence and the previous perturbation. As shown in Figure 5, this time the hump is more pronounced, and there is only one. Both perturbations were caused by traffic to port 80, primarily SYN-only flows. Two of the three /8 networks that had the most

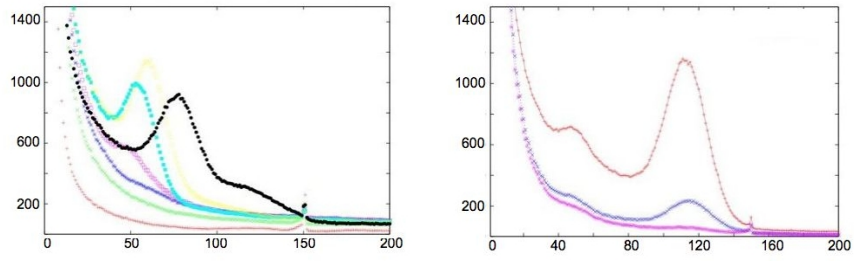

Fig. 4. 2004 perturbation appearance (February) and disappearance (June) details. The y-axis is the number of outside hosts that contacted x internal hosts.

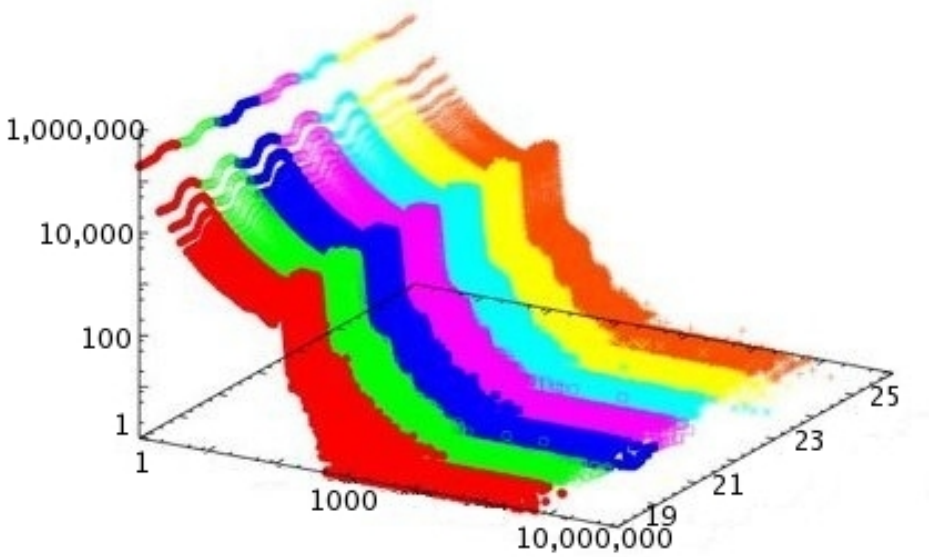

Fig. 5. Contact Surface for April 19 – 25, 2004. The y-axis is the number of external hosts that contacted x internals hosts. The z axis represents time.

sources in the first perturbation appear as the primary contributors to the second perturbation. The third /8 network also contains a large number of sources, however is not in the top three contributing /8 networks. The destination sets, however, were different, with 23% of the traffic targeting a different /8 network from the first perturbation.

3 Hypotheses

At the time this type of perturbation was first observed, the authors were interested in scanning activity. The initial contact surface graph displayed the connection behaviour of the entire monitored network. We hoped to determine whether scanning activity could be easily separated from legitimate network connections. We expected to find a large number of external hosts each of whom contacted a small number of internal hosts, and that this would represent legitimate traffic. We also expected that there would be a small number of external hosts who contacted a large number of internal hosts, and that this would represent scanning activity.

As we note in Section 2, what was actually observed was a power law relationship in the count of external hosts that contact a given number of internal addresses per hour in the central portion of the contact surface with outliers at the one destination per hour end of the line and a high degree of spreading at small source count end of the line. In retrospect, this linear relationship is not surprising. However, the perturbation was not expected and is particularly

interesting because of its variable nature — consistently present or absence for months at a time, with sudden onset and disappearance.

The perturbation was present in the earliest data that we analyzed (January 2003). The perturbation persisted until August 11, 2003, when it abruptly disappeared. Blaster was released on August 11, 2003, and this might be related to the change in behavior that we observed. At first, we thought Blaster might have caused data loss at our sensors, suppressing the data that gave rise to the perturbation. The nature of NetFlow collection using routers is such that substantial data can be lost under heavy load as NetFlow is sacrificed to routing under router overload. In addition, NetFlow is transmitted to the collection point using UDP which also suffers under network load. Had the losses been significant, the entire surface would have been displaced downward and, we suspect, the perturbation reduced, but not eliminated. It is also possible that the sources of the perturbation were taken off line by Blaster. If that were the case, we would expect the systems to resume their previous activity when brought back on-line. This did not happen. The most likely explanation is that the remote systems were also infected with Blaster, which caused them to be taken off the network, patched, and cleaned to remove all sources of malicious activity. We suspect that this cleaning removed the cause of the unusual behavior observed at our monitoring points and that the patching prevented the source from reestablishing the behaviour. We suspect a scanning worm of some kind, possibly exploiting a vulnerability in port 80 that was patched at the same time as the DCOM vulnerability[5] used by Blaster, but have not identified a specific candidate.

The 2004 perturbation appeared on 11 February and disappeared as June 1 arrived. In December of 2004, Alfred Huger of Symantec noted that the dates of the appearance and disappearance of 2004 perturbation exactly matched the onset and demise of the worm Welchia.B[6]. Welchia.B contained a "suicide" timer that accounted for its demise along the dateline between May 31 and June 1. We were able to persuade a colleague to examine a corpse of Welchia.B and discovered that the main scanning loop of the worm contained a "sleep" system call with a delay constant that appeared to account for the disturbance when the scanning rate and the percentage of the total IPv4 address space being monitored were taken into account.

At this point, we had a plausible explanation for our observations, but lacked a confirmation. Not too long afterwards, we took other positions and subsequently lost access to the data source, so that further analysis was not possible. In 2007, we learned from former colleagues that the perturbation phenomena had reappeared. Development of code to visualize the contact surface is part of an unrelated analysis project, and we decided to revisit the original problem to refine our hypotheses and determine whether we could develop a simulation that provided a plausible explanation for the earlier observations. Based on the

[5] `http://www.cve.mitre.org/cgi-bin/cvename.cgi?name=CAN-2003-0352`

[6] W32.Welchia.B.worm was a relatively minor threat, in the general scheme of things. See `http://www.symantec.com/security_response/writeup.jsp?docid=2004-021115-2540-99` for additional details.

earlier work, we developed several hypotheses that serve to focus the analyses and simulations.

Hypothesis 1. *The perturbation of the contact surface is caused by the presence of persistent scanning behavior (such as would be exhibited by a worm-infected host) with a fixed time delay between each scan probe. This delay is constant across the infected population.*

Note that this hypothesis implies a coordinated activity, however, the coordination may well be preprogrammed. All that is required is that each participant scan at the same rate.

Hypothesis 2. *The targets of the scanning are essentially random so that they are not easily observed without a network telescope with an aperture that encompasses substantial address space (several /8s or more).*

There is a tradeoff between the strength of the observed signal and the telescope aperture. For example, a single source emitting 1 randomly addressed probe per second would be seen about once every 4 minutes if the aperture is the equivalent of a /8 while it would be seen about once every 10 days if the aperture is a /24. If the scans are targeted so that the percentage of the total probes that are intercepted is disproportionate to the address space monitored, the signal strength increases.

We noted in Subsection 2.2 that we observed a small spike at 150 addresses that was consistent over time. While we did not investigate that spike further to determine the characteristics of the IP addresses generating the spike, we believe that it was due to scanning activity from several sources whose targets were largely within the monitored address space. This gives rise to an additional hypothesis.

Hypothesis 3. *Sharp spikes in the contact surface are due to a group of hosts that all scan addresses within the monitored address space at a fixed rate.*

Note that there are several limiting cases here. A small number of dropped packets are equivalent to a scan that largely, but not completely, targets the monitored address space. In addition, scans that target one or more complete subnets and are carried out so rapidly that all addresses are probed within the interval of the analysis (one hour in our case), will also generate spikes. If all probes from all sources scanning full subnets are observed, the spike becomes a point whose amplitude is the number of scanners and whose position on the target count axis reflects the size of the scanned subnets.

In the next section, we explore these hypotheses using a combination of simulation and analysis of data from a /22 network that we have monitored for several years.

4 Analysis and Simulation

The correlation between the onset and demise of Welchia.B and the 2004 perturbation provided evidence that the observations could be due to regular

behaviours of a small population of infected machines. The discovery of a timed scanning loop in Welchia.B provided a mechanism for regular behaviour. We decided to see if we could reproduce the observed perturbations in a controlled manner. In order to do this, we needed to generate appropriate background behaviour and perturb it according to our hypotheses. Because we were not analyzing flow data *per se* in constructing the contact surface, we can avoid the task of simulating millions of individual hosts and concentrate on the essential characteristics of the background traffic as seen in the contact surface. This admits some simplifications.

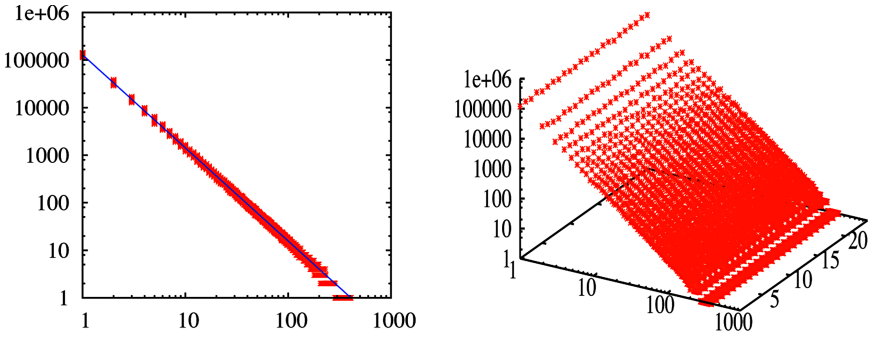

Fig. 6. Base traffic simulation, 2D (6a) and 3D (6b) views, 4% of IPv4 monitored (12 /8s), 40% noise spread, 24 hours with no perturbations. Vertical axis is number of outside addresses seen per hour. Horizontal axis is number of inside addresses targeted per outside address. Time is into the page in the 3D case.

In Section 2, we noted that there is always an outlier at the 1 host per hour portion of the fit. Far more outside hosts make single contacts than would be predicted by the model. A similar phenomenon occurs at the other end of the scale where a few (typically less than 5) hosts contact each of a very large number (thousands to hundreds of thousands) of destinations per hour. We believe that, in the undisturbed (no perturbation) case, that the contact lines represent three independent phenomena. The excess of single destination hosts represents a very low frequency noise component. For about a year, we have been analyzing NetFlow data from a local /22 network that is about 10% occupied. In a typical hour, on the order of five thousand external sources each contact single inside addresses. Extending the filtering period to a day gives an average of about 62 thousand, with a little over a million external addresses contacting a single internal address in the course of a month. For the year, there are slightly over 10 million external addresses that each contact a single inside address and more than half (5.4 million) are active in only a single hour of the observation period. For reasons that we are still investigating, the diurnal variations in this low frequency component are much smaller than those in the main portion of the traffic, and the diurnal variations in the regular data are suppressed by the

log scale presentation of their sum. Inspection of other end of the scale indicates that the spreading is due to a substantial number of bulk scanners who systematically probe the monitored network. We do not include either the low frequency or bulk scanning components in our base model.

The simulation is written in Snobol-4[7], enhanced with a Mersenne Twister random number generator[8]. The simulator produces scripts for the `gnuplot` graphics program. The simulation is constructed in two parts. The first uses the regression line, $e^{11.763367} \times x^{-1.957496}$, noted above and generates the expected number of sources contacting each destination count. Noise is added using a triangular distribution that spreads each point by a fixed percentage of its value (constant width on a log / log scale). It is important to note that the sole purpose of this portion of the simulation is to provide a **realistic appearing** base and **not** to emulate the processes that actually produce the base. The noise spread is included in the base so that the injected perturbations will not be completely obvious at low levels of disturbance. We have not included the diurnal and weekly variations found in the real data. Figures 6a and 6b show two and three dimensional views of the base traffic. These were created by setting the perturbation parameters to zero and assuming the fit parameter, 4% of the IPv4 address space associated with the regression line. It is interesting to compare Figure 2b with Figure 6a. In the central region, the figures are sufficiently similar so that we can claim that the base traffic generation is adequate. This base traffic is used for all of the subsequent simulations, normalized for different monitored percentages as necessary.

To simplify the simulation of the perturbation process, we assume that the perturbers scan at a constant rate, dictated by some delay loop. We also assume that they generate random IP addresses, over some portion of the Internet, up to and including all of the IPv4 address space, $0 \cdots (2^{32} - 1)$. If this is the case, our monitoring network will collect a fraction of the probes, based on the amount of address space being monitored and the percentage of probes that target that space. The simulator allows us to specify these parameters as well as the number of probers and the rate at which they probe. We have run a series of sensitivity analyses in which we vary the coverage, i.e. the percentage of the Internet being monitored, from 4 /24s to 20 /8s, assuming 1000 probers each sending 2 probes per second. Similar runs vary the number of probers or the probe rate while holding other parameters constant. The simulation is "brute force" in that we simulate each prober separately. For each scanner, we generate a random number in $\{0.0 \cdots 1.0\}$ for each scan it would have emitted during the hour. The number of observed probes is the count of random numbers with values below the monitored network percentage adjusted for the assumed probe range[9]. This number is the number of destinations reached by the prober and we add 1 to the appropriate base cell in the array holding the contact line.

[7] Phil Budne's C implementation, version 1.1 from `http://www.snobol4.org`

[8] `http://www.math.sci.hiroshima-u.ac.jp/~m-mat/MT/emt.html`

[9] If we assume that the probes target just the monitored network, or some portion of it, this count will approach the probe rate.

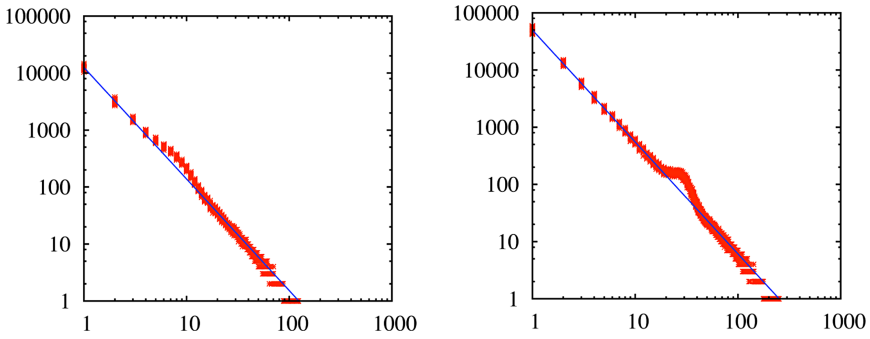

Fig. 7. Simulation of 1000 perturbers at 1800 probes per hour. 1 /16 (7a) and 1 /8 (7b) monitored.

Figures 7a and 7b indicate that a perturbing population of of 1000 sources at one probe per 2 seconds each *might* be visible if a single, relatively quiet /16 was being monitored but should be quite visible if a /8 is being monitored. This appears to confirm hypothesis 2 and provides guidance for future investigations of observed perturbations.

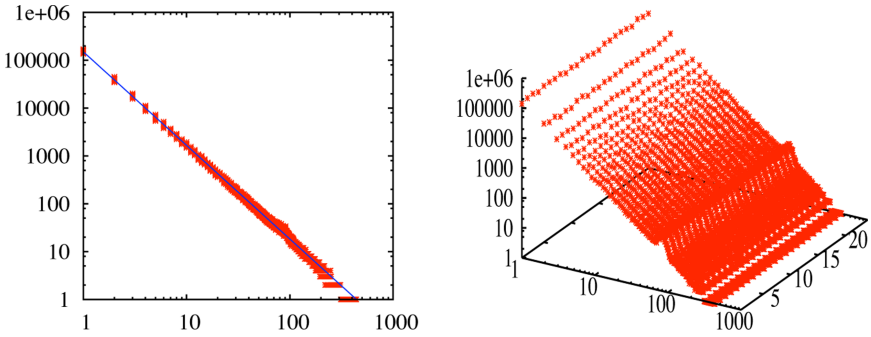

Fig. 8. Simulation of 100 (8a - 2D) and 1000 (8b - 3D) perturbers at 1800 probes per hour, 12 /8s monitored

Similarly, Figures 8a and 8b indicate that 100 sources at 1 probe per 2 seconds *might* be visible if 12 /8s are being monitored and that 1000 will be clearly visible. Note that the disturbance moves downslope and grows in amplitude as the coverage increases while the amplitude of the disturbance grows in place as the number of probers increases. In both cases, the width of the disturbance reflects the randomness of the interception process.

Figures 9a and 9b indicate that a disturbance caused by a population of 1000 probers each issuing one probe per 10 seconds (360 per hour) is barely visible in

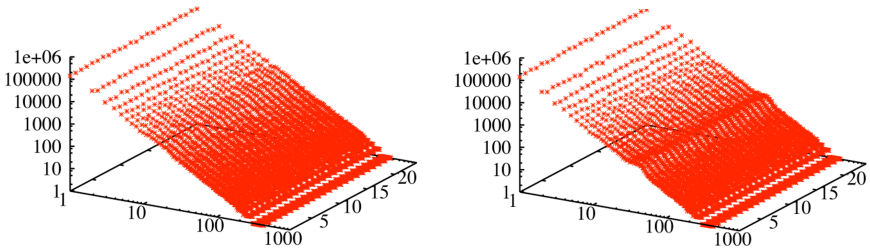

Fig. 9. Simulation of 1000 perturbers at 360 (9a) and 900 (9b) probes per hour, 12 /8s monitored

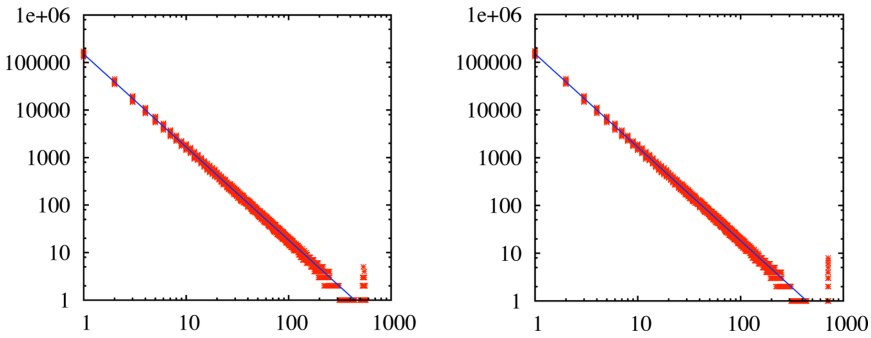

Fig. 10. Simulation of 20 perturbers at 720 probes per hour, 75% (10a) and 99% (10b) hit rate, 12 /8s covered

the artificial background, while a similar population is easily seen with a probe rate of one per 4 seconds. Again, the disturbance moves down slope and increases in amplitude with increasing probe rate. The above examples serve to describe the approximate limits of visibility for regular probes as a function of probe population and rate and observational coverage. In this part of the analysis, we assume that the percentage of probes seen is equal to the percentage of the address space that is being monitored. While this is true for a number of random scanning strategies, targeted scans will manifest differently. The simulated perturbations embody the strategies implicit in Hypothesis 1. The resemblance to the observed data is gratifying, but we cannot say conclusively that these are the only assumptions capable of producing the observed phenomena.

4.1 The Minor Spike

Figures 4a and 4b show minor spikes at the 150 contacts per hour point in addition to the broad disturbances seen earlier. We did not have an opportunity to analyze these at the time, but realized that the simulation also provides a plausible explanation for these as well. If the probing strategy is such that the

probes fall entirely within the monitored network, the disturbances sharpen as can be seen in Figures 10a and 10b. In this case, we are considering a population of 20 probers with 75% and 99% of the probes being seen in the monitored network. The probe rate is 720 per hour (1 every 5 seconds) per source. Note that, if the hit rate were 100% for all sources, the spike would become a single dot whose amplitude indicated the number of sources. We see exactly this behaviour in the /22 network that we have been monitoring. This simulation is consistent with Hypothesis 3.

4.2 Full Subnet Scanning on a /22

Figure 11 shows contact data for a month from the /22 that we have been monitoring. The data was first filtered to retain only traffic from an external source address to a destination addresses within the monitored network then sorted by flow start time. The data was then filtered using a Bloom filter so that only the first record for each unique source IP / destination IP pair was kept for subsequent analysis. The retained data was passed through to the SiLK `rwbag` program and a bag or multiset made for the source IP addresses. This bag counts the number of internal addresses contacted by each external address. Inverting this bag provides the data pairs used to produce the contact line. The line is similar to those composing the contact surface seen in the Figure 2a.

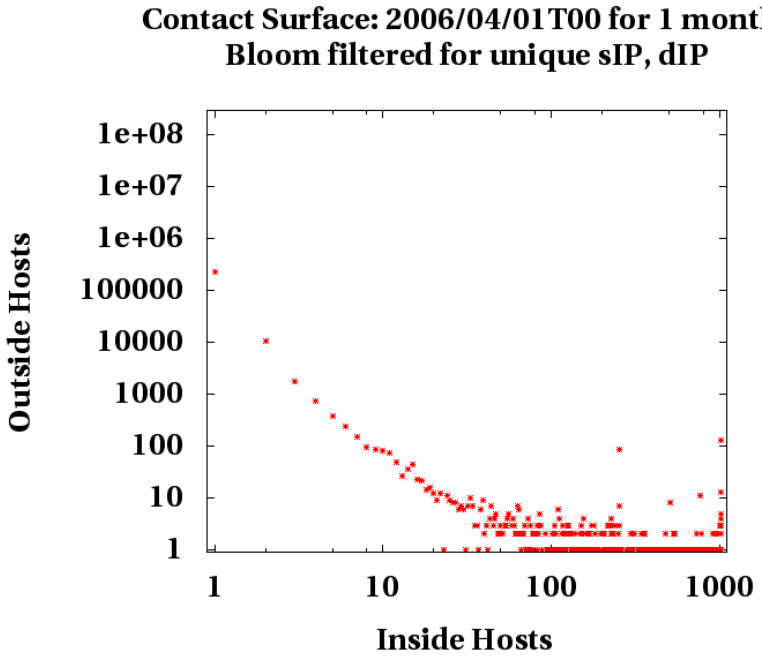

Fig. 11. Observed Contact line for a /22 in April 2006

The general shape of the curve is similar to those for the larger network. Several hundred thousand external addresses appear associated with only a single internal address. In other months, this number is as much as a million. At the bottom of the figure, one to three addresses contact most of the host counts between 50 or so and 1016[10].

Notable at the right hand end of the figure are high points, approaching 100 sources at 254 and 1016 destinations with smaller peaks at 508 and 762. A closer inspection shows that significant numbers of sources attempt to connect to one to four of the monitored subnets. Most probe all of the monitored addresses, but smaller numbers contact nearly all, missing only one or two, resulting in secondary points, as well. Inspection of the corresponding daily and hourly traffic shows that the full scans are distributed throughout the month and are not easily seen on a shorter time scale. These observations appear to be consistent with the limiting cases of Hypothesis 3. The singleton traffic is the subject of a current investigation. We note that during the 14 that months we have analyzed, nearly 13 million outside sources contacted the network. About 42% of these generated only a single flow record and over 90% generated 10 or fewer flows.

5 Related Work

The contact surface described in this paper was first shown in a paper by McHugh and Gates [3] on locality. However, at that time, the disturbance in the contact surface was only noted, but not analyzed nor was it presented as a subject for speculation. At that point it was not known that the perturbation would disappear in August 2003. It appeared again in a later paper by McHugh *et al.* [4], however at that point the cause of the perturbation was still unknown and, again, the disturbance was not analyzed. Since that time the suspected cause of the 2004 perturbation was discovered, and the contact surface was the subject of several invited presentations but not of any publications.

While we know of no other work that has demonstrated a contact surface, nor demonstrated observed effects of security-related phenomena on large-scale traffic analysis, some work related to the contact surface can be found in the network traffic analysis literature. Network traffic analysis traditionally has focused on an examination of traffic volumes or round-trip-times. For example, Paxson and Floyd found that WAN traffic was largely self-similar in nature [5], exhibiting fractal-like scaling behavior and a heavy-tailed distribution over various time scales. Feldmann *et al.* [6] studied this further, relating the impact of the local networks on the traffic characteristics to the physical construction of the network. Later, Chen modeled traffic volumes using an ARIMA (Auto-Regressive Integrated Moving Average) model [7]. He analyzed traffic based on subnetworks, such as analyzing all http traffic in isolation (rather than by LAN,

[10] While the /22 contains 1024 addresses, the maximum count that we see is 1016. Addresses 0 and 255 in each /24 do not appear, having been "absorbed" within the instrumented router.

as done by Feldmann *et al.*), and suggests aggregating each of these subnetworks together to better model overall network traffic. Traffic volumes have also been analyzed for security events (*e.g.*, denial-of-service attacks) and failures, using approaches such as signal processing [8]. Lee and Fapojuwo review several statistical techniques for analyzing network traffic [9]. However, these studies and others have not examined the interhost communication characteristics we observe in the contact surface.

Some work has been done on analyzing host-to-host communications. For example, Epsilon *et al.* [10] examined host-level network traffic characteristics for ATM traffic at an ISP, finding that host traffic is highly non-uniform (with a few servers accounting for the most traffic). They also analyzed connection information in terms of typical traffic volumes, however no analysis was performed on typical connection patterns (such as how many servers a typical client accessed). Sarvothan *et al.* [11] do a similar connection-level analysis, noting that there are two types of traffic — alpha and beta — where alpha traffic is dominated by a few flows transferring large amounts of data over high-bandwidth connections whereas beta traffic consist of the remaining flows with smaller data transfers and lower bandwidths. Lakhina *et al.* [12] examined OD (origin-destination) flows on backbone networks using PCA, however they defined the origin as a network ingress point and the destination as a network egress point, rather than as the source and destination hosts respectively. Lakhina *et al.* [13] also examined network traffic using clustering and entropy approaches on source and destination IP addresses, but did not combine the two as done in the contact surface. Dübendorfer *et al.* [14] analysed the effect of worm traffic on an internet backbone, where they aggregated the number of unique sources seen over time. However, they did not split this into the third dimension, aggregating again by number of destinations contacted. Karagiannis *et al.* [15] do examine network traffic at what they term the "social level", analyzing the communication of a single host with regards to the number of destination IPs contacted for particular types of traffic (*e.g.*, web, p2p, malware and mail), however they do not aggregate this information across multiple sources.

Visualization software has been developed to help administrators better understand their network traffic and better detect anomalies. Such visualizations have typically focused on host to host behavior, such as providing three dimensional graphs indicating the traffic relationships between external host, internal host and destination port (see, for example, [16] and [17]). Goodall *et al.* [18] developed a Time-based Network traffic Visualizer (TNV) demonstrating context and time for network traffic. TNV provides a visual representation of the network traffic between hosts, but does not aggregate in the form shown in the contact surface, instead focusing on traffic between individual hosts. Oberheide *et al.* [19] present a similar tool, showing traffic volumes per host, or non-aggregated interhost relationships. Interestingly, they show how the Dabber worm appears using their interface, along with traffic both before and after a slashdot event. However, as they do not look at the aggregated traffic, they do not observe the same phenomena presented in this paper.

6 Conclusions and Acknowledgments

We have developed a graphic representation for large scale Internet connection behavior and have used it to investigate two outbreaks of what appears to be synchronized activity by significant populations of scanning hosts. It appears that the synchronization arises from a "design time" choice of a delay constant in the scanning loop and that this allows a small population of scanners to create a pronounced disturbance in the midst of the activities of millions of others. We have explored the phenomena through simulation and believe that we have plausible explanations for a number of features that appear in observed contact lines. We realize that there is some risk in publishing this kind of result. It would be trivial to modify the scanning mechanism so as to avoid the observed phenomena. As part of our future work in this area, we will investigate the effect of such changes, noting that both increases and decreases work against the scanner, raising detectability or reducing effectiveness. Given access to suitable data, we expect to discover additional periodic phenomena, as well. In addition, the development of the techniques used to display and analyze this phenomena has aided us in performing more immediate tasks as well as serving to identify other research areas of interest. The reviewers of the paper made a number of helpful suggestions, including the performance of more detailed analyses of the 2003 and 2004 outbreaks. We agree that these analyses should be performed, but we no longer have access to this data and know of no comparable sources to which we might obtain access. Thus far, we have been unable to persuade individuals who have current access to collaborate. We want to thank Tom Longstaff for his encouragement when we were all at CERT, Michael Collins and Mark Thomas for the initial and continuing support of the SiLK tools. We owe a debt of gratitude to the late Suresh L. Konda for the vision that made our discoveries possible.

References

1. Gates, C., Collins, M., Duggan, M., Kompanek, A., Thomas, M.: More NetFlow tools: For performance and security. In: Proceedings of the 18th Large Installation Systems Administration Conference (LISA 2004), Atlanta, Georgia, USA, pp. 121–132 (November 2004)
2. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: Proceedings of the 1999 ACM SIGCOMM Conference, Cambridge, MA, USA, August 31 - September 3, pp. 251–262 (1999)
3. McHugh, J., Gates, C.: Locality: A new paradigm for thinking about normal behavior and outsider threat. In: Proceedings of the 2003 New Security Paradigms Workshop, Ascona, Switzerland, pp. 3–10 (August 2003)
4. McHugh, J., Gates, C., Becknel, D.: Situational awareness and network traffic analysis. In: Proceedings of the Gdansk NATO Workshop on Cyberspace Security and Defence: Research Issues, Gdansk, Poland, pp. 209–228 (September 2004)
5. Paxson, V., Floyd, S.: Wide area traffic: The failure of Poisson modeling. IEEE/ACM Transactions on Networking 3(3), 226–244 (1995)

6. Feldmann, A., Gilbert, A., Willinger, W.: Data networks as cascades: investigating the multifractal nature of internet wan traffic. In: Proceedings of the ACM SIGCOMM 1998 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Vancouver, British Columbia, Canada, pp. 42–55 (1998)
7. Chen, Y.W.: Traffic behavior analysis and modeling of sub-networks. International Journal of Network Management 12(5), 323–330 (2002)
8. Barford, P., Kline, J., Plonka, D., Ron, A.: A signal analysis of network traffic anomalies. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet Measurement, Marseille, France, pp. 71–82 (2002)
9. Lee, I., Fapojuwo, A.: Statistical methods for computer network traffic analysis. IEE Proceedings Communications 153(6), 939–948 (2006)
10. Epsilon, R., Ke, J., Williamson, C.: Analysis of ISP IP/ATM network traffic measurements. ACM SIGMETRICS Performance Evaluation Review 27(2), 15–24 (1999)
11. Sarvotham, S., Riedi, R., Baraniuk, R.: Connection-level analysis and modeling of network traffic. In: Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement, San Francisco, CA, USA, pp. 99–103 (2001)
12. Lakhina, A., Papagiannake, K., Crovella, M., Diot, C., Kolaczyk, E., Taft, N.: Structural analysis of network traffic flows. In: Proceedings of the 2004 Joint International Conference on Measurement and Modeling of Computer Systems (SIGMETRICS/Performance), New York, NY, USA, June 12-16, pp. 61–72 (2004)
13. Lakhina, A., Crovella, M., Diot, C.: Mining anomalies using traffic feature distributions. In: Proceedings of the ACM SIGCOMM 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, PA, USA, pp. 217–228 (2005)
14. Düberdorfer, T., Wagner, A., Hossmann, T., Plattner, B.: Flow-level traffic analysis of the blaster and sobig worm outbreaks in an internet backbone. In: Proceedings of the 2005 Conference on Detection of Intrusions and Malware and Vulnerability Assessment, Vienna, Austria, pp. 103–122 (2005)
15. Karagiannis, T., Papagiannaki, K., Faloutsos, M.: BLINC: Multilevel traffic classification in the dark. In: Proceedings of the ACM SIGCOMM 2005 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, Philadelphia, PA, USA, pp. 229–240 (2005)
16. van Riel, J.P., Irwin, B.: Inetvis, a visual tool for network telescope traffic analysis. In: Proceedings of the 4th International Conference on Computer Graphics, Virtual Reality, Visualization and Interaction In Africa, Cape Town, South Africa, pp. 85–89 (2006)
17. Lakkaraju, K., Yurcik, W., Lee, A.J.: NVisionIP: NetFlow visualizations of system state for security situational awareness. In: Proceedings of 2004 CCS Workshop on Visualization and Data Mining for Computer Security, Washington, DC, USA, pp. 65–72 (October 2004)
18. Goodall, J., Lutters, W., Rheingans, P., Komlodi, A.: Preserving the big picture: visual network traffic analysis with tnv. In: Proceedings of the 2005 IEEE Workshop on Visualization for Computer Security, Minneapolis, MN, USA, pp. 47–54 (October 2005)
19. Oberheide, J., Goff, M., Karir, M.: Flamingo: Visualizing internet traffic. In: Proceedings of the 10th IEEE/IFIP Network Operations and Management Symposium, Vancouver, BC, Canada, pp. 150–161 (2006)