Network Traffic Visualization

Fabian Popa
Seminar Innovative Internet-Technologien und Mobilkommunikation, WS 2008/2009
Institut für Informatik, Lehrstuhl Netzarchitekturen und Netzdienste
Technische Universität München
fabian.popa@mytum.de

ABSTRACT

This paper discusses Network Traffic Visualization, covering the motivation, challenges, and different approaches. In starts off by analyzing the relevant types of data needed, as well as collection methods employed in generating data sets. It then moves on to the ways in which these sets of network traffic information are mapped onto an image, in order to accentuate specific characteristics of the traffic. Sample architecture is presented for a distributed information collection system. In the second part of the paper, the focus shifts toward real-world systems and methods currently in use, providing sample outputs and emerging benefits. Finally, the paper extends a conclusion on the visualization methods discussed, revisiting specific applications, and glimpsing into the future of network traffic visualization.

1. INTRODUCTION

Computer networks, and specifically the Internet, are the foundation and enabler of our fast-moving information society. In our day to day lives, we are explicitly and implicitly relying on computer systems, connected in networks that range from home or corporate LANs (Local Area Network) to global and ever growing WANs (Wide Area Network). Weak, unstable networks, as well as unpredictable network activity, can severely damage or hold back operations of all magnitudes. But how can we make sure that a network is not vulnerable? This is where network traffic information comes in.

The analysis of network traffic data can provide indicators about the state of a network. By monitoring the network and looking for specific markers, anomalous behavior can be identified and addressed in a timely manner. Unfortunately, not all data bears relevance, and, while readable by a machine, it is usually difficult to interpret, in its raw form, by a human. Therefore, the right data has to be found and then converted into an accessible form.

Toward this purpose, methods have been devised to visually represent traffic information, making it human-readable, but also analyzable from a different standpoint, that of image processing. Weak points in the network, and malicious behavior such as DDoS attacks and scanning activities, are specifically targeted and highlighted. There are multiple reasons for visualizing network traffic information, which will be discussed in the following.

1.1 Motivation

Visualizing information is a technique that can encode large amounts of complex interrelated data, being at the same time easily quantified, manipulated, and processed by a human user. Therefore, it is an obvious candidate for the representation of network traffic information. Using specific techniques, the state of the network can be depicted in such a way that anomalous activities will display as objects inside the traffic image. These can be identified by image-processing algorithms, as well as the system administrator looking at the image. Furthermore, images can be compressed, enabling size reductions and faster communication and analysis of the data.

Indeed, not only information on anomalous and malicious activities is desired, but also on the connectivity and performance of the system. In this case, topological and geographical representations bear more meaning. In a topology map or network graph, specific measures can be applied, such as the critical paths between two subnets, or the shortest path between two peers. Furthermore, looking at a single node or subnetwork, notions like "reachability" are highly important and can be measured and displayed in an intuitive manner using a topological approach.

In the end, network traffic visualization aims at providing a clear overview of the state of a network or subnetwork, in a way that aids system administrators and network architects in maintaining the integrity, availability, and reliability of the network, as well as plan for capacity increases, new communication protocols, and expansions.

1.2 Challenges

Traffic visualization is effective in dealing with specific network issues, as studies discussed later in this paper show, but it faces important challenges, mostly due to the sheer size of the Internet.

Monitoring activities produce large volumes of data, which need to be efficiently communicated, stored, and processed. These become an even bigger concern when real-time representation is desired, although this is not always the case.

Furthermore, when providing visualizations for the human reviewer, they should be efficient and easy to understand, without the need for lengthy in-depth analysis. The right method and the right data for a specific task have to be found and combined to reveal the important aspects visually.

The structure, technology, and bandwidth of networks are changing at a fast pace. It is increasingly difficult to measure large networks, let alone analyze the traffic.

When looking at changes over a period of time, differences in traffic states need to be accentuated as clearly and effortlessly as possible.

It is important to choose the right data and methods of processing, but it is equally important to look at the right data. Finding the best combination of the two is not a light task.

2. METHODOLOGY

Network traffic data is acquired in a single or distributed environment, by different means. Specific to the application, it is then communicated and processed or directly processed, at which point visualization is employed. The resulting images can be passed on, archived, or further analyzed programmatically (e.g. object recognition in the image).

Different participants in the provision of the Internet Service have different concerns and would look at different network information. For instance, an ISP (Internet Service Provider) will look at data regarding the usage of the service by its users, and aim at optimally filling its network capacity and choosing its network partners (other ISPs or large network providers). In this process, it would consider the bandwidth and types of content

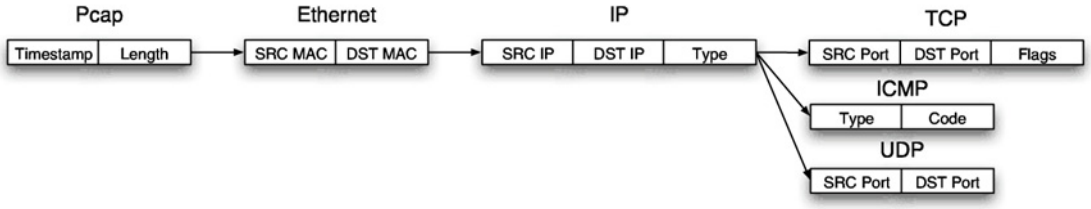

Figure 1. Packet header fields parsed for statistics gathering [3]

consumed by the users and reasonably plan capacity and pricing models. On the other hand, the regular Internet user might judge the performance of his or her connection only by looking at how promptly a given website is loading. While this may be a more naïve approach, it is nevertheless valid.
How would we go about measuring the traffic of a network properly? The steps to acquiring and preparing the data for visualization are discussed below.

2.1 Types of data

Some types of traffic data bear more meaning than others. Before we can look at how to obtain the required data, first we need to identify the right data for our purpose. To analyze behavior inside a network, the following traffic characteristics are typically considered:

- **Packet Header Fields (D1)**:
  source/destination IP, source/destination port, time to live (Figure 1). The packet header information characterizes the flow of the data packet through the network (direction, time). It provides parameters for visualization methods and statistical procedures.
- **Round Trip Time (D2)**:
  time elapsed for a packet to reach a destination address and return to the sender. This is a network connectivity and performance indicator.
- **Packet Hop (Routing) (D3)**:
  the route of a packet through the network nodes. The communication route between two peers can differ in one direction from the other. Therefore, the data packet may be directed on a longer path in one direction and may take longer to reach its destination.
- **Bandwidth (D4)**:
  bandwidth consumption for incoming/outgoing traffic indicates where the activities taking up the most resources are occurring. Consequently, based on the destinations/sources of the traffic from/to one address, network attacks can be identified and restricted.
- **BGP Tables (D5)**:
  a network router looks at the destination IP address of the data packet and uses a BGP (Border Gateway Protocol) Table of addresses to figure out the next hop toward the target.

There are additional metrics as well, but the aforementioned provide a sound depiction of the network state.
The Internet data cannot be recorded in its entirety, due to the giant scale. Therefore, it needs to be sampled and looked at over a specific timeframe. Now we need to choose a method to acquire the data.

2.2 Data acquisition

The basis for visualization is the data set, collected over a given timeframe. There are two ways in which traffic data can be collected:

- **active means**:
  specially crafted traffic is introduced into the network. It can be observed at receivers and it can trigger a response, which returns to the sender. Sample applications include estimating the bandwidth of an Internet link and determining the path connecting two computers.
- **passive means**:
  data is collected at strategic locations, without introducing new traffic into the network. Sample applications include "telescopes", listening to incoming packets and requests (traffic with no legitimate destination).

"In typical (and simple) cases, the *active* measurements can be used for direct quality investigation of end-to-end communications (traffic performance), while the *passive* method is used to collect figures for network-internal statistics (traffic load, traffic matrix, etc)". [10]
When looking at measurements regarding the Internet, simple approaches can often not be employed. Here, the need for distributed systems becomes evident. Such architecture is presented in the following paragraph.
With regard to the global Internet infrastructure, CAIDA [4] provides tools and analyses promoting the engineering and maintenance of robustness and scalability. They also provide measurement, topology, and routing data sets, which are the basis for some visualizations discussed later on.

2.3 System structure

Designing an effective data acquisition system can determine the speed, accuracy, and, most of all, reach of traffic measurements.
In a simple setting, a network survey (active means), for example, can be conducted from one location where traffic is sent from this one source to many destinations, recording the results. Figure 2 depicts such a setting.

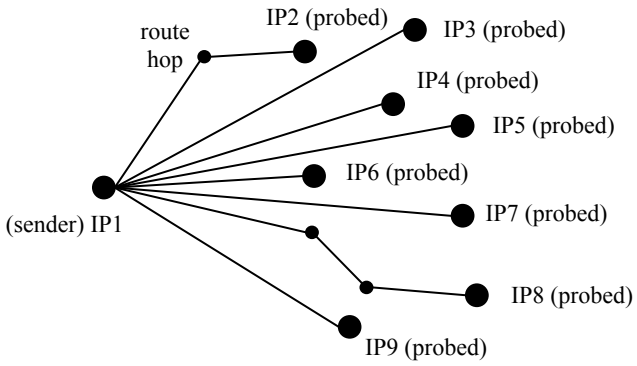

Figure 2. One location network survey. [no ref.]

However, given the sheer size of the Internet, even a straightforward measurement can require multiple senders. In practice, more frequently employed are complex distributed systems. This applies for passive data collection as well. Telescopes, for instance, could be installed in one or more locations, depending on the size or reach of the network. The bigger the scope (telescope's "lens"), the more accurate the inference about the overall state will be.
The configuration of a distributed system (Figure 3) typically contains the following components, although some modules are specific to a visualization technique discussed later on (2D plane conversion, Space-Filling Curves):

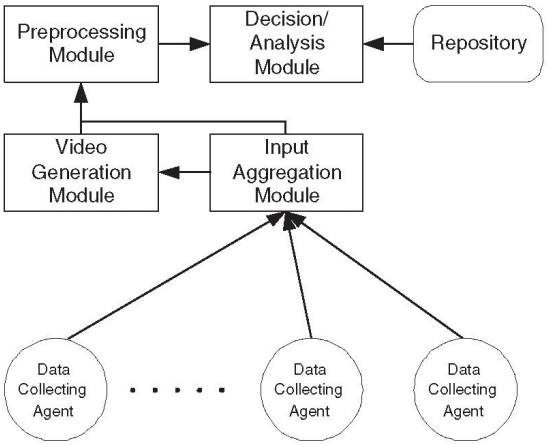

Figure 3. Distributed system design. [1]

a) **Data Collection Agent**:
these are distributed sensing modules that collect real time statistics from the locations they have been installed in. The gathered data is processed on-site. Here, the initial visualizations (images) are generated over a given time period, then aggressively compressed to save bandwidth. The images are sent to the aggregation module.

b) **Aggregation Module**:
after receiving the data (i.e. compressed images) from the agents, it is prepared for analysis. Normalization and synchronization operations are carried out, so that a consistent data set can be passed on to the video generation module.

c) **Stream (i.e. Video) Generation Module**:
the incoming multiple streams of data are converged into a single stream, that depicts the state of the system as a whole (i.e. video composed from the images received over successive timeframes). Information resulting from the formation process, such as peak intensity locations, is passed along.

d) **Preprocessing Module**:
objects in the stream (i.e. video) are identified and their characteristics are recorded (location, trajectory, speed, brightness, shape, size, etc.)

e) **Decision/Analysis Module**:
based on the information provided from the previous modules, it can be inferred if the activity is normal or anomalous. In the decision process, the engine draws historic information from a repository, where known patterns, special cases, and system history are saved. The Decision Module can automatically take action or notify the administrator of certain changes or happenings.

The basic architecture of collection (single/multiple location), preprocessing, analysis, and repository is present in one form or another in every distributed network traffic analysis system.
Looking at the network traffic visualization system as a whole, the Aggregation Module concludes the first step (data acquisition), by outputting the raw data set. The data can be processed automatically as-is, or it can be visualized. Visual data, as mentioned before, can be handled not only by a human reviewer, but also by image processing algorithms.
In the paragraph following, different traffic measurement and visualization systems are discussed and their imaging approaches evaluated.

3. VISUALIZATION

Different visualizations are adequate for different tasks, and work best with specific data sets. For the imaging (mapping) of Internet Address Space, 2D plane conversions have been heavily discussed, mostly because of two reasons: they provide a comprehensive overview of the state of the Internet (they can map the whole IPv4 address space), and they retain traffic properties when certain mapping approaches are used and the resulting images compressed. This achieves bandwidth and processing time savings. On the other hand, when looking at issues such as address reachability, graph-based representations are preferred. Furthermore, for human readability, geographical map overlays are popular, in conjunction with more freely-chosen, but still adequate, visualization techniques. For each of these categories, we will now discuss methodology and applications. We will follow the structure discussed: data set, collection method, visualization, inference (benefits).

3.1 Ant census of the Internet Address Space

Starting in 2003, researchers at ISI [5] have been collecting data about the Internet. As part of this work, they have been probing all addresses in the allocated IPv4 Internet Address Space for their reachability. This is a type of active "one location network survey" application (Figure 2). The researchers at the ISI have sent an ICMP ping message to all the addresses in the IPv4 address space. ICMP ping is an "echo request" packet. If the packet reaches its destination, it will trigger an ICMP "echo response". The sender listens for this reply, and records Round Trip Time (**D2**), as well as packet loss. In this case, the quality of response from a destination IP constitutes the data set.
Naturally, because the IPv4 (Internet Protocol, version 4) provides 2^{32} (around 4 billion) addresses, choosing the right way of visualizing the address space is of great importance.
The researchers have chosen a layout first suggested in the popular web-comic xkcd [6]. Here, the one-dimensional, 32-bit address space is converted into two dimensions using a Hilbert Curve, as shown in Figure 4.

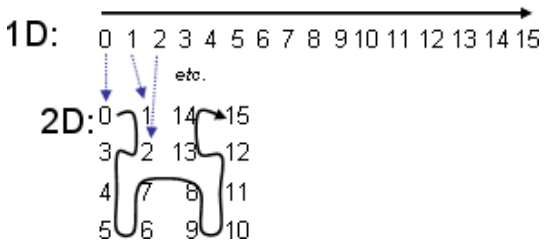

Figure 4. Hilbert curve 1D to 2D conversion. [5]

The Hilbert Curve was first described by the German mathematician David Hilbert in 1891, and presents a number of properties relevant to the mapping of IP Address Space. First, it is space-filling, which means that it will fill a "square unit" entirely (if the 1D data has 2^n points, the resulting 2D image is always square). Secondly, it is fractal, which allows zoom-in and zoom-out without resolution loss. Thirdly, it is continuous, meaning that consecutive points in 1D will be consecutively mapped in 2D (never breaking the curve). And fourth, and most important, it preserves locality. The curve keeps adjacent addresses physically near each other. Subnets will be visually represented as clusters. The Hilbert curve bears strong benefits to Internet Traffic Visualization and is discussed in paragraph 3.2 as well, but in a different application.

Internet addresses are allocated in blocks of consecutive addresses. The map constructed by ISI [5] shows who controls each of the 256 numbered blocks corresponding to /8 subnets ($2^{32}/2^8$ = ~16 mil. addresses). Block number *n* represents the 16 million addresses of the form *n*.-.-.- .

As an example, Figure 5 shows a subset of the ISI Internet map. Two blocks (196/8 and 199/8) are allocated geographically, while 198/8 is used by many groups, and 197/8 is still unallocated.

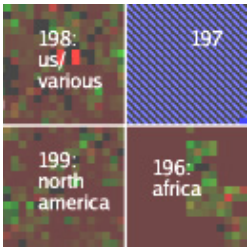

Figure 5. Subset of the Internet allocation map. [5]

Geographic allocations can reduce routing table sizes (**D5**), and round trip time (**D2**), delivering better network performance. The map provides a good overview in planning for new address allocations, by depicting the taken and available blocks. Given that it is a "one location network survey", the map also depicts the reachability of all the Internet from a specific source. If that source is an important content provider, for instance, it might be worth such an investigation.

When visualizing network information through Space-Filling Curves (3.2), in this case Hilbert Curves, color coding is usually employed to depict the data. In summary, Hilbert Curves map the Internet Address Space, and the color intensity of each pixel infers the amount/quality of data for the subnet that the pixel represents.

In this ISI visualization, each point's color coding depicts the average ICMP "echo response" of a /16 subnet (65.536 addresses). The brighter the point, the more replies were received.

3.2 Space-Filling Curves Mapping

The Hilbert Curve is a Space-Filling Curve (SFC). Figure 6 shows other SFCs, which can be employed in network data visualization. The mapping of 1D data to a 2D plane using SFCs is thoroughly discussed in [1]. In the paper a set of different SFCs are utilized to map the statistics collected to images that emphasize traffic patterns. Anomalies such as large scale DDoS (Distributed Denial of Service) attacks and scanning activities are identifiable, due to the enhanced locality of SFC clustering.

The Hilbert Curve bears one more advantage, which is explained in the paper [2]. Aggressive compression can be applied on the resulting Hilbert images, without major loss of traffic information, but with high savings in space and bandwidth, outperforming other SFCs.

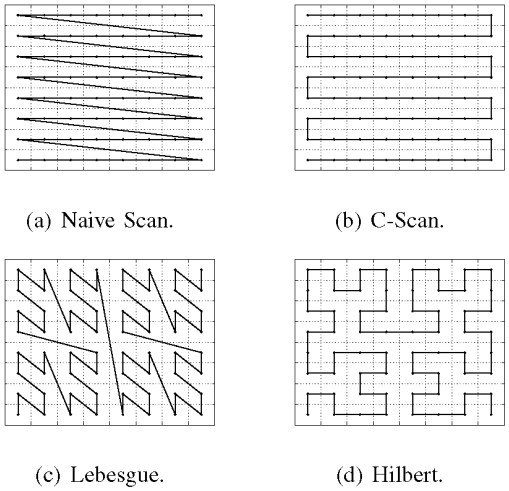

Figure 6. Space-Filling Curves. [1]

The paper [1] puts forward a visualization method for detecting anomalous network activities with the help of SFC mapping. In a distributed architecture (Figure 3), statistics based on packet header fields (**D1**) are obtained by the collection agents. The paper does not directly specify the data collection means, but it can be inferred that passive means are employed, as new traffic is not introduced into the network, but it is rather listened for in incoming packets. The data set is constituted by the number of packets arrived during an investigated time slot. Incoming packets are sorted, based on their source IP's lowest byte, into 256 classes. The paper points out that the lowest byte is the one with "the most interesting characteristics". The 256 classes are the 1D space, while the number of packets in each class will determine the color intensity.

Figure 7 shows a comparison of normal traffic and anomalous traffic, using this visualization data and method.

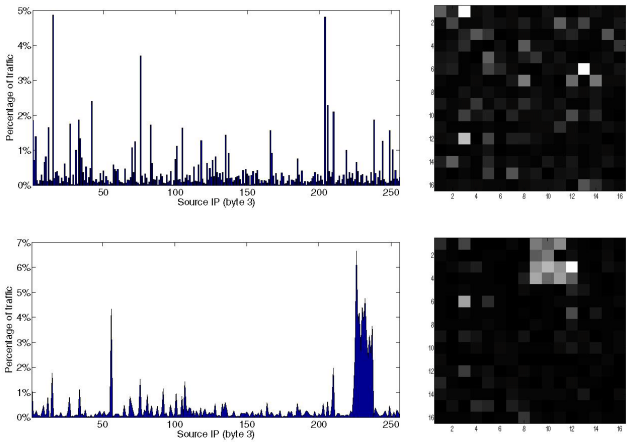

Figure 7. Sample histogram and Hilbert images of normal(top) and anomalous (bottom) traffic activity [2]

This method was chosen because of the curves' ability to preserve locality in converting traffic to intensities. The paper stresses that locality is a crucial in making attacks such as address scanning or DDoS distinguishable, "even if they are dispersed and not perfectly contiguous". This feature can be clearly observed in Figure 7, in the histogram, where most of the traffic comes from the range 225-240, as well as in the image, where due to the locality of the classes 225-240, a visual object has been formed in the cluster of classes.
Such visual objects will withstand aggressive compression, as it is pointed out in [2]. The size reductions of the images will help in processing large scale traffic, but also in communication between agents and the central analysis engine.
The visualization method can be extended to include more than one packet header characteristic. In that case, the intensity value of each pixel on the image would the number of data samples that contained the set of values <v1,v2,v3,…,vn>. In the previous example, the set of values contained, in fact, one value, the source IP of the incoming packets (<sourceIP>). Possible data sets would be <sourceIP,destIP>, or <sourcePort,destPort>, or <sourceIP,sourcePort,destIP,destPort>, etc.

The visualization procedure is the following:

1. the data set (i.e. packet IP headers) is recorded by the sensing agents in their specific locations and principal fields are extracted (source/destination addresses and ports, etc.)
2. the timeframe data is partitioned into windows (i.e. 2 min each).
3. for each window, the histogram is calculated (frequencies of each field value tuples). This is where the data is serialized (n-D to 1-D). Serialization means that different n-D tuples are numbered in a consecutive order (<v1,v2,v3> is 1, <v1,v2,v4> is 2, etc.).
4. the image is generated for each window, mapping the pixel intensities from the frequencies in the histogram using SFCs (1-D to 2-D). The serialized tuples each get a pixel on the image, which will be colored with an intensity directly related to the number of occurrences of that given tuple in the time window.

On the efficiency of SFCs in Network Traffic Visualization and anomaly detection, the paper [2] offers a discussion, based on a DDoS attack sample data set.
Analyzing the errors resulting from lossy compression of the images obtained with SFCs, it concludes that using space-filling curves, especially Hilbert mapping, has very small traffic data loss upon aggressive compression. It goes on to point out that, even after decreasing the space requirements of the image by a factor of 1024, the Hilbert mapping made the image robust enough and capable of resisting the ruining degradation of quality. The Hilbert curve has consistently the lowest error of all the mappings. Image sizes resulting from compressing Hilbert achieved smaller errors, while guaranteeing the smallest sizes.
Considering the DDoS trace analyzed in [2], the original space used to store the traffic information was ~3.5 MB, while the compressed and then 4x downscaled images required only ~1 KB (excluding the common header of the images). However, this storage saving comes at the expense of a small error value, and loss of individual packet information.
We can now see the importance of Space-Filling Curves in the visualization of network traffic information, and specifically, the strongly positive properties of Hilbert Curves. The methods discussed in this paragraph produce images which make anomalous network activities or attacks visible to the naked eye, as well as to image-processing algorithms, which can identify and process the visual objects.

3.3 Skitter

Moving away from 2D plane conversions, we will discuss Skitter data, a graph-based visualization tool and data set developed by CAIDA [8] for actively probing the Internet in order to analyze topology and performance. Its specific goals include the measuring of Forward IP Paths (**D3**), Round Trip Time (**D2**), and persistent routing changes (**D5**). It also offers a visualization of network connectivity, such as in Figure 8.
Skitter employs active means in measuring network characteristics. It is a type of "one location network survey", which records the unidirectional IP path from the source to multiple destination IPs probed. A unidirectional path is the path a packet takes from the source to the destination IP, passing through other IP devices along the way. No routing is imposed on the packets by the source, so that the network will determine the route the Skitter probes take.
Skitter accomplishes its goal in an ingenious manner. Because it would not receive responses from each IP device that the packet travels through on the way to its destination, the application is, in fact, probing each hop along the path by incrementing the time-to-live (TTL) in the IP packet header (Figure 1). This way, every hop will send a ICMP TIMEXCEED message in response to a packet with an expired TTL. The application receives the message, increments the TTL and sends it this time directly to that hop (IP device). The packet will get diverted to another hop which will send the ICMP TIMEXCEED and so on, until it reaches its destination.
The visualization Skitter uses is a graph-based one, where communication is depicted as a link between nodes. In this sample, data is plotted onto a globe in 3D, for easier handling.

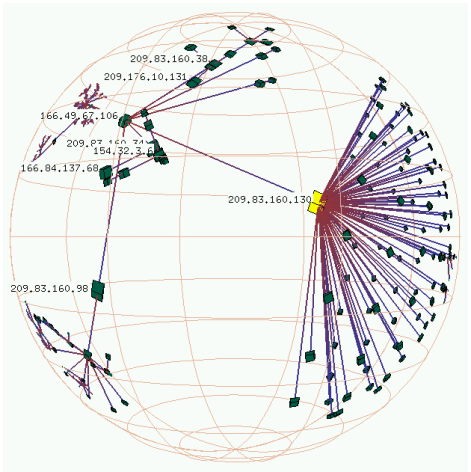

Figure 8. Sample visualization from Skitter data. [8]

The yellow square in Figure 8 is the source IP address. All other nodes are hops or destinations.
Skitter has been used to identify critical paths (i.e. routers or network nodes that might be vulnerable points), and to map dynamic changes in Internet topologies (by looking at Skitter data collected over time).

3.4 Akamai real-time Web Monitor

Finally, we take a look at geographical overlays and examples of nonconventional visualization methods.
Akamai is one of the largest CDNs (Content Delivery Network) in the world. They monitor their global private network around the clock. With this real-time data, they can identify the global

regions with the greatest attack traffic, cities with the slowest Web connections (latency), and geographic areas with the most Web traffic (traffic density).

In the case of attack traffic, Akamai collect data on the number of attempted connections, and on source and destination IPs and Ports (**D1**) of packets flowing through their network (passive measurement). Malicious activity generally comes from automated scanning trojans and viruses, searching for new computers to infect, by randomly inquiring IP addresses. The number of attacks in the last 24 hours in a region is depicted by tones of red of varying intensity covering that specific region on the map.

The network latency between most major cities is measured via automated scripts. These are tests consisting of connections, downloads, and ICMP pings (active measurement). Two latency quantifications are provided (in milliseconds), as vertical bars, for each monitored city: the absolute current latency, and the relative latency, in comparison with its historical average latency.

The third visualization, network traffic, is the amount of data being currently requested and delivered, by geography. This is perhaps the most important measurement for Akamai, as it is the basis of their revenue. The values are provided as percentages of global network traffic. Again, a color overlay is used, in relating the metric to different countries.

Figure 9. Akamai in the last 24 hours: network attacks (top), latency (middle), traffic (bottom).[7]

Although the data sets are significant, they are still limited to a private, more controlled network, and can only suggest the state of the whole Internet.

The geographical overlays and maps are used to clarify the data and present it in an appealing way. Regions are displayed as countries.

4. CONCLUSION AND OUTLOOK

Monitoring traffic activity is a necessity in ensuring the health of networks. Traffic Visualization encompasses the tools and metrics necessary for such a task. In this paper, we have broken down visualization systems and looked at types of data relevant for specific tasks, data acquisition and visualization methods, as well as real-world examples, usage scenarios, and benefits.

The current techniques, albeit effective, can only be applied on small networks for rapid information provision.

Improvements in processing power, and advances in analysis, modeling, visualization, and simulation tools, particularly those capable of addressing the scale of the Internet, will enable system administrators and network architects to plan for the next-generation Internet, a safer, more reliable and interconnected place.

5. REFERENCES

[1] T. Samak, A. El-Atawy, E. Al-Shaer, and M. Ismail. A novel visualization approach for efficient network-wide traffic monitoring. End-to-End Monitoring Techniques and Services, 2007. E2EMON apos; 07. Workshop on Volume , Issue , Yearly 21 2007-May 21 2007 Page(s): 1 - 7

[2] T. Samak, S. Ghanem, and M. Ismail. On the efficiency of using space-filling curves in network traffic representation. Computer Communications Workshops, 2008. INFOCOM. IEEE Conference on Volume , Issue , 13-18 April 2008 Page(s): 1 - 6

[3] K. Abdullah, C. Lee, G. Conti, and J. Copeland. Visualizing Network Data for Intrusion Detection. Proceedings of the 2002 IEEE Workshop on Information Assurance and Security, United States Military Academy, West Point, NY, 17-19 June 2002

[4] CAIDA: Cooperative Association for Internet Data Analysis. http://www.caida.org

[5] Information Sciences Institute (ISI) at the University of Southern Carolina: ANT censuses of the internet address space. http://www.isi.edu/ant/address/

[6] xkcd: A webcomic of romance, sarcasm, math and language. #195 Map of the internet. http://www.xkcd.com/195/

[7] Akamai Real-time Web Monitor http://www.akamai.com/html/technology/dataviz1.html

[8] CAIDA Skitter http://www.caida.org/tools/measurement/skitter/

[9] B. Irwin, N. Pilkington. High-level Internet Traffic Visualization using Hilbert Curve Mapping. VizSEC ’07 Presentation – 29 October 2007.

[10] Information Society, Traffic Measurement and Monitoring Roadmap http://www.ist-mome.org/documents/traffic_ngni.pdf