# Visual Traffic Monitoring and Evaluation

Robert F. Erbacher
Department of Computer Science, LI 67A
university at Albany-SUNY
1400 Washington Avenue
Albany, NY 12222, USA
erbacher@cs.albany.edu

## ABSTRACT

As computer networks and associated infrastructures become ever more important to the nation's commerce and communication, it is becoming exceedingly critical that these networks be managed effectively. Current techniques, which rely on manual or log based analysis, are too slow and ineffective to handle the explosive growth of network infrastructures. We have developed visualization techniques geared towards aiding the analysis of network based infrastructures such that network managers can quickly identify usage characteristics of the network and reallocate bandwidth or restructure portions of the network to better improve connectivity. In this fashion, bottlenecks can be quickly identified along with their cause so the issues can be remedied expeditiously. The techniques can also be used for long range infrastructure planning and network misuse detection.

**Keywords:** Information Visualization, Computer Networks, Traffic Monitoring

## 1. INTRODUCTION

Large scale telecommunication infrastructures provide the foundation on which e-commerce and other computing applications are based. However, monitoring the traffic on such large scale systems to determine the effectiveness of the current infrastructure is a daunting task given the complexity of current infrastructures and interactions between the systems and user based traffic. Visualization provides an effective mechanism for monitoring traffic patterns and evaluating the traffic patterns as they relate to the current infrastructure, particularly in the identification of bottlenecks, failures, and wasted resources. This becomes of particular value when rapid assessment is needed. Computational or log based mechanisms for analyzing such system can require substantial time to perform the analysis and are often used in current scenarios to determine, after the fact, what went wrong. Our visualization techniques enable the analyst, by invoking the power of the human visual system, such that the state of the infrastructure can be identified instantaneously, providing immediate feedback if problems should occur. This would then allow the analyst to identify and implement a solution before customers or other users of the infrastructure become impacted by the problem and react negatively. Such problems can occur during localized periods of exceptional loads, failure of infrastructure components such as routers, denial of service attacks, internal misuse, etc.

The visualization techniques we are developing provide a dynamic view of the localized network infrastructure in conjunction with information on remote access, system load, and network topology. We use perceptually based representations that are easily understood and allow problems to be immediately identified. This is done using a glyph based approach that allows information from multiple sources to be combined into a single visual display and automatically correlates the information. The environment, at its basic level, provides details relating to the criticality of systems, network load and capacity, local and global peaks, and temporal relationships of the bandwidth usage. At a greater level of detail the visualizations can provide details of the types of packets that are being transmitted, allowing for identification of misuse, either internal or external.

In commercial applications it is critical that customers have fast access to the organization's web sites. Should access be limited due to poor performance then the network managers must be able to identify the cause and source of the problem such that it can be corrected. This may be the result of insufficient capacity for the number of transactions, internal misuse in which employees are downloading material not relevant to their employment, or external misuse geared directly towards disrupting the day to day business of the organization.

## 2. MOTIVATION

This work is derived from our work on intrusion detection [3] and our realization that there is a great need for network monitoring tools, not only to monitor for intrusions but to monitor network usage in general. Our university provides a useful environment for this work due to its own inherent needs and inability to keep up with the activities of the students.

Analysis of log information for the university's primary UNIX server reveals that there are as many as 200 users accessing the system simultaneously. During a one week period there are more than 25000 successful connections from over 2500 different hosts. The university's primary UNIX server provides e-mail and compute resources for the university at large. However, the university's network infrastructure is insufficient to handle this connectivity load, providing 10 Mb/s connectivity with 100 Mb/s connectivity in areas with specific needs. External connectivity provides 100 Mb/s.

The issue that arises is the determination of bandwidth usage through each network and subnet. Are there subnets that are constantly saturated or exceed their bandwidth capacity? Are there subnets that achieve localized peaks that exceed capacity? If so, why and when? What subnets are under utilized? Is all utilization related to university activities? What type of traffic is generated? Would it make sense to move resources closer to the external network connection or off-site, e.g., HTTP sites? These details are critical for management of the network infrastructure. The answers to such questions will aid an organization in determining how to reallocate bandwidth, determine where upgrades are required, and identify inappropriate usage. The goal is to provide better connectivity overall for the entire organization, its employees, and customers.

Typical network usage consists of summaries in the form of Figure 1. This summary describes the network usage for a single subnet. Such textual representations can provide overviews of network usage. Fully managing the network from such summaries would require an enormous amount of time due to the difficulty in discerning relationships and meaning from text. Currently, the summary is only provided once per day. Full analysis would require the summary be provided at regular intervals, e.g., hourly or minutely. Summaries would then need to be provided for all sub network as well as the network as a whole. The volume of textual information that would subsequently be generated would make the analysis process unfeasible.

Figure 2 shows a graph generated by the network managers at the university. The graph shows overall usage of the external network connection over time. It is clear that the university's outbound network connection is completely saturated with no break. This has proven detrimental to students attempting to access the university's resources from external locations.

```
Ethernet Analysis from 2.2 minutes of observation:
Packets counted: 36000
Average pkt length: 383 bytes
Sampling periods: 144
Packets / sec    Kilobytes / sec     Protocol
-------------    ---------------     --------
173.48  64.0%      47.76  47.0%      NFS
 26.53   9.8%      19.53  19.2%      X0
 20.38   7.5%      18.41  18.1%      TCP Dest 822
 16.48   6.1%       4.48   4.4%      TCP Dest 2049
  4.78   1.8%       2.77   2.7%      http (World Wide Web)
  4.28   1.6%       2.68   2.6%      UDP Dest 2049
  3.76   1.4%       0.32   0.3%      Telnet
  2.23   0.8%       0.67   0.7%      UDP Dest 793
  1.82   0.7%       0.20   0.2%      UDP Dest 862
  1.12   0.4%       0.12   0.1%      TCP Dest 5432
  1.11   0.4%       0.15   0.1%      Domain
  0.94   0.3%       0.81   0.8%      TCP Dest 775
  0.90   0.3%       0.15   0.1%      SNMP
  0.90   0.3%       0.35   0.3%      UDP Dest 795
  0.78   0.3%       0.48   0.5%      UDP Dest 6970
  0.65   0.2%       0.04   0.0%      ARP
  0.29   0.1%       0.41   0.4%      TCP Dest 9100
 10.50   0.0%       2.24   2.2%      Other
 =====              =====             =============
270.94              101.57            Total

Packets / sec    Kilobytes / sec     Endpoint Networks
-------------    ---------------     -----------------
531.91  98.2%      99.74  98.2%      xxx.xxx.*.*
  1.58   0.0%       0.05   0.0%      Other
 =====              =====             =============
270.94              101.57            Total
```

**Figure 1:** Textual network usage summary.

However, the graph does nothing to identify the network usage internal to the university. While the graph has been used to qualify the need for upgrading the external bandwidth, what about the internal bandwidth? Additional information is needed to facilitate the development of long range plans for network infrastructure upgrades throughout the *entire* university.

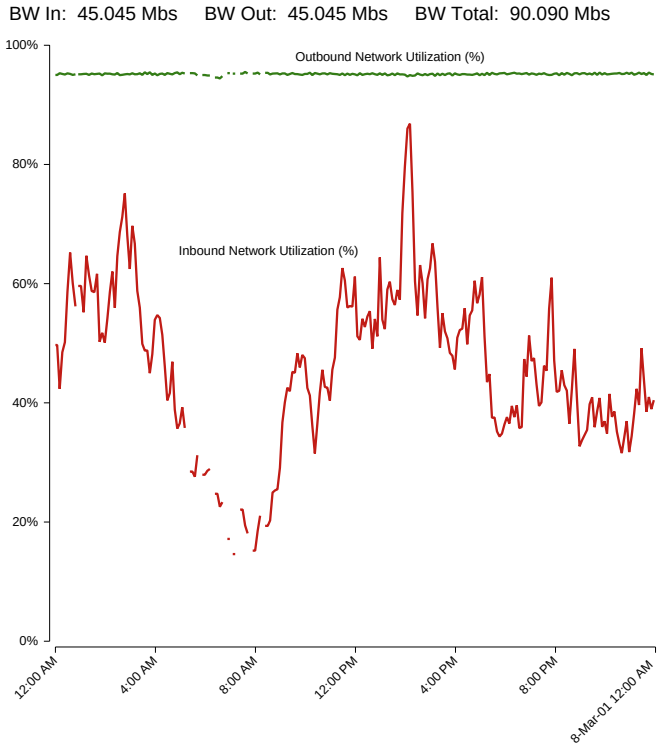


**Figure 2:** Network usage graph.

Besides basic network usage it is also critical to identify the type of data that is being transmitted through each subnet. This allows the organization to decide what types of data should be supported and aids the organization in eliminating inappropriate usage that may be overly consuming bandwidth with no benefit to the organization. It can also be used to identify out right abuses, such as the scenario in which CIA employees where misusing computers to run an illegal chat room. Network traffic monitoring would allow packets related to such activity to be identified quickly before harm is incurred.

Figure 2 does bring up another issue, namely, how usable is the university's connectivity for the student's population? Individuals who are internal to the university's subnet are able to access the university's resources without too much difficulty. However, individuals who attempt to connect from the outside will find that there are significant delays when attempting to perform even the simplest task. There can be delays of 5-10 seconds before keystrokes are echoed onto the remote terminal. As the diagram shows, this is not a matter of getting the input from the remote user to the system but rather getting the results of the input back to the user on the remote system. These limitations on student access to the university's resources led to the desire to improve bandwidth through upgrades to the network infrastructure.

## 3. PREVIOUS WORK

There has been concern over the representation of network information for quite some time, particularly when *network* is taken as a generic term. Computer networks have received not been ignored but have not received the level of attention needed to resolve all of their inherent issues. The visualizations that have been generated to assist in the task have been limited, particularly concerning the number of attributes that can be represented, making the current techniques insufficient for our needs. For example, the work by Eick et al. [8] strictly deals with e-mail and subsequently resolves many fewer nodes and attributes than is needed for our bandwidth analysis work. As an added concern we must monitor the direction of data flow as this can significantly impact our analysis. Additionally, the node placement algorithm they incorporated was based on

the weight of the link between the two nodes and does not directly reflect the network hierarchy. This can be confusing and can make it difficult to associate the visualization with the network infrastructure.

Becker et al. [9] developed the SeeNet environment which provides linkmaps for visually representing the number of connections between two network nodes. It can help identify when a node may be overloaded and shows the network's behavior. This is important when a crisis occurs and usage increases dramatically, e.g., after a California earthquake. Understanding the consequences of events, so telephone companies can be prepared for changing demands, is imperative. This technique was designed for geographical networks, i.e., telephone networks, and only shows connections between nodes. It does not represent the used or available bandwidth, the type of information transmitted, the criticality of the nodes, etc.

The work by Koutsofios et al. [6] is similar to that of Becker's except relying on the use of hierarchical graphs to represent the huge numbers of connections that are seen in telecommunications applications. Again, the application is for geographical arrangements and the number of represented attributes is limited, focusing instead only on the individual connections. In our case overall usage is our focus with the ability to pick out individual activity should it be deemed inappropriate.

Livelink [10] is an environment for visualizing and measuring the web. By probing web accesses the authors gather statistics as to the number of hits web sites are receiving. This information is presented in statistical form as charts and graphs. An extension to the environment provides a more graphical representation. In this advanced form, the approximate location of network nodes is represented, showing geographical association between web sites. The visual representation of each node is then presented in such a way as to reveal the activity of the site. Each node can represent several parameters simultaneously. In this example, the representations only include system statistics. Attributes of the network itself are not incorporated.

Finally, the work by Estrin et al. [7] is designed to visually simulate individual packets to aid in the development of protocols. It does not deal with visualization of the network at large and the issues developed once a protocol is deployed and the effectiveness of the network must be measured overall. The techniques do incorporate more realistic node placement, though on a very small scale. With the advancement of graphing theory and algorithms [11] developing appropriate node placement algorithms should prove feasible for our environment as well, even when dealing with huge numbers of nodes.

## 4. DATA COLLECTION

The majority of the data collection is done by the router. The router can easily keep track of the inbound and outbound traffic for a given subnet and even for individual workstations on that subnet. It is also perfectly positioned to forward that data to the network manager. Additional data collection is usually necessary within the subnet itself to collect intra subnet traffic. The majority of this data is traffic bound for any servers on the subnet. Consequently, we collected data from any servers on the subnet as well. This provided us with a good view of the data traffic on the subnet. There will still be some missed data resulting from traffic targeted between machines on the subnet but the picture we will get will be sufficient for analysis purposes as the amount of targeted traffic between workstations should be minimal. This data collection provides the environment with the bandwidth usage characteristics. The latency for individual connections is collected using traceroute and ping. This information can only be collected from system to which we have access.

When attempting to identify the network organization to be represented visually we found that we will need to provide some site specific information to the environment in order for it to determine the organization of the network infrastructure. In particular, we provided a configuration file identifying routers on the network and the IP addresses associated with them. We also included primary servers in this list for simplicity. The remaining organization of the network can be inferred from the traceroute information and IP addresses. Systems identified to be on the same subnet by their IP addresses and having the same number of hops to a given router are assumed to be switched to that router. We cannot directly derive details of switches but they should not hinder traffic flow. The environment can determine if a switch is overloaded by examining the inbound and outbound traffic of all systems connected to that switch.

The process of automatically identifying aspects of the network organization for workstations assists the network manager in determining if systems are being illegally connected to the network by monitoring the destination address of packets on the network and flagging systems that have not been properly added to the NIS database.

## 5. VISUALIZATION TECHNIQUES

Our goal when developing the visualization techniques was to incorporate perceptual issues [4, 5] such that the visualizations could be easily perceived and interpreted, particularly when there is a large amount of information being animated over time and interesting events must be pulled from the visualization quickly. Additionally, the appropriate presentation of the information [1, 2] is critical. The effectiveness of the presentation is highly dependent on how the individual will perceive the visualization but also on any preconceived notions the user may have. Ensuring that our visualization environment does not falsely represent information according to the preconceived notions of our expected user base is critical.

The visualization techniques are based on a glyph metaphor in which attributes of the glyph are used to represent characteristics of the bandwidth usage needed by the network manager. Currently, the data parameter to visual attribute mappings of the glyph are fixed though in the future we expect to make them fully customizable. The glyphs used in the environment are shown in figure 3. Figure 3a shows the basic glyph providing the network bandwidth used by a node in both the outbound and inbound directions from that node. As the graph in figure 2 showed, it is critical that we differentiate between inbound and outbound traffic.

(a) (b)

**Figure 3:** Basic glyph organization

The directed line shows the direction of the data flow being measured. The black borders show the *full* capacity of the network connection, with the border being thicker for 100 Mb/s connections than for 10 Mb/s connections, and the intensity of the gray interior shows the percentage of currently used bandwidth. The hashes along the directed line are representative of the type of data being transmitted, if available, and the quantity of that type of data. The hashes represent the following data types in order, starting from the node itself:

- Ping average. Representation of the average ping for all users. Should the system become bogged down and the ping time begin to increase dramatically the visual representation will be obvious. Using the average also eliminates outliers, i.e., individuals whose connection is very poor on their end or are very remote.
- NFS data. Since much of our data is internal to individual departments there is generally an enormous amount of this type of data.
- Telnet
- HTTP
- TCP
- UDP
- SNMP

Notice that the currently supported data types are those already available from the available network monitoring tools, as seen in figure 1. The tools should be easily extensible to incorporate additional details, such as IRC data, etc. Also notice that we are concerned with total bandwidth usage over each link and do not concern ourselves with the ultimate destination of that data. As with maximum flow type graphs we are concerned with usage or over usage of individual links and not with links that the flow will contribute to down the line.

Figure 3b shows a glyph used to represent additional information, if available. This is particularly relevant for the university's principal server, where we can collect such information. The lines extending from the perimeter of the glyph are

representative of the number of users on the system. Each line represents ten users; due to the number of users on the system consistently representing the users individually is unfeasible. The thick center ring is representative of the current system load. These details are useful in determining how critical the network bandwidth to a particular system is. Given a minimally active system, having a saturated network is not as critical as having a saturated network feeding a heavily used network. One final representation that we use is a single thick circle to represent known routers as the routers in general should not have users accessing them directly.

Figure 4a shows an example of the environment in operation. Normally the visualization would be animated, showing changes in bandwidth usage over time. The accuracy of the representation is completely dependent on the network monitoring intervals. As mentioned previously, the textual information is collected daily. This leaves huge gaps during the day when monitoring is not performed. This opens up the possibility that significant changes in the behavior of network usage is missed. Essentially what is seen in such sporadic monitoring is the normal background noise of the network, ignoring peaks that will likely occur throughout the day or when more users attempt to access the system.

In order to assist with determination of the time of day the full display includes a border around the screen. The intensity (gray level) of this border is white at noon and black at midnight. An additional yellow border is displayed in association with PM. This assists in determining if the intensity is increasing or decreasing at any particular point in time. The example in Fig 4. has a very light border without a yellow border so it is clearly approaching noon at the point this image was taken.

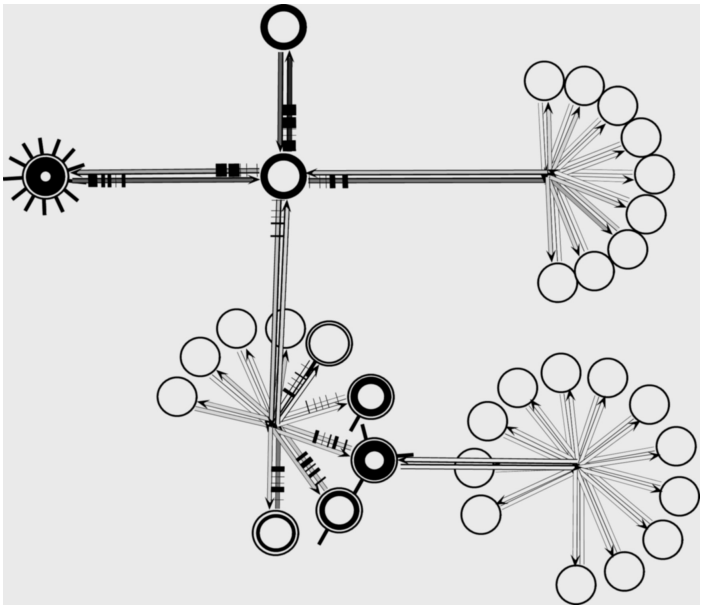

**Figure 4:** Visualization examples.

In the diagram, the larger nodes are the principal servers for which additional information is available. The smaller nodes are workstations and remote users for which such information is not available. This diagram shows activity in the CS department, the main university server, the main university laboratory, and the university's primary routers.

The example clearly shows the clustering expected from a switched network. Each cluster is centered on a switch or groups of switches. The cluster in the bottom right represents a student lab of xterms for which we did not collect data. Those systems are run off of one of the servers in the CS department. This server can be identified by following the links from the center of the xterm cluster. The cluster in the middle of the bottom is the main CS department cluster, which consists of several servers and workstations. The cluster on the right is the main university computer laboratory. Note that the since this data was collected outside of a normal semester the number of users in the labs is relatively small. The number of users on the

university's main server, left most node, is always extremely high since it also serves employees. Finally, in the center can be seen the two main routers for the university.

## 6. USER INTERFACE

The user interface, Figure 5, provides basic capabilities for controlling the environment. Along with a typical menu, the user interface contains convenient buttons for taking snapshots and animations of the environment. This is critical if the network manager needs to capture information to validate reorganization or upgrade proposals. The exact time of day is also provided in digital form in the user interface. A representative of the time of day is needed within the visual display to maintain the user's focus on that single display so that activity is not missed. The remainder of the interface is occupied by VCR like controls that allow the user to control the execution rate of the simulation; speeding it up, slowing it down, pausing it, and rewinding it as necessary.

The environment was developed in C++ with OpenGL and GLUT. The interface was developed using Tcl/Tk. The scripts were written in bash. The implementation is completely system independent While the visualization was actually done on a PC the use of the cygwin environment eliminated the need for any system specific code. The application can therefore be ported directly to any environment without modification.

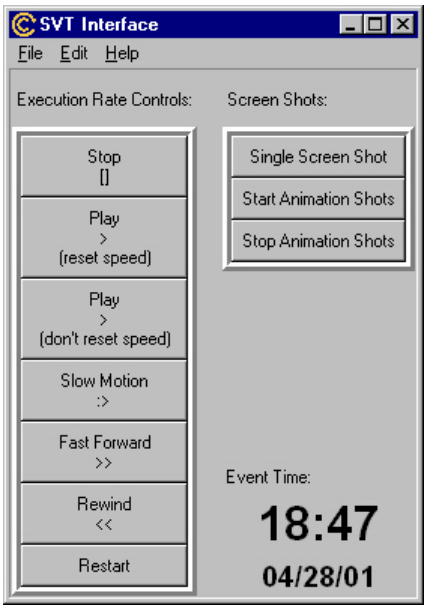


**Figure 5:** User interface for the visualization environment.

## 7. FUTURE WORK

The environment could be greatly improved in many respects. The visualization itself should incorporate user selectable levels which force links to become highlighted in red when their usage exceeds a certain level of the capacity, perhaps 95%. This visual key would greatly assist in identification peaks and insufficient bandwidth.

The interaction provided with the environment needs to be enhanced greatly. The ability to probe the environment so that when the user selects a node details of that node are shown would greatly improve the usability of the environment. We envision the probing capability providing details about the system, specifics of the network capacity and bandwidth used, as well as more detailed graphics statistics as to how the network is being used by that system. This would require participation by individual departments to fully provide the needed data but should be feasible. The visualization merely shows the average ping time for all users accessing the system. This provides an overview of network connectivity. The network manager will need the ability to select a system and quickly identify the ping times for all users and identify whether or not they are local.

Work needs to be done to analyze the impact of the monitoring agents on the performance of the systems they are running on as well as on the network itself. Clearly, our tools will be having some impact, we have yet to measure that impact.

## 8. CONCLUSIONS

By visually representing the statistical information associated with network usage and directly associating that information with the network layout our environment greatly enhances the ability for a network manager to asses the effectiveness of the network infrastructure and plan long range infrastructure management as well as deal with short term and immediate crisis. The additional details the environment is capable of providing ensures that the network bandwidth is being put to the best use and provides the details needed to locate and eliminate waste and misuse should they occur. The visualization techniques make the network usage and limitations clear beyond a doubt, providing validation for requested upgrades.

## REFERENCES


1. Edward R, Tufte, *The Visual Display of Quantitative Information*, Graphics Press, 1983.
2. Edward R. Tufte, *Visual Explanations*, Graphics Press, 1997.
3. Robert F. Erbacher and Deborah Frincke, "Visualization in Detection of Intrusions and Misuse in Large Scale Networks ," *Proceedings of the International Conference on Information Visualization '2000*, London, UK, July, 2000, pp. 294-299.
4. Markus Gross, *Visual Computing, The Integration of Computer Graphics, Visual Perception and Imaging*, Springer-Verlag, 1994.
5. William R. Hendee and Peter N.T. Wells, *The Perception of Visual information*, Springer-Verlag, 1994.
6. Eleftherios E. Koutsofios, Stephen C. North, Russel Truscott, and Daniel A. Keim, "Visualizing Large-Scale Telecommunication Networks and Services," *Proceedings of the IEEE Visualization '99 Conference*, IEEE Computer Society Press, San Francisco, CA, pp. 457-461, 1999.
7. Deborah Estrin, Mark Handley, John Heidermann, Steven McCanne, Ya Xu, and Haobo Yu, "Network Visualization with Nam, the VINT Network Animator," *IEEE Computer*, Vol. 33, No. 11, pp. 63-68, November 2000.
8. Stephen G. Eick and Graham J. Wills, "Navigating Large Networks with Heirarchies," *In Visualization '93 Conference Proceedings*, San Jose, California, pp. 204-210, October 1993.
9. Richard Becker, Stephen Eick, and Allan Wilks, "Visualizing Network Data," *Readings in Information Visualization: Using Vision To Think*, Stuard Card, Jock D. Mackinlay, and Ben Shneiderman, editors, Morgan Kaufman Publishers, pp. 215-227, 1999.
10. Tim Bray, "Measuring the Web," *Readings in Information Visualization: Using Vision To Think*, Stuard Card, Jock D. Mackinlay, and Ben Shneiderman, editors, Morgan Kaufman Publishers, pp. 469-492, 1999.
11. Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*, Prentice-Hall, 1999.