Joint EUROGRAPHICS - IEEE TCVG Symposium on Visualization (2002), pp. 1–8
D. Ebert, P. Brunet, I. Navazo (Editors)

A Visual Technique for Internet Anomaly Detection

Soon Tee Teoh* Kwan-Liu Ma* Xiaoliang Zhao† S. Felix Wu*

*Department of Computer Science, University of California, Davis
†Department of Computer Science, North Carolina State University

Abstract

The Internet can be made more secure and efficient with effective anomaly detection. In this paper, we describe a visual method for anomaly detection using archived Border Gateway Protocol (BGP) data. A special encoding of IP addresses built into an interactive visual interface design allows a user to quickly detect Origin AS changes by browsing through 2D visual representation of selected aspects of the BGP data. We demonstrate that each visually spotted anomaly agrees with actual anomaly on record. It is clear that this visual approach can play a major role in an anomaly detection system.

1. Introduction

The Internet has become indispensable to the functioning of individuals and organizations, including government, businesses, schools, and even emergency services. However, the very nature of the Internet, which relies on interconnectedness and autonomy, makes it prone to unintentional machine or human errors as well as malicious attacks. It is therefore of utmost importance to learn about and understand these harmful events. Monitoring the Internet to recognize anomaly allows us to gain valuable understanding about the Internet so that we can take appropriate action in a timely manner.

In computer network security, anomaly detection is the process of searching for behavior deviating from normal network use. Most existing anomaly detection methods are based on statistical analysis, where user normal profiles are expressed as sets of statistical measures [8, 11, 12]. That is, a set of "normal" data is first analyzed to derive representative characteristics of normal use, which are then compared against the characteristics of unknown data to disclose abnormal behaviors. This comparative analysis forms the basis of anomaly detection.

In this paper, we describe a visual-based approach to the anomaly detection problem. Our approach does not need a "normal" data set and mainly relies on the superior visual processing capability of the human brain to detect patterns and draw inference. Starting with no prior knowledge of what shape or form the anomalies take, we use visualization as the key tool for discovering the intrinsic properties of normal and abnormal data.

We have developed a visual representation along with a set of interaction techniques for the user to visually browse through archived Border Gateway Protocol (BGP) [13] data to quickly detect anomaly in Origin AS changes [17]. These changes can indicate either configuration errors or intentional attacks of the Internet.

Section 2 introduces BGP, Origin AS changes, and their implications to Internet security. Section 3 describes in detail the visual-based anomaly detection method. Finally, we report our findings and the lessons learned from the visual analysis of archived BGP data over 480 days.

2. BGP Data and Origin AS Changes

The Internet is a network of networks. Each network within the Internet is identified by its IP address prefix. For example, the University of California's (UC) Davis campus network is identified as 128.120.0.0/16, which means every host in the UC Davis campus network shares the same first 16 bits. One or more networks within a single administrative domain is referred to as an Autonomous System, or AS for short. Each AS is assigned a unique AS number. For example, the AS number for the UC Davis campus network is 6192. Informally, we could say AS 6192 owns the IP prefix 128.120.0.0/16.

Each AS connects with one or more other ASes. Between

two ASes, inter-AS routing protocols are used to exchange network reachability information so that eventually routers know how to forward data packets to the correct destination. Border Gateway Protocol (BGP) [13] is the current standard inter-AS routing protocol. BGP routers exchange the network reachability information in the format of BGP routes. A BGP route lists a particular IP prefix (destination) and the path of ASes used to reach that prefix. The last AS in an AS path is referred as the Origin AS of that prefix. For example, the BGP route "128.120.0.0/16: (6079,11423,6192)" means that the IP prefix 128.120.0.0/16 could be reached by first going to AS 6079, then to AS 11423, and finally to AS 6192. AS 6192 is the Origin AS of the IP prefix 128.120.0.0/16.

Apparently, the Origin AS should be the owner of the IP prefix. Thus, the Origin AS for a particular prefix should remain same all the time unless the ownership changes. However, due to some faults like router misconfiguration or intentional attacks, we may observe abnormal Origin AS changes through the BGP routing table, which contains all the recent BGP routes. For example, AS 6192 originates the IP prefix 128.120.0.0/16 all the time, except, on one day, we observed that a different AS started to originate the same IP prefix too. We could ask if it is due to valid network operation or due to an attack. In the latter case, the routing system could be adversely affected and data packets could be delivered to the wrong place.

We obtained the archived daily BGP routing data over 480 days from the Oregon Route Views server [1]. Then we collect all the changes to the Origin AS of an IP prefix. We believe that examining these Origin AS changes exposes router errors and attacks.

3. A Visual-Based Approach

Traditional statistical anomaly detection methods search for patterns by using primarily automatic mechanisms. In contrast, a visual anomaly detection method is based on interactive data exploration. Goldstein et al. [6] describe data exploration as an iterative and interactive process initiated and directed by people. Previous efforts in visual techniques to aid data mining [7] include [4], [9] and a method based on clustering [14]. Girardin [5] uses self-organizing maps to to help analyze network activity. Atkison et al. [3] propose detecting network intrusion by running data through an information retrieval system and visualizing the result.

There are three goals of our visualization system. The most important one is for the user to be able to quickly identify anomaly in the data. However, it is not enough merely to discover that an anomaly has occurred. Therefore, two additional goals are to enable the user to quickly understand the nature of the anomaly and to identify its source. This is so that the user can know where to focus further investigation and take corrective action. With appropriate visual metaphors, these two additional goals can be more easily achieved than with automatic, non-visual techniques. This key advantage of data exploration over data mining is mentioned in [6].

Ahlberg and Shneiderman [2] promotes visual-based methods as a viable approach to information-seeking due to the ability of humans to recognize features in visual displays and recall related images to identify anomalies. Girardin [5] states that human perception can notice even features which are not expected. This is especially important when the user has no idea in advance about the characteristics of normal and abnormal behavior.

Lee [10] states that a shortcoming of statistical anomaly detection methods is that normal behavior changes over time, and the detection system has difficulty adapting to the change. In the visual method, the human user is more able to recognize gradual, normal changes in behavior, and distinguish that from genuine anomalies.

In traditional statistical methods, it is a challenge to set threshold values such that false positives are minimized while not missing true positives. With the visual approach, we relegate the responsibility of making fuzzy judgment of what is normal/abnormal to the user [5]. Furthermore, the user can judge whether a detected anomaly is important or is just an isolated case, whereas an automatic method would just raise flags based on a rigid set of criteria.

3.1. An interactive visualization process

Anomaly detection by visual data exploration consists of 3 steps.

1. data are collected and filtered.
2. data are mapped to appropriate visual properties.
3. the user interacts with the data, possibly going back to 2.

The visual anomaly detection method is an iterative process. The anomaly detection method has to be performed with different parameters in order to achieve success. Interactive visualization provides an efficient means of trying out different combinations of variables to watch, as well as different mappings from data to visual properties. With interactive visualization, the human user can very easily guide the iterative process in the most promising direction.

It is crucial to provide the user with the tools to interactively change parameters, focus on certain details, and animate the data over time. Interactivity allows consecutive image frames to give the user a coherent mental picture. Our design of the user interface adheres to two main principles given in [15]:

1. rapid, incremental and reversible actions, and
2. immediate and continuous display of results.

These guidelines facilitate intelligent and productive human interaction for anomaly detection. In order to achieve interactive display rates despite the large size of the data, we

need to use efficient data structures. We also have to provide the means for viewing at different levels of detail.

4. Visualizing Origin AS Changes

In this section, we describe in detail the design of our visual anomaly detection system for analyzing Origin AS changes. An Origin AS Change is an entry in the form (*Prefix,AS,Date,Type*). *Prefix* is the IP prefix whose Origin AS has changed. *AS* is a list of the associated AS(es) of the change. *Date* is the date on which the change occurred. *Type* is the type of the change.

4.1. Types of Origin AS changes

Origin AS changes are classified into 4 main types and then further classified into 8 types in total. The 4 main types are:

1. B-type: An AS announces a more specific prefix out of a larger block it already owns
2. H-type: An AS announces a more specific prefix out of a larger block belonging to another AS
3. C-type: An AS announces a prefix previously owned by another AS
4. O-type: An AS announces a prefix previously not owned (and therefore owned by ICANN by default)

A Multiple Origin AS (MOAS) conflict occurs when it appears as though an IP prefix originates from more than one AS. MOAS conflicts could be a symptom of a fault or an attack [17]. The C-type and O-type changes are further classified by whether they involve Single Origin AS (SOAS) or MOAS:

1. CSM: C-type change from SOAS to MOAS
2. CSS: C-type change from SOAS to SOAS
3. CMS: C-type change from MOAS to MOAS
4. CMM: C-type change from MOAS to MOAS
5. OS: O-type change involving SOAS
6. OM: O-type change involving MOAS

The 8 types are thus these six and the B-type and H-type changes.

4.2. Mapping IP prefixes

Each IP prefix maps to one pixel on a square. The mapping is done in a traditional quad-tree manner. Figure 1 shows this mapping. In a quad-tree, a square is repeatedly subdivided into 4 equal squares. In mapping a 32-bit prefix to a square, we use start with the first two most significant bits of the address to place the IP address in one of the 4 squares in the second level of the quad-tree. We then use the next two most significant bits to place the IP prefix in the appropriate third level square within this square. We do this repeatedly until we can place the prefix in a square the size of a single pixel. The prefix is mapped to that pixel.

Due to the limitations of a computer screen space, we use

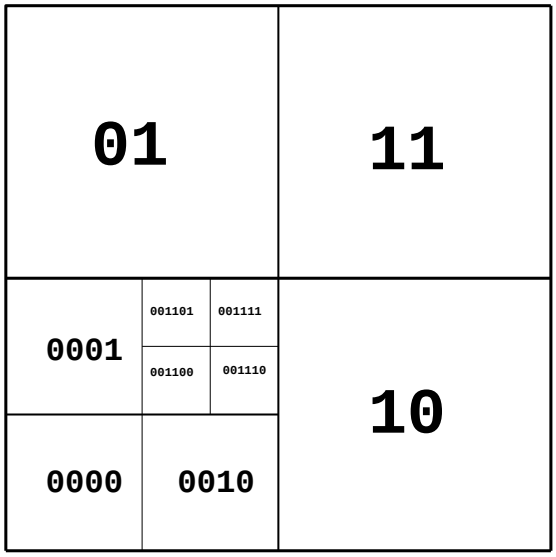

Figure 1: *Quadtree coding of IP prefixes, show top few levels of the tree, and the most significant bits of the IP prefixes represented by each sub-tree (sub-square).*

a 512 X 512 pixel square to represent the entire 32-bit IP prefix space. With only 512 X 512 pixels, many IP prefixes map to the same pixel. Despite that, a 512 X 512 square is sufficient in spreading out the IP addresses in our data. With an additional level of zooming into a portion of the data, we can view individual IP prefixes. Figure 2 shows additional windows zooming into the main window showing the entire IP prefix space. In the main window, a pixel is colored yellow if an Origin AS Change occurred on the current day, and colored brown if a change occurred on a previous day. In the detail windows, a colored square is shown for each Origin AS change. The position is determined by the IP prefix, the size by the mask, and the hue by the type of the change. Each of the 8 different possible types of Origin AS change is mapped to one unique hue. The brightness of each square depends on the day the change occurred, with the current day's data being the brightest. This example shows the data over a 416-day window from January 1, 2000 till February 19, 2001. To show only one day's data, the user can set the window to one day.

This is a sensible mapping from IP prefix to screen space because IP prefixes sharing similar more significant bits would be in close proximity on the screen. In the detail windows, each IP prefix is shown as a square or a rectangle. The size of the rectangle indicates the size of the block of IP addresses; prefixes with a smaller mask get mapped to larger rectangles.

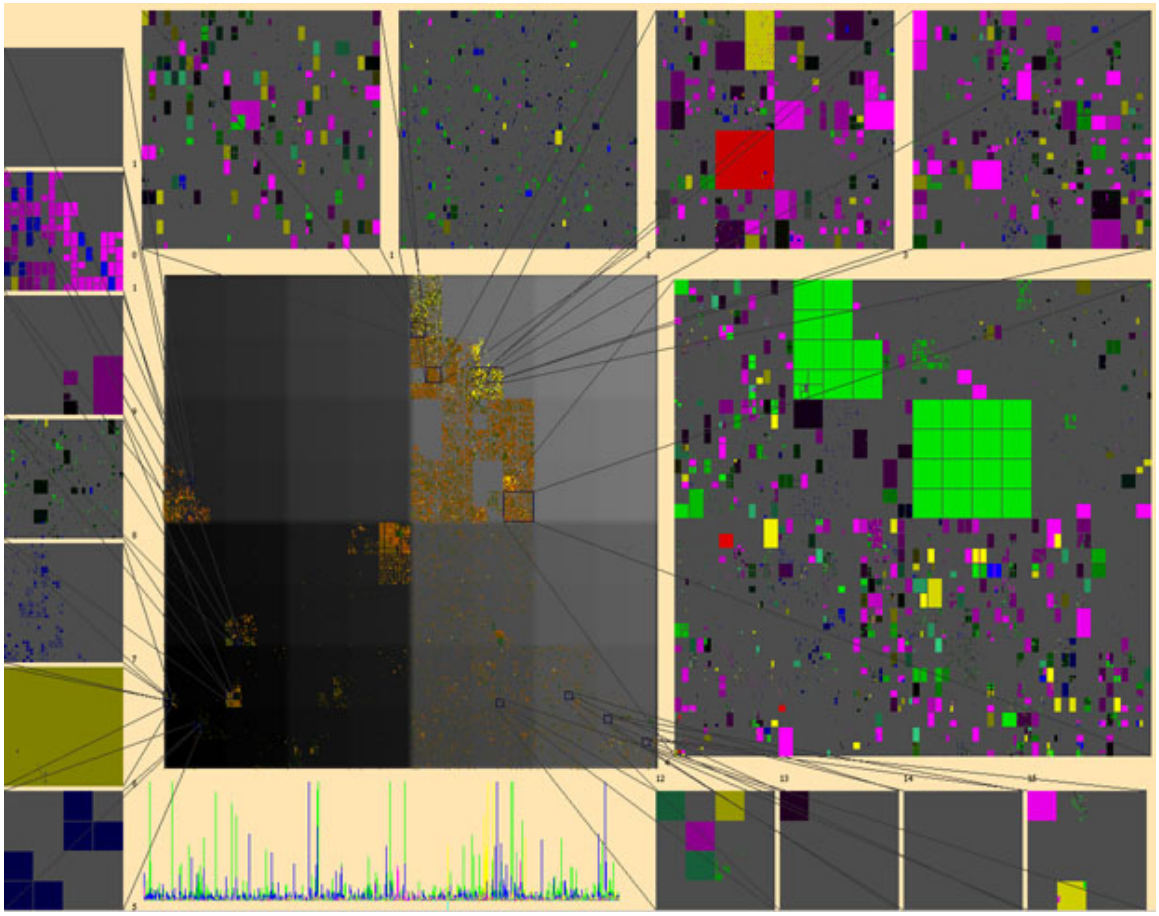

Figure 2: *Visualization of data for 416 days up till February 19, 2001. The main window shows the quadtree mapping of the entire space of 32-bit IP address. A pixel is colored yellow if an Origin AS Change occurred on the current day (February 19, 2001), and colored brown to green if a change occurred on a previous day (January 1, 2000 through February 18, 2001). In the windows showing detail, a square is used to depict each change, with hue determined by the type of the change, brightness determined by how long ago the change occurred (present day data shown the brightest), and size determined by the mask of the prefix. The background of the main window is shaded according to the IP prefix the pixel represents. The brighter the pixel, the larger the IP prefix represented.*

4.3. Relationship between prefix and AS

Next, the relationship between a prefix and its associated AS number needs to be represented. To achieve this, we draw 4 lines surrounding the IP square. An AS number is mapped to a pixel on one of the 4 lines. We draw a line from an IP address to an AS number if there is an Origin AS change involving that IP address and that AS number. This mapping takes advantage of the user's acute ability to recognize position, orientation and length. Figure 3 shows the visualization of the IP-AS relationship of Origin AS Changes of a typical day. Once again, the color of each line is based on the type of change it represents.

Since there are more AS numbers than pixels, more than one AS number maps to a pixel. Again, we provide zooming features for the user to differentiate between AS numbers which map to the same pixel in the main display. The lines representing changes for the AS in focus is shown with brighter and more saturated colors than other changes. This effectively highlights the AS, fading the other changes into the background. This is shown in Figure 4, where the pink (OS-type) lines emanating from one AS are highlighted among thousands of lines.

4.4. Animation and other features

For the time dimension, we show one day's data at a time, and allow the user to animate the visualization, each frame showing consecutive day's data. With this "movie" display,

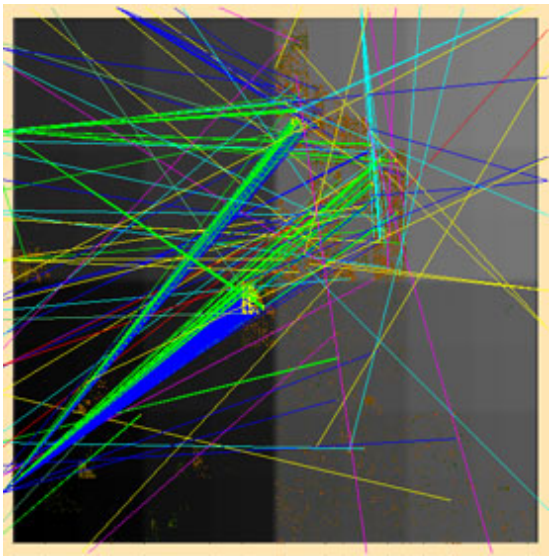

Figure 3: *Data on a typical day (September 24, 2000). For each change, a line is drawn between the IP prefix and the AS involved. Each line is colored according to the type of the change. On this day, there are many H-type (blue) and B-type (green) changes originating from a single AS to a few blocks of IP addresses.*

the user can detect temporal patterns. To assist our memory of patterns from previous days, we allow a user-defined window of a certain number of days prior to the currently shown date. Data from these previous days are displayed, but with darker, less saturated colors, so that the current day's data stands out.

For the convenience of the user, we also provide textual display of the IP address or AS number represented by the pixel clicked by the user. Other features for convenience include a slider bar to tell the date of the current data shown, and also to allow the user to choose the date to show. With the time line is a simple plot of the total number of changes of each type on each day. The plot is in the lower left of Figure 2. The current date is also displayed in text. The user can also change the date shown by typing the desired date.

By choosing parameters like what IP prefixes to zoom in on, which AS numbers to focus on, which type of changes to view etc., the user can view vastly different information. Depending on the combination of chosen parameters, the user can see the overall pattern of the data, or the user can focus attention on very specific parts of the data. Different choices would reveal different anomalies and information.

Figure 2 shows the Origin AS Changes accumulated over 416 days (from January 1, 2000 to February 19, 2001). We observe that the Changes occurred in localized areas. An area on the square corresponds to a block of IP addresses sharing the same prefix. It is also observed that different areas have different characteristics. For example, Changes on the lower right (128.0.0.0/8) tend occur in larger blocks (16-bit masks)

4.5. Anomalies detected

To validate the visualization approach for anomaly detection, we had a couple BGP experts use our tool to detect potential problems (faults and attacks) since January 1, 2000. Both of them agree that our visualization tool provides a much improved interface than the tool they were using previously, and is helpful in debugging the network.

We classify the detected anomalies into three different catagories: measure intensity (the number of MOAS conflicts we observed, regardless of MOAS types), AS anomaly (unusual behavior per AS), and animation correlation (special correlation relations across the time domain).

4.5.1. Measure Intensity

Normally, the amount of MOAS conflicts in the Internet is limited. When some serious faults/problems happened, the number of dots (in the 3D figure) or colored lines (in the AS view) would increase significantly. While it is possible that some ISPs had some dramatic network topology (or configuration) changes in one single day, it is very valuable to monitor the health of the network through this measure.

For instance, on September 18, 2001, while the Nimda/CodeRed-II worms are spread around the Internet, we can clearly observe a surge on the intensity measure for MOAS conflicts. Furthermore, since the attack was widespread around the whole Internet, we can observe that many ASes simultaneously have contributed the problems.

On another instance, on June 14, 2000, many MOAS conflicts appear in the picture. After careful analysis, 40% of the CMS conflicts are caused by AS 1591 and 35% for the prefixes 204.208.x.x.

4.5.2. AS Anomaly

One very useful feature of our visualization tool is the capability to identify a small number of problematic ASes because most of the practical BGP problems today only involve one or two ASes.

In Figure 4, the entire square is covered with blue lines (H-type changes). In addition, some pink lines (OS-type changes) emanating from a single AS are very noticeable. This is in contrast with the more common observation of a H-type changes involving close IP addresses and a single AS, for example in Figure 3. From the picture, we easily discover this anomaly since it is highly unusual that so many H-type changes occurred on one day involving so many different ASes. It turns out that AS 7777 misconfigured their

routers, announcing many prefixes, including many with 32-bit masks, which is not supposed to happen. In Figure 4, H-type lines are only drawn from the IP prefix to their previous Origin ASes and not to their new Origin AS, which is AS 7777. However, OS-type lines are drawn to their new Origin AS, which is AS 7777, since their previous Origin AS is null (see Section 4.1). This example shows that although the picture may have many lines crossing and obscuring each other, anomaly can still be detected. To overcome the clutter to get specific information regarding an individual or a group of IP prefixes and ASes, the user can select those prefixes or ASes to focus on, as mentioned in Sections 4.2 and 4.3. Other ways to avoid visual clutter are discussed in Section 4.6.

In Figure 5, yet another example appears on January 20, 2001, where we observed, through the tool, that AS-8708 falsely (most likely misconfiguration) announced 29 /32 prefixes. While, on the same day, AS-6463 injected 41 CSS conflicts against AS-15290. The later might be normal though because AS-6463 belongs to AT&T Canada Telecom Services, while AS-15290 belongs to AT&T Canada IES. But, it is interesting that our tool shows a potential topology change within the same service provider.

4.5.3. Animation Correlation

The most interesting aspect of our tool is to discover "correlation" relations via the animation of the BGP data sets. Figure 6 shows a large number of changes due to AS 15412 erroneously announcing prefixes belonging to many different ASes on April 18, 2001. The next day, changes were made to correct the error, shown by Figure 7. Although Figures 6 and 7 look disorderly, an identical pattern is easily observed because the changes involved the exact same prefixes and ASes, once again demonstrating the effectiveness of human pattern recognition. In fact, this storm of on and off CSM and CMS problems have occurred since April 6, 2001. The animation helps the system administrators to discover not only a problem has occurred but also how one type of MOAS conflict affects another type.

Other anomalies observed include private AS number leakage on September 18, 2000, and many days with high type-O activity. We have not found explanations for many of these observations. With more investigation, and further exploration with the visualization tool, we will be able to find out why these changes occurred.

4.6. Alternative representation

Another way to overcome visual clutter is presented in Figure 8. It shows an alternative representation of the data for August 14, 2000, (original representation in Figure 4). In this representation, each Origin AS change is mapped to a point on a horizontal plane in the same quad-tree manner we described. The vertical position of the horizontal plane is based on its associated AS number. Each change is shown as a cube in the position. The cube is colored according to the type of change, as before. Once again, anomaly is revealed because many different ASes are involved. In this mapping, there are no lines crossings. However, from our experience, the original 2D representation is still better at showing certain features, for example the same AS originating many far-apart IP prefixes. The user can navigate through this 3D representation by operations such as rotation, translation and zoom/pan.

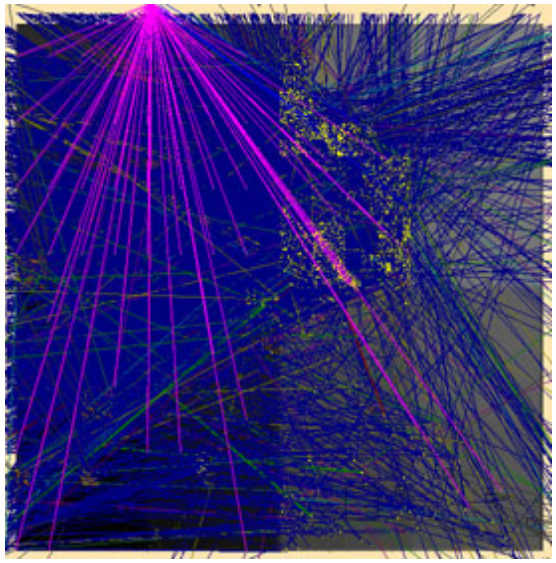

Figure 4: *Data on August 14, 2000. An anomaly is observed despite visual clutter. Many B-type changes involving different ASes and IP prefixes occurred. Some OS-type (pink) changes are highlighted. These OS-type changes all involve AS 7777 and far-apart IP prefixes. This also indicates a fault.*

Projecting the cubes onto two perpendicular vertical planes gives us yet another alternative visual representation of the data. The projected images of each day's data can reveal patterns of anomaly. In Figure 8, the cubes are projected to grayscale values onto two planes in the background. Figure 9 shows the result of projecting the cubes in Figure 8 onto squares colored by change type. The Figure shows only the projected images and not the cubes themselves. The anomalous pattern of regularly-spaced pink (OS-type) squares is especially obvious.

5. Conclusions and future work

We have demonstrated the principles and effectiveness of using visualization as a tool for anomaly detection, and for revealing the source and nature of the detected anomalies. We believe that visual-based approach will be widely adopted, improving the security and efficiency of the Internet.

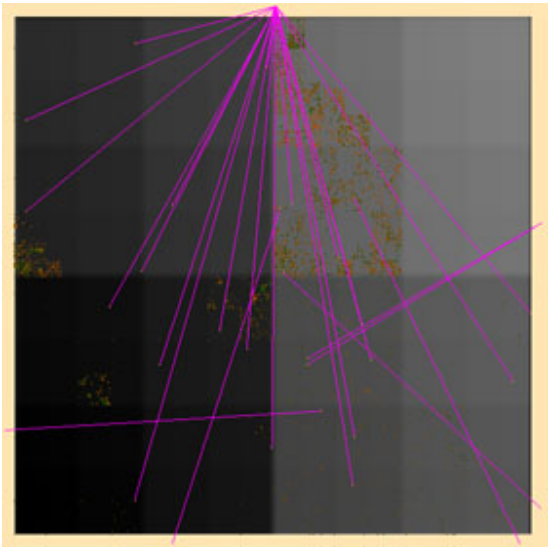

Figure 5: *OS-type changes on January 20, 2001. Many involve AS 8708 and IP prefixes with 32-bit masks.*

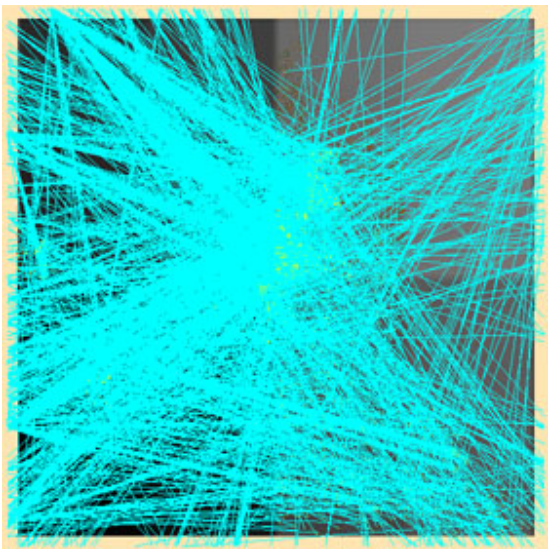

Figure 6: *CSM-type changes on April 18, 2001.*

One limitation of the current approach is that it is not easy for the user to quickly find out which ASes cause frequent changes over non-adjacent days. It is also not easy to quickly notice which AS-IP pairs occur frequently, or occur in a periodic manner. More data preprocessing incorporating statistical methods could help identify these phenomena and highlight these ASes, IP prefixes or AS-IP pairs to draw the user's attention during interactive visualization.

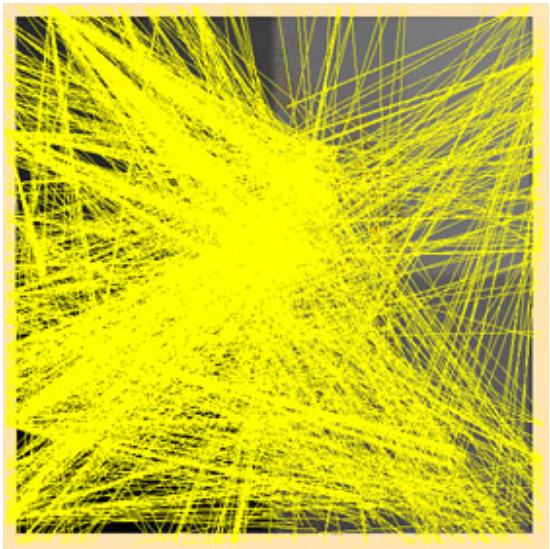

Figure 7: *CMS-type changes on April 19, 2001. Pattern identical to CSM-type changes on the previous day (see Figure 6).*

Acknowledgments

This work has been sponsored in part by NSF PECASE, NSF LSSDSV, and DOE SciDAC. We thank them for their support.

References

1. University of Oregon Route Views Project. *http://www.antc.uoregon.edu/route-views/* 2
2. C. Ahlberg and B. Shneiderman. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. *Proceedings CHI'94: Human Factors in Computing Systems*, pp 313–317, Boston, Massachusetts, 1994. 2
3. T. Atkison, K. Pensy, C. Nicholas, D. Ebert, R. Atkison, and C. Morris. Case Study: Visualization and Information Retrieval Techniques for Network Intrusion Detection. *Joint Eurographics-IEEE TCVG Symposium on Visualization (VisSym01)*, Ascona, Switzerland, 28-30 May 2001. 2
4. R.J. Brachman, F. Halper, P.G. Selfridge, T. Kirk, L.G. Terveen, A. Lazar, B. Altman, D.L. McGuinness, A. Borgida, and L.A. Resnick. Integrated Support for Data Archaeology. *International Journal of Intelligent and Cooperative Information Systems*, 1993. 2
5. L. Girardin. An Eye on Network Intruder-Administrator Shootouts. *Proceedings of the Workshop on Intrusion Detection and Network Monitoring (ID'99)*, USENIX Assoc, Berkeley, CA, USA, 1999. 2

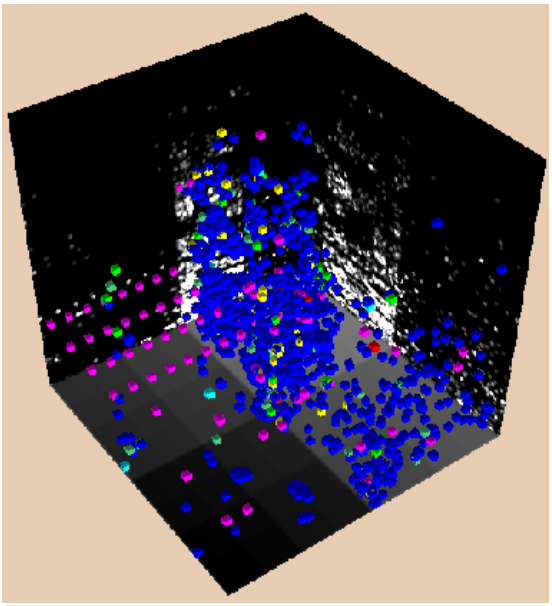

Figure 8: *3D representation for August 14, 2000 (original representation for the same day's data in Figure 4). Each AS change is represented by a cube with coordinates determined by IP prefix and AS number. Each cube is colored by its change type. Two planes in the background show a grayscale projected image of each cube. This picture shows an extraordinary number H-type (blue) changes involving different ASes. In addition, there are also some OS-type (pink) changes arranged in a regular pattern on one horizontal plane. This corresponds to AS 7777 announcing those prefixes. This is clearly a fault.*

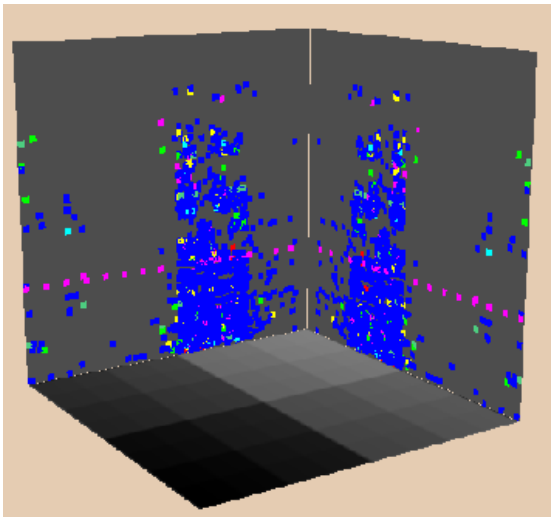

Figure 9: *Another way of looking at the 3D representation of the data on August 14, 2000, showing the projection of the cubes (of Figure 8) onto two perpendicular planes and coloring the projected image according to change type. The cubes are not shown. On this particular day, the OS-type (pink) changes and H-type (blue) changes obviously shows a pattern, in agreement with the other visual representation.*

6. J. Goldstein, S.F. Roth, and J. Mattis. A Framework for Knowledge-Based, Interactive Data Exploration. *Journal of Visual Languages and Computing*, pp. 339–363, December 1994. 2

7. M. Holsheimer and A. Siebes. Data Mining: The Search for Knowledge in Databases. *Report CS-R9406*, ISSN 0169-118X, Amsterdam, The Netherlands, 1991. 2

8. T. Lane. Hidden Markov Models for Human/Computer Interface Modeling. *Proceedings of the IJCAI-99 Workshop on Learning about Users*, pp 35–44, 1999. 1

9. H.Y. Lee, H.L. Ong, and L.H. Quek. Exploiting Visualization in Knowledge Discovery. Proceedings of the First International Conference on Knowledge *Discovery and Data Mining*, pp 198–203, Montreal, Quebec, 1995. 2

10. W. Lee. A Data Mining Framework for Constructing Features and Models for Intrusion Detection Systems. *PhD Thesis, Columbia University*, June 1999. 2

11. T. Lunt, A. Tamaru, F. Gilham, R. Jagannathan, P. Neumann, H. Javitz, A. Valdes, and T. Garvey. A real-time intrusion detection expert system (IDES) - final technical report. *Technical report, Computer Science Laboratory, SRI International, Menlo Park, California*, Feb 1992. 1

12. T. Lunt. Detecting intruders in computer systems. *Proceedings of the 1993 Conference on Auditing and Computer Technology*, 1993. 1

13. Y. Rekhter and T. Li. A Border Gateway Protocol 4 (BGP-4). *RFC 1771*, 1995. 1, 2

14. W. Ribarsky, J. Katz, T.Y. Jiang, and A. Holland. Discovery Visualization Using Fast Clustering. *Report GIT-GVU-99-14, IEEE Computer Graphics and Applications*, **19**(5), 32–39, 1999. 2

15. B. Shneiderman. *Designing the User Interface: Strategies for Effective Human-Computer Interaction: Second Edition*. Addison-Wesley Publ. Co., Reading, Massachusetts, 1992. 2

16. R. Spence. *Information Visualization*. ACM Press, 2000.

17. X. Zhao, D. Pei, L. Wang, D. Massey, A. Mankin, S.F. Wu, and L. Zhang. An Analysis of BGP Multiple Origin AS (MOAS) Conflicts, *SIGCOMM Internet Measurement Workshop 2001*. 1, 3