A Visual Technique for Internet Anomaly Detection

Soon Tee Teoh, Kwan-Liu Ma, S. Felix Wu
Computer Science Department
University of California, Davis
email: teoh,ma,wu@cs.ucdavis.edu

Xiaoliang Zhao
Computer Science Department
North Carolina State University
email: xzhao@unity.ncsu.edu

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 anomalous Origin AS changes by browsing through visual representations of selected aspects of the data. We demonstrate how our system can be used to detect errors on the Internet and to discover the source and reasons for each fault detected. Our visual approach can play a major role in an anomaly detection system.

KEY WORDS
Information visualization, Network Security, Data Exploration, Anomaly Detection

1 Introduction

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. Monitoring the Internet to recognize anomaly allows us to gain valuable understanding so that we can take corrective action in a timely manner. 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 [9, 12, 13]. That is, a set of "normal data is first analyzed to derive representative characteristics of normal use. These are then compared against the characteristics of unknown data to disclose abnormal behaviors.

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) [14] data to quickly detect and explain anomaly in Origin AS changes [17]. These changes can indicate either configuration errors or intentional attacks on the Internet.

Section 2 gives background on 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 that every host in the UC Davis campus network shares the same first 16 bits: 128.120. 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.

Each AS connects with one or more other ASes. Between two ASes, the Border Gateway Protocol (BGP) [14] is used to exchange network reachability information so that eventually routers know how to forward data packets to the correct destination. BGP routers exchange the 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 to 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.

The Origin AS for a particular prefix should remain the same unless the prefix's ownership has changed. However, due to valid network operations or 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. 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 [1] server. Then we collected 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 technique to aid data mining [7] include [4, 8, 10] and a method based on clustering [15]. Girardin [5] uses self-organizing maps 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 two goals of our visualization system. The first goal is for the user to be able to quickly identify anomaly in the data. The other is 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 further action. With appropriate visual metaphors, this 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 [11] 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 tranditional 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 a visual representation of the data, and possibly going back to 2.

An inherently iterative process is required in order to lead to successful discoveries. The user must be allowed to explore various aspects of the data in different parameter spaces. With interactive visualization, the human user can very easily conduct the iterative process towards the most promising direction. It is thus crucial to provide the user with the tools to interactively change parameters, focus on certain details, and animate the data over time.

Our design of the user interface adopts two main principles given in [16]:

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

These guidelines facilitate intelligent and productive computer-human interaction. In order to achieve interactive display rates despite the large size of the data, we need to design efficient data structures which also support viewing the data 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 occured. *Type* is the type of the change (described below).

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 involved 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 SOAS
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 6 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 start with the first two most significant bits of the address to place the IP address into 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 a 512×512 pixel square to represent the entire IP prefix space. With only 512×512 pixels, each pixel represents all IP prefixes sharing the same unique first 18 bits. Since IP prefixes have masks at most 24 bits long, at most 64 different IP prefixes map to the same pixel. Figure 2 shows additional windows offering closeup views of several different portions of the main window, which shows the entire IP prefix space. From the visualization, a 512×512 square is sufficient in spreading out the IP addresses in our data. The figure also shows that with the additional level of zooming into a portion of the data, individual IP prefixes can be distinguished.

In the main window, a pixel is colored yellow if an Origin AS Change occurred on the current day (February 19, 2001), and colored brown if a change occurred on a previous day (January 1, 2000 through February 18, 2001). 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, brightness determined by how long ago the change occurred (present day data shown the brightest), 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 background of the main window is shaped according to the IP prefix the pixel represents. The brighter the background, the larger the IP prefix represented. This example shows the data over a 416-day window from January 1, 2000 to February 10, 2001. To show only one day's data, the user can set the window to one day.

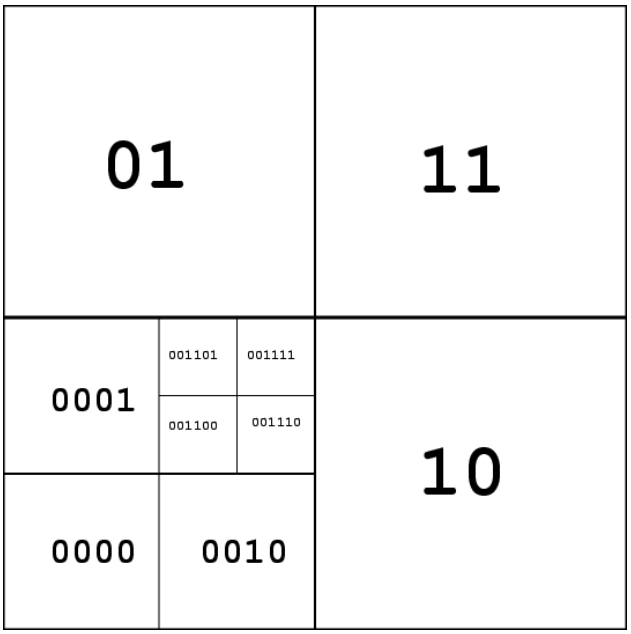

Figure 1. Quad-tree encoding of IP prefixes, showing the top levels of the tree and the most significant buts of the IP prefixes represented by each sub-tree (sub-square).

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.

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 place 4 lines surrounding the IP square, and an AS number is mapped to a pixel on one of the 4 lines. A line is then drawn from an IP prefix to an AS number if there is an Origin AS change involving that IP prefix 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 on 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 on 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 amanating from one AS are highlighted among thousands of lines.

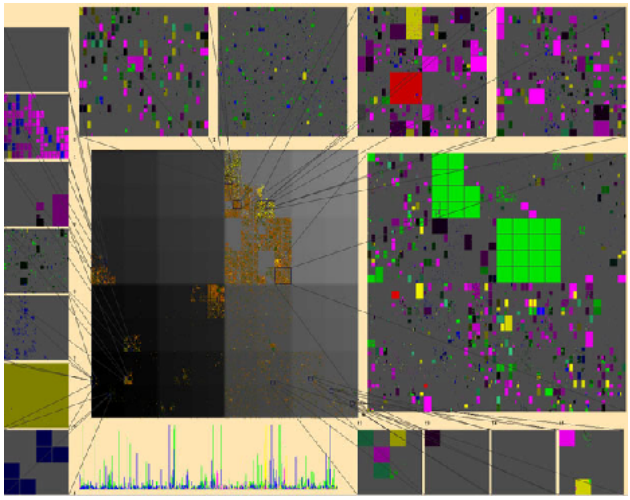

Figure 2. Visualization of data for 416 days up till February 19, 2001. The main window (large window on the left) shows the spread of IP prefixes involved in Origin AS changes over the entire IP prefix space. In the main window, multiple IP prefixes map to a single pixel. Zoom windows resolve the IP prefixes completely.

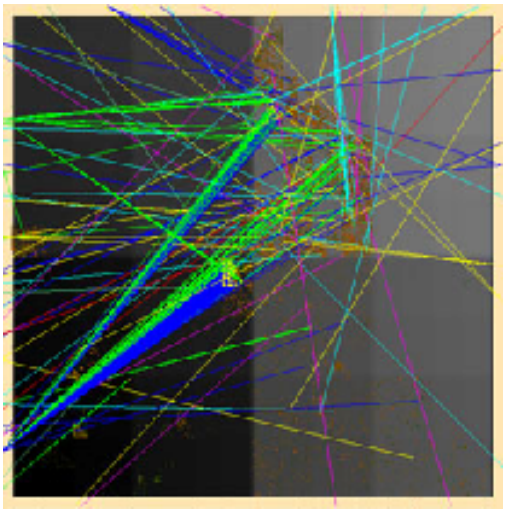

Figure 3. Data on a typical day (September 24, 2000)

4.4 Animation and Other Features

Our design shows one day's data at a time, allowing the user to animate the visualization, each frame showing consecutive day's data. With this "movie" display, 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 stand 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. We provide a slider bar to tell the date of the current data shown, and for the user to specify the date shown. A simple plot of the total number of changes of each type on each day is shown with the bar. 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.

4.5 Anomalies Detected

To validate the visualization approach for anomaly detection, two BGP experts used our tool to detect potential problems (faults and attacks) since January 1, 2000. With our tool, they were able to gain much insight into the BGP data. In this section, we describe some of the network errors they were able to detect and explain because of the visual system.

Many anomalies were detected simply because there are abnormally large numbers of MOAS conflicts on the same day. However, these anomalies can be just as easily detected by non-visual methods. Therefore, we focus on two categories of anomalies non-visual methods would have difficulty in detecting and analyzing: AS anomaly (unusual behavior per AS), and animation correlation (special correlation relations across the time domain).

4.6 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 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 moer common observation of H-type changes involving close IP prefixes and a single

AS, typified by 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 it routers, announcing many prefixes. 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 or ASes, the user can select those prefixes or ASes to focus on, as mentioned in Sections 4.2 and 4.3.

4.7 Animation Correlation

A very important aspect of our tool is to discover "correlation" relations via the animation of the BGP data sets. Figure 5 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 in Figure 6. Altough Figures 5 and 6 look disorderly, a nearly identical pattern is easily observed because the changes involved the exact same prefixes and ASes. This pattern causes the user to understand that the events on April 19 are corrective actions and not a new fault, demonstrating the effectiveness of human pattern recognition.

Other anomalies observed include private AS number leakage on September 18, 2000 and many days with high O-type 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.

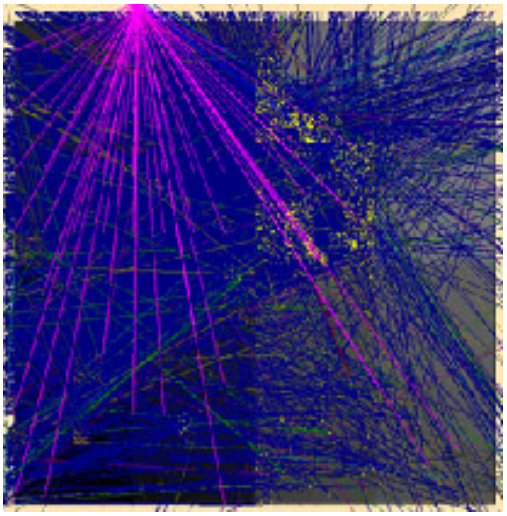

Figure 4. Data on August 14, 2000. Anomaly detected despite visual clutter. H-type (blue) lines involving many different ASes and IP prefixes indicate abnormal activity.

5 Conclusion and Future Work

We have demonstrated that, appropriate visual representation of data built into an interactive data exploration system is effective in detecting anomaly in Internet data. Not only that, but the visual system is also useful in revealing the source and nature of the detected anomalies. In the design of our visual anomaly detection system, we have given the user an overview of the data and allowed the user to focus on parts of the data. We have shown actual examples of how the visualization is able to easily detect and explain certain anomalies.

One limitation of the current approach is that it is not easy for the user to quickly find out which ASes or AS-IP pairs cause frequent or periodic changes over non-consecutive days. More data preprocessing incorporating statistical methods could help identify and highlight these phenomena during interactive visualization.

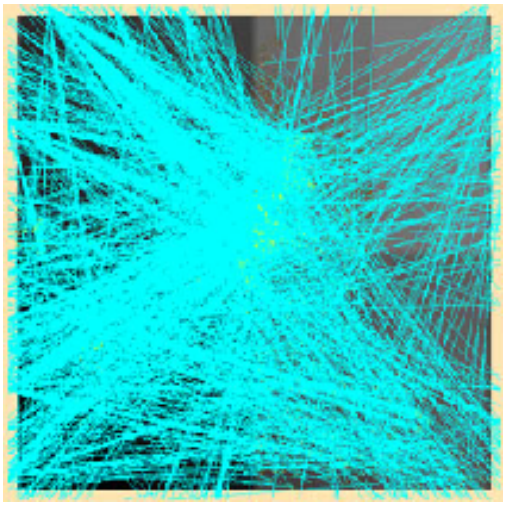

Figure 5. CSM-type changes on April 18, 2001

6 Acknowledgements

This work has been sponsored in part by NSF PECASE, NSF LSSDSV and DOE SciDAC.

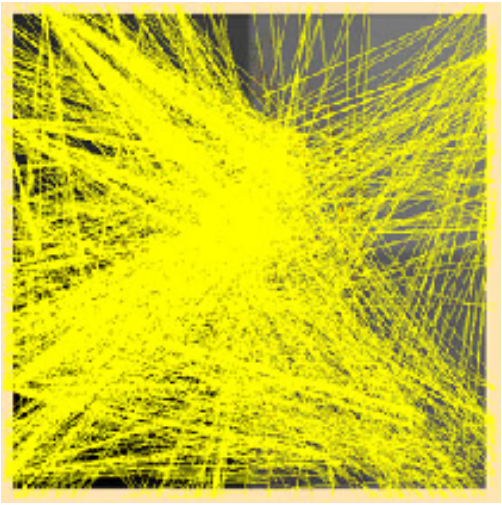

Figure 6. CMS-type changes on April 19, 2001. Pattern identical to CSM-type changes on previous day (see Figure 5).

References

[1] University of Oregon Route Views Project. *http://www.antc.uoregon.edu/route-views/*

[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*, Boston, Massachusetts, 1994, 313–317.

[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 TVCG Symposium on Visualization (VisSym01)*, Ascona, Switzerland, May 2001, 28–30.

[4] R.J. Brachman, F. Halper, P.G. Selfridge, T. Kirk. L.G. Terveen, A. Lazar, B. Altman, D.L. McGuiness, A. Borgida and L.A. Resnick. Integrated Support for Data Archeology. *International Journal of Intelligent and Coorperative Information Systems*, 1993.

[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.

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

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

[8] D. Keim and H-P Kriege. Visualization Techniques for Mining Large Databases: A Comparison. *Transactions on Knowledge and Data Engineering, Special Issue on Data Mining*, 1996.

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

[10] 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*, Montreal, Quebec, 1995, 198–203.

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

[12] 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, CA, February 1992.

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

[14] Y. Rekhter and T. Li. A Border Gateway Protocol 4 (BGP-4), *RFC 1771*, 1995.

[15] 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), 1999, 32–39.

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

[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*.