Configurable IP-space maps for large-scale, multi-source network data visual analysis and correlation

Citation: Scott Miserendino, Corey Maynard, William Freeman, "Configurable IP-space maps for large-scale, multi-source network data visual analysis and correlation," Proc. SPIE 9017, Visualization and Data Analysis 2014, 901705(3 February 2014); doi: 10.1117/12.2037862

As Published: 10.1117/12.2037862

Publisher: SPIE-Intl Soc Optical Eng

Persistent URL: https://hdl.handle.net/1721.1/137539

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Configurable IP-space maps for large-scale, multi-source network data visual analysis and correlation

Scott Miserendino, Corey Maynard, William Freeman

Scott Miserendino, Corey Maynard, William Freeman, "Configurable IP-space maps for large-scale, multi-source network data visual analysis and correlation," Proc. SPIE 9017, Visualization and Data Analysis 2014, 901705 (3 February 2014); doi: 10.1117/12.2037862

SPIE.

Event: IS&T/SPIE Electronic Imaging, 2014, San Francisco, California, United States

Configurable IP-space maps for large-scale, multi-source network data visual analysis and correlation

Scott Miserendino*[a], Corey Maynard[a], and William Freeman[a]

[a]Northrop Grumman Corporation, 8666 Veterans Highway, Millersville, MD, USA 21108

ABSTRACT

The need to scale visualization of cyber (IP-space) data sets and analytic results as well as to support a variety of data sources and missions have proved challenging requirements for the development of a cyber common operating picture. Typical methods of visualizing IP-space data require unreliable domain conversions such as IP geolocation, network topology that is difficult to discover, or data sets that can only display one at a time. In this work, we introduce a generalized version of hierarchical network maps called configurable IP-space maps that can simultaneously visualize multiple layers of IP-based data at global scale. IP-space maps allow users to interactively explore the cyber domain from multiple perspectives. A web-based implementation of the concept is described, highlighting a novel repurposing of existing geospatial mapping tools for the cyber domain. Benefits of the configurable IP-space map concept to cyber data set analysis using spatial statistics are discussed. IP-space map structure is found to have a strong effect on data clustering behavior, hinting at the ability to automatically determine concentrations of network events within an organizational hierarchy.

Keywords: IP-space maps, cyber, situational awareness, networks, web application

1. INTRODUCTION

Today's network analysts and cyber warriors lack a flexible, common operating picture (COP) of the cyber domain. The goal of a cyber COP is to get all critical information into the hands of decision makers in an easily understood format that provides an appropriate context to aid in interpreting that information. For network analysts and cyber warriors, this critical information can consist of a wide variety of network events such as intrusion detection alerts, network flow records, host-based security alerts, firewall and proxy logs, maintenance notices, and software health and status messages. Visual analytics can play a valuable role in summarizing, correlating, and contextualizing these large, complex datasets common to computer network operations.

Cyber visualization tools should allow decision makers to quickly understand, correlate, track, and perhaps act upon activities within this unique domain. One major challenge for a cyber COP that we seek to address is the need for a visual representation of the domain that is equally applicable to all missions, data types, and network configurations, and yet is customizable enough to meet the needs of individual users. In addition, approaches to cyber domain visualization must facilitate sharing information and analytic results across multiple organizations. Finally, the visual presentation and arrangement of the data must add analytic value beyond what is achievable with simple tables and charts.

Another major challenge faced by those users seeking broad cyber situational awareness, such as Internet service providers (ISP) and government agencies like the Department of Homeland Security (DHS) and the Department of Defense (DoD), is the need to visualize cyber data across a multitude of network enclaves. In most cases, these enclaves are not owned or operated by the organization attempting to gain situational awareness across the domain. This network data is either shared or observed in transit with little to no information known about either the data's source or destination network. To achieve a cyber COP suitable for these users, the visualization methodology must not require detailed information about the network structure, such as routing topology and end host identity or function. An example use case based on the planned Anonymous hacking group attack on May 7, 2013, against multiple U.S. and international entities is discussed in Section 5[1].

This paper presents a novel, generic approach to visualization of cyber domain information to enhance understanding of network events and to enable a common view of cyberspace, which we call IP-space maps. Our approach produces scalable, user-definable visualizations of the internet protocol space (IP-space) that handle a dynamic range from several

*Corresponding Author: Scott Miserendino; scott.miserendino@ngc.com; phone: 410-923-8444

Visualization and Data Analysis 2014, edited by Pak Chung Wong, David L. Kao, Ming C. Hao, Chaomei Chen, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 9017, 901705
 · doi: 10.1117/12.2037862

tens of hosts to the global Internet containing several billion hosts. Our IP-space maps address the problems of simultaneously handling multiple missions and cyber data types across large numbers of network enclaves without requiring detailed knowledge of enclave routing topology or host identity. Our implementation translates some of the most successful concepts and tools developed for visualization from the geospatial domain into the cyber domain. By first visualizing the entire cyber domain, or regions of interest within that domain, any dataset with an IP address field can be layered on top of the IP-space maps just as datasets with latitude and longitude fields can be layered on geospatial maps. Dataset context is made clearer by controlling or reconfiguring the organization of the IP-space maps. Furthermore, clusters within the dataset can be visually and analytically identified relative to the structure of the map. Cluster detection and identification make it easier for users to detect the contextual focus of the dataset, for example, the part of the U.S. critical infrastructure being targeted by a hacker group or botnet. A web-based user interface allows multiple users or even multiple organizations to leverage a single software installation. An application programming interface (API) allows users to easily incorporate the visualization technique into existing cyber situational awareness products.

This paper is organized as follows: Section 2 examines existing approaches to cyber data visualizations and highlights some of their limitations. Section 3 describes the underlying theory for constructing configurable IP-space maps and associated data layers. Section 4 details our implementation of an IP-space mapping tool, including our incorporation of existing geospatial mapping software and visualizations of several test datasets. Section 5 discusses the application of spatial statistics to this new type of map. Finally, we present our conclusions in Section 6.

2. EXISTING CYBER DATA VISUALIZATION APPROACHES

There are three primary classes of visualizations applied to cyber data that attempt to add context to the data: geospatial maps, network graphs, and IP-space views. Geospatial maps provide physical location and context to the datasets by using IP geolocation transformations to convert IP addresses to corresponding latitudes and longitudes. Network graphs are used to show physical and logical network connection context. Finally, IP-space views attempt to contextualize the data set relative to the organization of the cyber domain. Visualization approaches that rely solely on the data, such as tables and charts, do not add any context to aid in data interpretation and will not be addressed here.

2.1 IP geolocation and geospatial mapping

Because geospatial maps have both a meaningful and familiar frame of reference, many cyber visualization tools incorporate them. To use the geospatial domain for cyber data set visualization, however, requires that network elements first be geolocated. The process of geolocating network elements, typically referred to as IP geolocation, is subject to many difficulties. Today's networks do not allow for wide-spread, reliable, high-resolution geolocation. IP geolocation is subject to many sources of error depending on the technique used to estimate the host's physical location[2]. Commercially available IP geolocation databases[3] are typically accurate to no better than 40 km. The best academic attempts at active IP geolocation, measurements of network traffic instead of a database lookup to determine location, report accuracy of 30 km using the landmark-latency technique[4]. In some special circumstances, web servers have been geolocated to within 1 km using a variant of the landmark-latency technique, but these results are atypical and require dense landmark concentrations around the target IP[5].

IP geolocation does not make sense for some classes of network devices. Virtual machines (VM) that co-exist in the same server or are spread across multiple servers have no distinct location. Network elements on satellites or those that have IP addresses tied to satellite providers cannot be readily visualized on traditional geospatial maps.

The relatively coarse resolution of these geolocation techniques causes network elements across an entire city to be binned together on the geospatial maps regardless of their nature, function, or owner. This binning behavior can result in the gross false correlation of network element data and can severely limit the usefulness of geospatial displays for cyber situational awareness. To avoid or mitigate the challenges of IP geolocation, logical visualizations of the domain are often used.

2.2 Network graphs

Another traditional approach to cyber domain visualization is connection-based. The location of elements in the visualization is driven by their connections to other elements in the form of network graphs[6]. The application of this approach, however, relies on a detailed understanding of the network's topology. In some cases, this is a reasonable assumption, for example, networks under the user's administrative control. For large enterprise networks or networks divided by many administrative domains, however, this topological data can be difficult to discover and maintain. If the goal is to visualize data going to or from a massive network, then the network graph approach is limited. Since access to most network domains is unavailable, a thorough connection-based approach to the display of global datasets is unrealistic[7].

Network graph-based visualizations must also solve the problem of how to place the nodes and edges within the visualization. There exists no absolute location for any network element. Typically, the elements are dynamically arranged to minimize visual clutter and/or to focus the user's attention on a particular part of the graph. Some tools allow the user to select from a variety of algorithms for generating the graph layout. In some layout algorithms, such as force-directed, as nodes and edges are added, removed, or moved in these graphs, the relative position of all the elements change. This forces users to reorient themselves every time data changes, limiting the "at-a-glance" understanding of the domain and data. This visual instability is in stark contrast to geospatial displays, where map elements maintain their relative positions. Furthermore, available computing resources may limit the application of layout algorithms to very large graphs, for example, those in excess of tens of thousands of nodes.

Another challenge faced by a connectivity-based approach to IP-space visualization is visual clutter. As nodes and their associated connections are added to network graphs, the number of elements being visualized can grow at a quadratic rate. This growth of visual information, or clutter, quickly overwhelms a user's ability to make sense of it. Problems with visual clutter can arise with well under one hundred nodes and must be addressed for graphs with over one hundred nodes. Clustering to a reasonable number of nodes is one common approach used to de-clutter network graphs. Other approaches (view distortion, sampling, edge displacement, radial layouts, balloon layouts, etc.) can be used to minimize the impact of clutter; however, none of these solutions completely eliminate it[8].

2.3 IP-space visualizations

Every externally reachable element on a network must have a unique IP address in some subnet just as every physical object must have a unique position in physical space. Because every network element must be uniquely addressable, IP-space serves as a natural domain for network-centric data. Although less common than geospatial and network graphs, there exist a variety of approaches that use the organization of IP-space as the context for visualization. IP-space can be treated as a bounded, discrete, one-dimensional domain, which, due to its size, must be mapped to a two or three-dimensional domain for the purposes of visualizing the data. The differences and potential value and drawbacks of IP-space approaches are based on the choice of mapping function.

Teoh introduced a mapping method based on using a quad-tree where a square is iteratively divided into 4 equal squares. IP addresses are located in the quad-tree based on two bits of the IP address being used at each iteration, starting with the two most significant bits in the address[9]. In a similar approach, Ohno uses just the first two octets of the IP address to create an IP Matrix[10]. A fractal approach using a Hilbert curve was introduced by Munroe[11] to achieve a nesting of the IP-space by subnet block and was later adopted by others for data visualization using heat maps[12,13]. In all cases these existing IP-space mapping techniques are independent of any attribute of the IP address being mapped and are based solely on the numeric representation of the address, limiting their contextual value. Micro IP address block assignments, address assignment delegation to ISPs, IP address recovery and reassignment, and growth of total assigned space to an organization over time have lead many organizations to own multiple discontiguous portions of the IP space that cannot be rendered together using these methods.

Others have investigated the use of 3D visualizations to represent the cyber domain. The CyberNet project used a city metaphor to create a virtual 3D world representing cyberspace[14]. Lau used a three-axis space to create a scatter plot of network events with one axis dedicated to IP address[15]. 3D data visualizations, however, can prove difficult to work with because of data overlays on 2D screens.

3. IP-SPACE MAPS

Our work generalizes the concept of hierarchical network maps (HNmaps)[16,17] to add additional user-definability required for a cyber COP and to improve information sharing through a web-based implementation, a multi-user interface, and a dataset-agnostic input file type. Hierarchical network maps are treemaps[18] of the IP-space where leaf nodes are characterized by the number of IP addresses in a subnet and the levels of the tree are fixed as continent, country, ASN, and IP prefixes. The tree levels determine the nesting of labelled, rectangular boxes within the visualization. The nesting effect can be seen in Figure 1. For example, the element of North America at the content level will have a child node of Canada, and in the visual display the box representing North America will contain a box representing Canada. HNmaps simultaneously optimize display space utilization, layout preservation, geographic awareness, and rectangle aspect ratio. IP-space maps generalize the HNmap concept of hierarchical trees of IP-space by allowing any user-defined hierarchy ending in an IP subnet level, including those unbounded by a geospatial context.

IP-space maps have several other distinct and novel features compared to traditional treemaps and HNmaps. First, IP-space maps visualize a domain, not a particular dataset. Hence, these maps can simultaneously support visualization of many disparate data sources and data types (e.g., IDS event data, traceroute data, and patching compliance ratings for an organization). This is akin to being able to plot multiple lines on the same set of axes. Traditional treemaps are limited to visualizing at most two simultaneous attributes and HNmaps only one. Treemaps visualize these attribute values by assigning one attribute to determine block area and optionally another to determine block color (e.g., number of IP addresses in a subnet may be used to determine size while total bytes transferred to the subnet may determine color). IP-space maps, however, can use a rich combination of dots, lines, polygons, and icons, each with potential variations in size, line width, and color for each independent data source, as shown in Figure 1. Second, for time-varying datasets, a treemap approach can be difficult to understand because changes in the attribute value determining block size can cause significant rearrangement of the visual display. This rearrangement can disorient users and force them to spend time locating familiar or interesting data elements. By treating the IP-space as a domain, time variant geometries on an IP-space map show up in the same location each time, allowing the user to develop familiarity with the visual layout and reducing the time required to visually analyze the data.

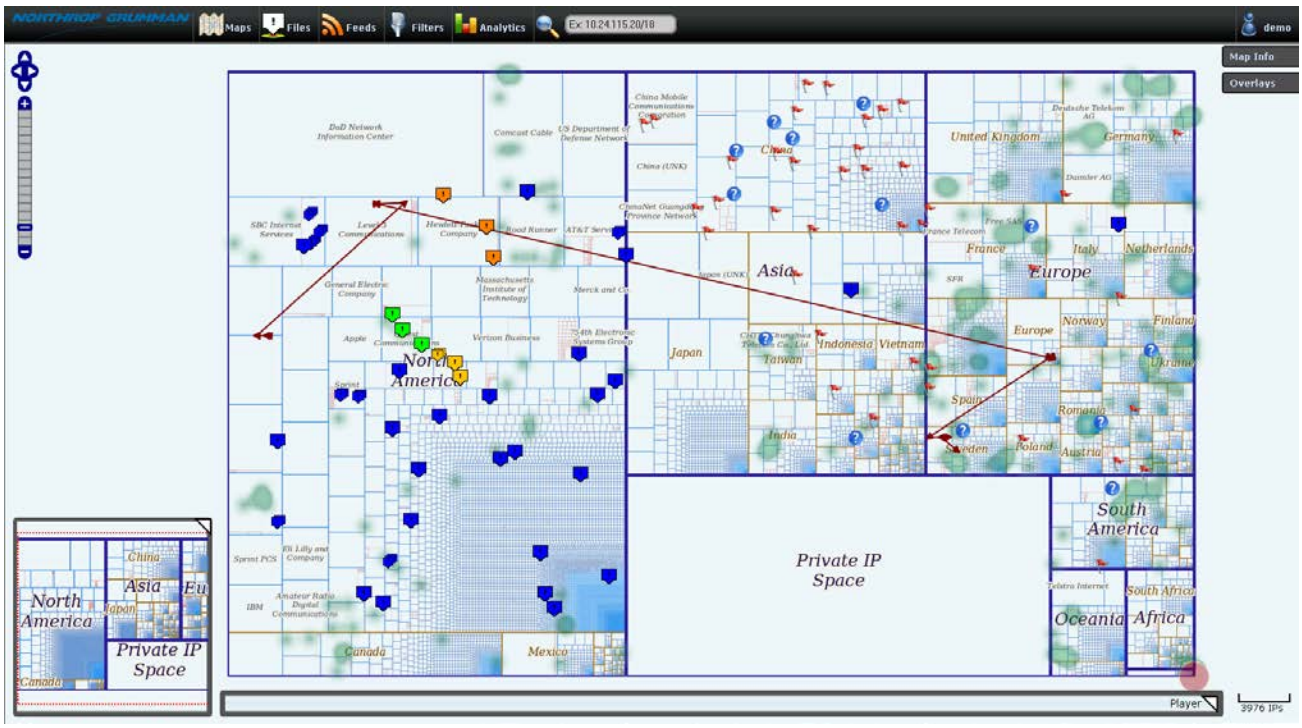


Figure 1. Global IPv4-space map with multiple data set overlays. Square placemarks are IP-addresses registered to four individual companies, each in a different color. Heatmap shows locations of Tor exit nodes. Line shows a traceroute. Red flags and question mark icons show IP addresses with out-of-date web browsers and unidentified web browsers, respectively.

3.1 Configurable IP-space maps

Each IP address or block of addresses, known as a subnet, is "owned" by a series of progressively larger, more encompassing entities. IP addresses are owned whether they are globally routable or part of a private network. By imposing an organizational schema onto the full IP-space for a network, the location of individual elements becomes meaningful. The globally routable IP-space is naturally regulated in this hierarchical manner[19]. Each globally routable IP address must be allocated by one of the five regional internet registries (RIRs). The five RIRs allocate blocks of IP address to large Internet service providers and other micro-end users. ISPs and other independent service providers often subdivide RIR allocations to smaller organizations and report that information to the RIRs as SWIPs (Shared WHOIS Project data). By basing an organizational schema on the naturally occurring ownership data provided and maintained by the RIRs, globally routable IP-space can be meaningfully visualized without the need for connectivity data using the IP-space map concept.

The default IP-space map levels for globally routable addresses are continent, country, ISP, organization, and host/subnet. During visualization, the host/subnet level is not used, but the collection of subnets within an organization does help to determine the location of individual IP address on the IP-space map. While this default IP-space hierarchy structure provides some geospatial or geopolitical context to the map at the highest levels, it provides little value not previously achievable with a traditional geospatial map. Many other hierarchical organizations of the space are possible that can add additional context.

Often companies or organizations will allocate portions of their IP-space (whether globally routable or private) to individual sub-organizations, agencies, buildings, functions, or departments according to some hierarchical structure. Using IP-space maps, users can view their network data relative to these custom organizational structures. Using these specialized maps, perhaps in conjunction with global maps of the public IP-space, network analysts get a complete contextual understanding of who is the source or destination of particular network events.

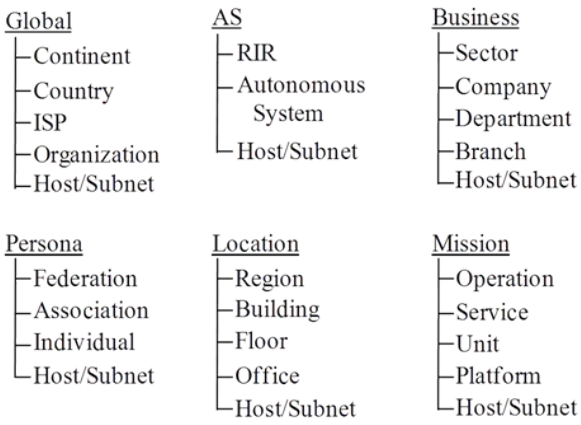


Figure 2. Potential network hierarchy structures for IP-space maps

Applications of reconfigurable treemap hierarchies have been shown to be beneficial in analyzing non-cyber related, multivariable data sets by allowing the user to construct hierarchies from the various dimensions of the data[20]. Here we demonstrate the value of a variety of potential organizational structures for the IP-space, some of which are shown in Figure 2 alongside the default global structure. Using the RIR autonomous system assignments, we created an Autonomous System (AS)-based structure. The business structure starts by organizing the space into business sectors, then individual companies, then department, branches, and so on within those companies. By selecting a number of representative companies from the Department of Homeland Securities sectors of the U.S. critical infrastructure, we have develop a prototype IP-space map of U.S. critical infrastructure globally routable network assets. The persona structure is based around individuals who own a variety of network assets (such as desktops, laptops, virtual machines, etc.) and grouping the individuals by their associations and then federations of those associations. This type of structure would be useful for analyzing hacker groups, terrorist organizations, or advanced state-sponsored network threats. A

location hierarchy is also possible where a large organization is broken up into its physical locations first by region, then building, then floor, and finally down to an individual office. For military users, a mission hierarchy may make the most sense with the IP-space broken down into operations, services, units, and finally platforms. Using an organization's IT asset management databases, the data for these maps can be generated. In the case of dynamic IP address assignment, the entire address block is assigned to the appropriate parent element. Based on their user-configurable structures, IP-space maps apply operational context to the visualization of network events detected by host-based security systems and network intrusion detection systems.

3.2 Cyber mark-up language (CML)

Network data must be layered onto the IP-space maps for them to be useful as part of a cyber common operating picture. Network data is often multi-dimensional, including fields such as IP address, time, port, protocol, metadata about the network traffic, or results of analytic processing. We developed a cyber mark-up language (CML) tailored to the cyber domain based on the approach of input file formats for geospatial maps, such as shapefiles and KML. CML is an eXtensible markup language (XML) file that focuses on how the data should be visualized rather than the structure of the data itself. Like KML, it is based on the concept of geometries plotted over the rendering of the background map. Supported geometry types include placemarks (a single IP address), lines (a list of IP addresses), and polygons (a subnet or block of IP addresses). CML allows users to decide geometry styling parameters, such as color, line widths, and icon for placemarks. Unlike KML, it also allows an optional, generic, real-valued number to be assigned to a geometry that can be used to automatically control styling.

The choice to focus on visualization parameters rather than data structure has two major benefits for a cyber COP. First, it allows users to share as little or as much detail about the geometry as they wish. Sharing is a major goal in cyber security, yet it is often hampered by legal and policy restrictions. With every organization having a different set of rules and regulations governing the type of data it can share, minimizing the requirements on data structure is critical. CML only requires the geometries' IP addresses. Second, the focus on visualization parameters ensures that regardless of the selected implementation of the IP-space map, the resulting visualizations of the same CML file will be consistent. This makes CML portable and provides users across implementations a common look to the data.

4. IMPLEMENTATION

We selected a web application implementation for the IP-space map concept to minimize software deployment and maintenance complexity as well as to provide an opportunity for integration with other web-based interfaces of cyber security tool. The system architecture, shown in Figure 3, consists of five main components: client-side JavaScript, server-side PHP scripts, a PostgreSQL database, a map-making module, and an image tile server. The system is designed based on a repurposing of existing open source software originally designed for creating geospatial map widgets for websites. To provide a familiar interface for user interaction, we used popular user interface concepts such as geometry-based data layers and map image tiles based on the open source software products.

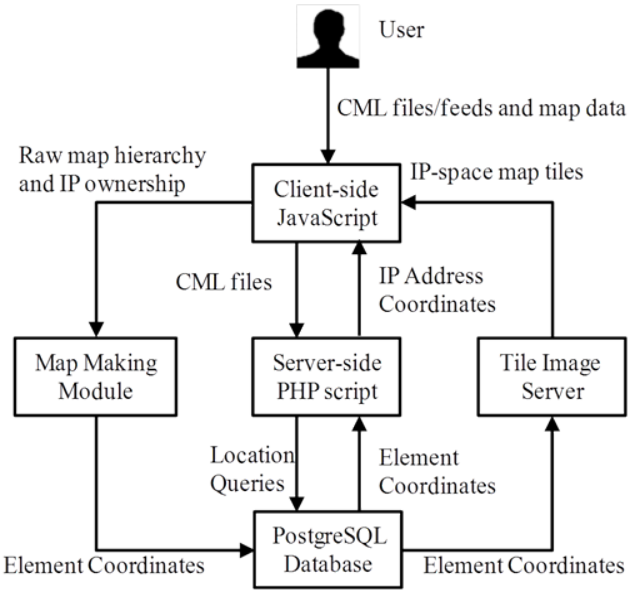


Figure 3. Diagram of the web-based system architecture for IP-space maps

4.1 Constructing IP-space maps

The map-making module is used to automatically construct an IP-space map and enter its data into the PostgreSQL database. IP-space maps are represented as a set of tables in the database, one for each level in a map's network hierarchy. The elements of each table contain a label or name for an element, a reference to an element in its parent's level, optionally a consecutive set of IP addresses belonging to that element, the total number of IPs associated with the label, and a set of box coordinates assigned by the treemap algorithm. All first-level elements reference a root node. All last-level elements contain the IP address set. The same label may be used for multiple non-consecutive parts of the IP-space.

The raw map level data stored in the database consists of a list of elements for each level and the consecutive parts of the IP-space owned by that element. The number of IPs owned for each unique label is calculated and stored along with its percent of its parent element's IP-space. A squarified treemap algorithm[21] is used to assign each unique label a portion of the total map space. The squarified treemap algorithm was selected to limit high-aspect ratio boxes, making it easier to apply labels on the map.

In a departure from traditional treemap implementations, the starting area to be partitioned is not based on a physical viewing area. Since treemap algorithms attempt to completely fill the space provided, they must drop elements once an element's allocated size gets below a certain threshold. At only one pixel per IP address, the entire IPv4 space would take a display wall of approximately 82,900x51,900 pixels. Since physical displays of appropriate size are not practical, a virtual display area is used with an area at least as large as the total number of IP addresses in the map. The virtual display is kept at an approximate 1:1 ratio of pixels to number of IP addresses. The use of a virtual display allows the treemap to scale but must be processed for display over the web on computer monitors.

4.2 Image tiles and zoom

The virtual display is handled using the same approach taken to scaling digital geospatial maps. The process includes dividing the map into image tiles of fixed size (256x256 pixels in our case) by creating tile sets for each resolution level. These tiled images are cached after generation for speed in later retrieval and then stitched together client-side based on the user's current zoom level and region of view. Because of the popularity of geospatial mapping widgets, a variety of open source projects exist to handle these functions; we elected to use GeoServer (http://geoserver.org) because of its implementation of open data standards and its stable development.

GeoServer must be specially configured to generate treemap tiles. Each level in the network architecture requires a styling file that controls how the boxes for that layer will be rendered. We visually separate hierarchy levels by color, border line thickness, and label font size. Higher level tiers are given thicker borders and larger label font sizes. We control label rendering as a function of goodness of fit within the element's rectangular box as rendered at a particular zoom level. Finally, at low zoom levels, deeper levels of the hierarchy have the opacity of their borders reduced. This gives the user a sense of where additional structure exists while limiting the visual clutter. A composite image is created from individual renderings of each level. See Figure 1 for an example of a global IP-space map based on a continent (purple borders), country (orange borders), ISP (blue borders), and organization (red borders) network hierarchy.

We have generated IP-space maps of a variety of sizes and tier structures. Our five-level global IP-space map is based on ownership data purchased from MaxMind, Inc. which contains IP allocations at the country, ISP, and organization level. We also generated a three-level map based on AS allocations with data provided by MaxMind, Inc. We constructed a four-level map based on a subset of the global space focusing on the U.S. Department of Homeland Security's 18 sectors of the U.S. critical infrastructure (CI). Finally, we created a five-level map of private IP-space based on a simulated network. Table 1 describes some of the key performance and scalability metrics for each map. Partition time describes how long the map-making module takes to process and generate the structure for the entire map. Precaching time describes how long it takes to precache the first seven zoom levels of each map. Partitioning and caching metrics are based on a quad-core CPU with 2 GB of RAM and are provided to give a rough sense of scaling efficiency. Partitioning and precaching performance is a complex combination of factors, but it typically trends with the number of hierarchy levels and the number of elements within each hierarchy level and is relatively independent of the number of IP addresses in the map.

Table 1: IP-space Maps Scaling and Performance Metrics

Map	Max. Tier Elements	Num. IP Addresses	Partition Time (min)	Precaching Time (min)
Global	3,264,961	4,027,582,434	70.0	5.5
AS	34,083	2,370,344,140	0.9	0.8
CI	9,142	369,189,238	0.6	1.4
Private IP	102	709	0.5	1.3

4.3 Locating IP addresses on IP-space maps

The server-side PHP scripts and PostgreSQL database are responsible for calculating the positions of IP addresses on a particular map. The database holds records of the coordinates and associated subnets of all of the boxes generated by the map-making module treemap algorithm. An individual IP address location on a map is based on the coordinates of the bounding box of the parent element containing that IP address on the lowest level of the network hierarchy. The element of interest may contain other non-contiguous portions of the IP-space as well. All IP addresses in the parent element are therefore arranged in numerical order from lowest to highest and rastered from left to right across the box. The center coordinates for each address are such that an integer number of rows and columns exist across the bounding box with the last row or column cell size adjusted to account for rounding in boundary box dimensions relative to the number of IP addresses contained. The center coordinates of an individual IP address can be calculated based on the location of that address in the ordered list of all addresses assigned to its parent element.

4.4 Data layer visualization and map selection

Data layers consist of geometry sets described in the CML format. The client-side JavaScript ingests either static CML files or dynamic feeds of CML-formatted messages and sends them to the backend web server. The server returns the location on the active map of all IP addresses present in that dataset. The client-side JavaScript uses these map locations to plot the geometries on the IP-space map just as latitude and longitude are used to plot geometries on a geospatial map. We employ existing open source geospatial mapping software, called OpenLayers (http://openlayers.org/), to handle the data layers. The client-side code also supports several other data visualization options for the CML files, including time lines, tables, and video-like playing of time-stamped geometries.

An atlas feature is used to facilitate switching between various IP-space maps and a geospatial map. When a new active map is selected, the data layers must be reloaded since IP address locations depend on the map. The atlas provides the user a preview of all the maps for which they have been granted access.

5. SPATIAL STATISTICS AND DATA CLUSTERING ON IP-SPACE MAPS

Through the detection of clustering within and across cyber datasets on the IP-space maps, users gain an understanding of what part of their organization or network is being targeted by malicious actors, undergoing malfunction, or needing update or repair. For example, if intrusion detection events are disproportionally located in a particular department of a company, perhaps that department's IT administrator is not properly configuring firewalls and maintaining software patches. For users of a cyber COP, the visual clustering could indicate coordinated targeting of particular sectors of the U.S. critical infrastructure. Section 5.1 discusses an example analysis based on the planned distributed denial of service (DDoS) attacks of the Anonymous hacking group in May 2013. Statistical and deterministic cluster existence and identification algorithms from spatial statistics can be used to supplement the visual identification of clusters under a variety of definitions of distance, similarity, and minimum cluster density.

5.1 Use-case: Anonymous hacking group opUSA attack plans

To demonstrate how IP-space maps can be used to identify the focus of activity within a set of network events, we examined the target list from the Anonymous hacking group's opUSA DDoS attacks on May 7, 2013[1]. The target list includes 140 URLs for websites belonging to institutions primarily in the United States. Since the geospatial focus of the attacks are known, we use the IP-space maps to explore the logical focus of the attacks. Our goal in the analysis is to demonstrate the ability of IP-space maps to aid users in quickly identifying the focus of large-scale coordinated cyber attacks by visually identifying clusters of target IP addresses. For network security analysts monitoring the U.S. Critical Infrastructure IP-space, Figure 4 shows how opUSA would have looked if the attack victims or their Internet service providers shared information indicating they were experiencing a DDoS attack. In this case, the full attack list was published in a public forum prior to the attack, but in most coordinated cyber attack scenarios the target list is unknown to security analysts, and attacker intent must be inferred by observing reports from the victims or through monitoring by service providers.

Visual analysis of the opUSA target list using IP-space maps shows two major clusters of activity. The first step in analyzing the target list is to resolve the target URLs to IP addresses by using public domain name services. The IP addresses were combined into a single CML file and the CML file was loaded on our U.S. Critical Infrastructure IP-space map. The visual display was changed from individual placemarks to a density-based heatmap where high placemark density is shown as red and low density is shown as green. Figure 4 shows two concentrations of targets. One concentration occurs across the Banking and Finance sector. The second shows a concentration in the Information Technology sector with a particular focus on Akamai IP-space. Zooming in on the concentration areas reveals a broad attack across many financial institutions but a more limited effect in the Information Technology sector. Targets in the Information Technology sector are located primarily in companies that provide web-hosting services. Knowing these areas of focus for the attack allows network security analysts to provide timely, targeted warnings to other members of these sectors that they may become targets as well.

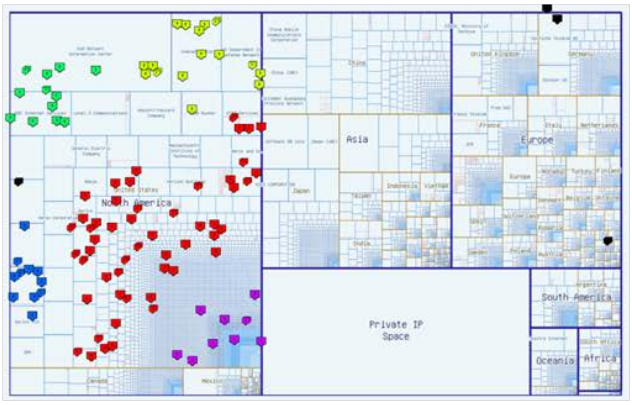
Comparing the results of the IP-space analysis to the target URL list shows that web sites for companies within the Banking and Financial sector do make up the vast majority of opUSA targets. Examination of the URLs alone, however, does not show a concentrated targeting of web-hosting service companies. The IP-space analysis shows that because many of the targeted banks use the same web-hosting service, Akamai, that service is indirectly affected by the attacks.

Figure 4. Data point density heatmap of opUSA target IP addresses plotted over an IP-space map representing elements of the U.S. critical infrastructure

5.2 Automated cluster detection

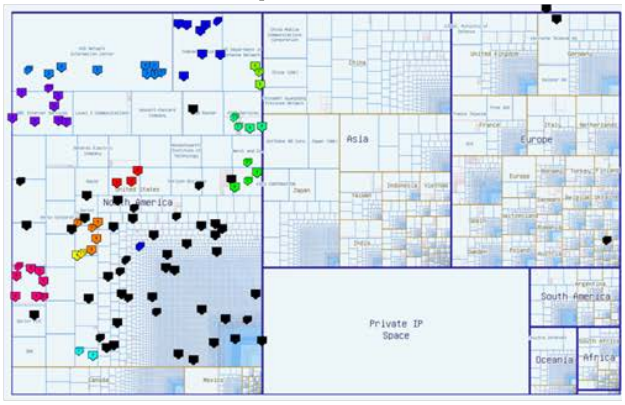
Whether for cluster detection or cluster identification, IP-space maps require specialized distance functions since geospatial distance is not applicable. While IP-space maps allow users to visually identify clusters of network events, there is added value to automating the process. By using data cluster formation as a measure of effectiveness, automated analytical cluster detection can help identify what map structure is best with which to view a particular data set. In addition, the visual identification of clusters of data points on the map primarily relies on data point proximity. IP-space maps, unfortunately, can produce a distorted sense of distance. In reality, all elements of the same tier are in some sense equidistance from each other, yet on the IP-space map some must be located further away than others. For example, all ISPs in the United States can be viewed as equally separated from one another, yet on the map in Figure 1 Comcast Cable is located next to DoD Network Information Center and far from IBM. This is a result of the treemap layout algorithm used to generate the map and is a fundamental limitation of the approach. With automated cluster detection and identification algorithms, however, non-Euclidean distance measures more appropriate to the nature of cyberspace can be used to aid users in finding truly clustered data points.

Euclidean Distance



Number of clusters: 5 Moran's I: +0.509

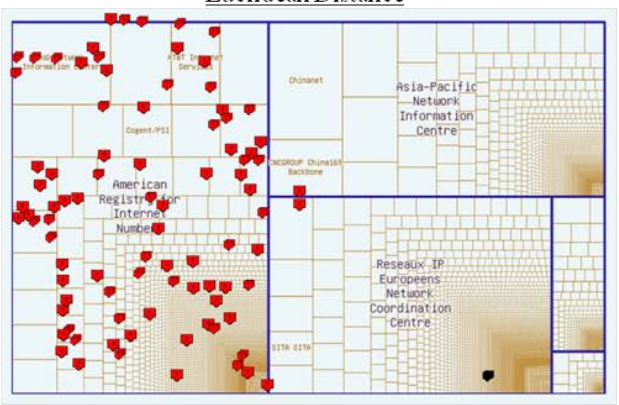
Map Tier Distance



Number of clusters: 11 Moran's I: +0.041

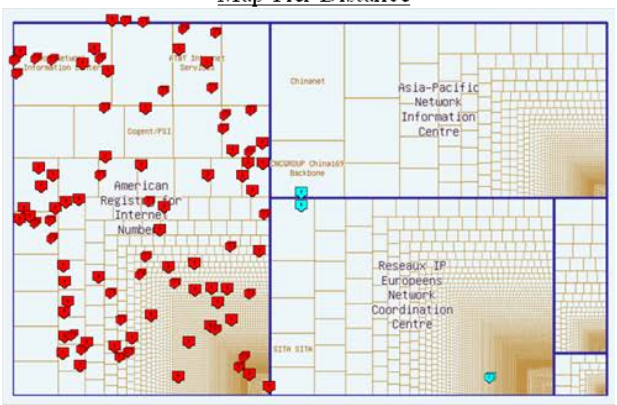
Autonomous System IP-space Map

Euclidean Distance



Number of clusters: 1 Moran's I: +0.513

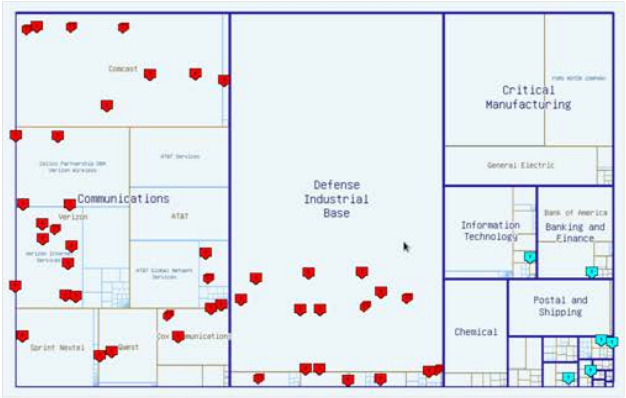
Map Tier Distance



Number of clusters: 2 Moran's I: +0.023

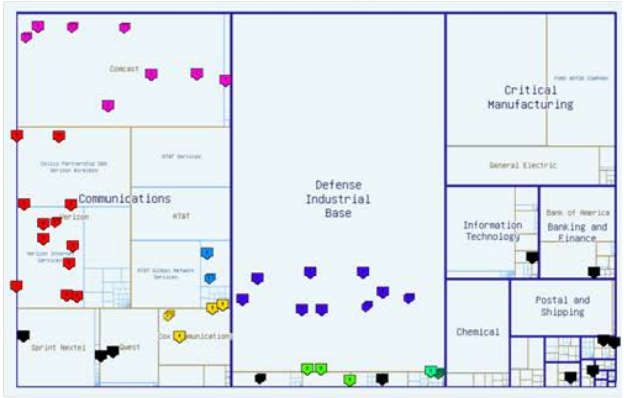
Critical Infrastructure IP-Space Map

Euclidean Distance



Number of clusters: 2 Moran's I: -0.045

Map Tier Distance



Number of clusters: 7 Moran's I: +0.087

Figure 5. Clustering analysis of web sessions on three IP-space maps. Each identified cluster is a different color. Unclustered data points are considered noise and are colored black.

Clustering behaviour of data in a space is described by a spatial autocorrelation statistic developed by Patrick Moran, known as Moran's I[22]. Moran's I provides a global spatial clustering statistic where -1 indicates perfect dispersion, 0 indicates random arrangement, and +1 indicates perfect correlation. It is an indicator of whether clustering exists, but it does not separate or identify the clusters within the data set. Moran's I is defined as

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2}, \quad (1)$$

where N is the number of data points, w is a weight matrix, and X is the value associated with the data point ($\bar{X}$ is the average of X). Data points are limited to placemark geometries in a CML file. The weight matrix is based on a variety of distance measures users can select and is defined as

$$w_{ij} = \begin{cases} 1, & i = j \\ dist^{-1}(i,j), & i \neq j. \end{cases} \quad (2)$$

The weight matrix has the property of being symmetric ($w_{ij} = w_{ji}$), having values between 0 and 1, and inversely scaling with the distance between different points. We implemented three distance measures. The first is based on the Euclidean distance between the points on the virtual display. The second is based on the minimum number of hops between the IP addresses in the IP-space map's network hierarchy. Third, we used border gateway protocol (BGP) advertisements to determine the minimum number of AS hops between two publically routable IP addresses.

The Euclidean and network hierarchy distances are dependent on the IP-space map. Euclidean distance-based clusters are the most obvious on the visual display of the map, but they disregard the implied separation of the network hierarchy levels just as geospatial distance disregards geopolitical borders, which are often meaningful. The network hierarchy distance can be used to find the map that best aggregates the data or to identify the organizational structure to which the data best aligns. The BGP distance is independent of the map and is an indication of how close the IP addresses are in terms of routing to one another over the Internet.

Figure 5 shows the results of analyzing the Euclidean and network cluster detection and identification on a set of 147 IP addresses observed to be connecting to Northrop Grumman public web servers. The IP addresses were identified from a three-hour capture of netflow data from a Northrop Grumman Internet point of presence. For each address the total number of bytes transferred during the session is used as the value. The figure illustrates the effect of map selection and distance measure on clustering. The results show much better clustering on the global and autonomous system maps due to a preponderance of traffic having addresses within the United States, where the American Registry of Internet Numbers is responsible for allocating AS numbers. The connections were randomly spread among companies within the U.S. critical infrastructure.

5.3 Cluster identification

To aid the user in identifying clusters within cyber datasets, we implemented a version of the density-based spatial clustering of applications with noise (DBSCAN) algorithm[23]. As in the implementation of Moran's I, a suitable distance measure must be selected. The user can select from the same three distance measures discussed in Section 5.2. We applied the DBSCAN algorithm to our Northrop Grumman netflow dataset on several IP-space maps. Figure 5 shows the results of the DBSCAN cluster identification using a neighborhood radius, *Eps*, of 5 for the network hierarchy distance measure and 5000 for the Euclidean distance measure. Minimum points per cluster, *MinPts*, of 3 was used in all cases. Unclustered points, or noise, are colored black. IP addresses not on a particular map were excluded from the clustering analysis.

Results show that the map tier-based distance measures cause cluster groups to be formed within second- or third-level network hierarchy elements. Euclidean distance measures produce clusters across multiple hierarchy-level boundaries.

6. CONCLUSION

Traditional approaches to cyber data set visualization have often relied on showing where events may have originated and not on who originated those events. Logical views focus on visualizing who, but suffer from reliance on network topology and have limited scalability. In this paper, we present IP-space maps that maintain the scalability and data agnostic features of geospatial mapping yet incorporate the ownership focus of logical views. Furthermore, these IP-space maps provide users the flexibility to define multiple hierarchical structures through which their data can be visualized. Significant reuse of web-based geospatial map software was incorporated into our implementation providing users a familiar interface. We defined multiple distance functions over the IP-space maps and demonstrated the application of techniques to analyze spatial statistical clusters to cyber data sets. We explored the effect of map hierarchical structure on clustering behaviour. We believe that web-based configurable IP-space maps are a valuable addition to cyber common operating pictures and network security tools.

In the future, we plan to explore automated methods for discovering hierarchical structures from which to generate the IP-space maps. In particular, we wish to examine map structures based on dynamic statistics of network behavior; for example, basing the first level of a map on the frequency a network has exchanged traffic with other organizations in the past. This will allow for quick identification of data exchanges to organizations for which data is not typically exchanged and may help in identifying unauthorized data transfers.

ACKNOWLEDGMENTS

The authors thank Andrew Bishop, Carl Shek, Shane Lester, Henry Chen, and Chris Lubas for their contributions to the development of the IP-space map prototype.

REFERENCES

[1] C. Thompson, "Anonymous to US: 'We Will Wipe You Off the Cybermap'," CNBC.com, 6 May 2013. <http:// www.cnbc.com/id/100712145 > (9 July 2013). www.cnbc.com/id/100712145

[2] J. Muir and P. van Oorschot, "Internet geolocation: Evasion and counterevasion," *ACM Computing Surveys (CSUR)* 42(1), 4:1-4:23 (2009).

[3] MaxMind, "GeoIP City Accuracy for Selected Countries," 2012, <http://www.maxmind.com/app/city_accuracy> (25 February 2013). www.maxmind.com/app/city_accuracy

[4] B. Wong, I. Stoyanov and E. G. Sirer, "Octant: A Comprehensive Framework for the Geolocalization of Internet Hosts," Proc. 4th USENIX conference on Networked systems design & implementation, (2007).

[5] Y. Wang, D. Burgener, M. Flores, A. Kuzmanovic and C. Huang, "Towards Street-Level Client-Independent IP Geolocation," Proc. 8th USENIX conference on networked systems design and implementation, (2011).

[6] I. Herman, G. Melancon and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," IEEE Transactions on Visualization and Computer Graphics 6(1), 24-43 (2000).

[7] R. Oliveira, D. Pei, W. Willinger, B. Zhang and L. Zhang, "In search of the elusive ground truth: the internet's as-level connectivity structure," Proc. ACM SIGMETRICS (2008).

[8] W. Cui, "A Survey on Graph Visualization," Ph.d. thesis, Hong Kong University of Science and Technology (2008).

[9] S. Teoh, K.-L. Ma and X. Zhao, "A Visual Technique for Internet Anomaly Detection," Proc. International Conference on Computer Graphics and Imaging (2002).

[10] K. Ohno, H. Koike and K. Koizumi, "IP Matrix: An Effective Visualization," Proc. Ninth International Conference on Information Visualization (2005).

[11] R. Munroe, "Map of the Internet," xkcd.com, December 2006, <http://www.xkcd.com/195/> (2 March 2013). www.xkcd.com/195/

[12] "IPv4 Census Map," The Cooperative Association for Internet Data Analysis (CAIDA), October 2007. <http://www.caida.org/research/id-consumption/census-map/> (2 March 2013). www.caida.org/research/id-consumption/census-map

[13] The Measurement Factory, "Gallery of IPv4 Heatmaps," 2009. <http://maps.measurement-factory.com/gallery/> (2

March 2013). maps.measurement-factory.com/gallery

[14] C. Russo Dos Santos, P. Gros, P. Abel, D. Loisel, N. Trichaud and J. Paris, "Mapping Information onto 3D Virtual Worlds," in Proc. IEEE International Conference on Information Visualization, (2000).

[15] S. Lau, "The Spinning Cube of Potential Doom," Communications of the ACM 47(6), 25-26 (2004).

[16] F. Mansmann and S. Vinnik, "Interactive Exploration of Data Traffic with Hierarchical Network Maps," IEEE Transactions on Visualization and Computer Graphics 12(6), 1440 – 1449 (2006).

[17] F. Mansmann, "Visual Analysis of Network Traffic: Interactive Monitoring, Detection, and Interpretation of Security Threats," Ph.d. thesis, University of Konstanz (2008).

[18] B. Johnson and B. Shneiderman, "Tree maps: A space-filling approach to the visualization," Proc. IEEE Visualization, (1991).

[19] B. Donnet and T. Friedman, "Internet Topology Discovery: A Survey," IEEE Communications Surveys & Tutorials 9(4), 56-69 (2007).

[20] A. Slingsby, J. Dykes and J. Wood, "Configuring Hierarchical Layouts to Address Research Questions," IEEE Transactions on Visualization and Computer Graphics 15(6), 977 - 984 (2009).

[21] M. Bruls, K. Huizing and J. van Wijk, "Squarified Treemaps," Proc. Joint Eurographics and IEEE TCVG Symposium on Visualization, (1999).

[22] P. A. P. Moran, "Notes on Continuous Stochastic Phenomena," Biometrika 37(1/2), 17-23 (1950).

[23] M. Ester, H. Kriegel, J. Sander and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise," Proc. 2nd International Conference on Knowledge Discovery and Data Mining, (1996).