Examining Correlation Techniques to Improve Strategic Decision-Making Through Advanced Cyber Situational Awareness

Dr. Preston Frazier, Dr. Risa Lin, Dr. Dennis McCallam
Northrop Grumman Corporation
2691 Technology Drive
Annapolis Junction, MD 20701
USA

preston.frazier@ngc.com

ABSTRACT

As the scope and scale of cyber-attacks continue to increase, numerous organizations around the World are presently facing the inability to make astute decisions from the vast amount of data they are collecting in the cyber domain. Stakeholders from cybersecurity analysts to senior-level government officials need a cyber common operating picture (C-COP) to make their decisions based on the cyber data collected. The need for advanced cyber situational awareness (CSA) to ascertain, conceptualize and predict potential cyber threats and vulnerabilities in real-time is critical. The techniques showcased in a C-COP provide an acute understanding of a cyber-attack to enhance the decision-making of analysts, network operations center (NOC) leaders, and senior executives. Through the use of novel correlation techniques, the context of malicious activities are easily and quickly revealed to allow informed decision-making on how to counter, mitigate, and/or defend against discovered attacks. The dominant methods for correlating and visualizing information in the cyber domain are based on geospatial data, network topology, or Internet Protocol (IP) space. However, these methods pose several limitations to cyber situational awareness, including geolocation accuracy, visual clarity, and consistency of representation. We present a recently developed approach that generalizes the concept of hierarchical tree maps to the correlation and visualization of cyber data by leveraging the concept of "ownership" within IP-space. The resulting IP-space maps have been used successfully for CSA in use cases such as cyber incident response, forensic investigation of cyber-attacks, and information security continuous monitoring (ISCM). Our hierarchical IP-space maps visualize a domain, not a particular dataset and thus, can simultaneously display many disparate data sources and data types as overlays to support a C-COP. Decision makers need a specialized and customizable visualization to quickly understand, correlate, track, and be able to act upon activities within IP-space. Our implementation also supports improved information sharing across public and private entities by using a web-based, multi-user interface and an open data format called Cyber Markup Language (CML).

1.0 OVERVIEW

Among the many issues discussed at the 2016 NATO Summit in Warsaw was a defining moment that officially recognized cyberspace as a "domain of operations." NATO's missions are increasingly net-centric as the modern security environment relies on communications and information systems for command and control of weapons, intelligence, communication and logistics. The amount of effort and technical skill required to mount a successful attack is decreasing and the value associated with successful attacks is increasing. Applying NATO's core strategies of collective defence, crisis management, and cooperative security to cyberspace will require

more effective cyber defence capabilities and integration of cyber defence into Alliance operational and strategic planning. At the forefront of cyber mission-critical capabilities is the need for a flexible, common operating picture (COP) of the cyber domain that utilizes tools and techniques that heighten cyber situational awareness (CSA). The techniques showcased in a cyber common operating picture (C-COP) should provide an acute understanding of a cyber-attack to enhance the decision-making of technical analysts, network operations center (NOC) leaders, and senior management.

The need for advanced cyber situational awareness to ascertain, conceptualize and predict potential cyber threats and vulnerabilities in real-time is critical. Lessons learned from critical infrastructure protection and the modernization of industrial control systems (ICS) clearly demonstrate that "security by obscurity" is no longer an adequate strategy. Cybersecurity frameworks such as "Defense in Depth" and the "cyber kill chain" are commonly deployed for corporate enterprise security, but emphasize perimeter-focused thinking that is being exploited by our adversaries. Furthermore, computing and networking architectures are evolving with the rise of software-defined networking, container technologies, and the Internet of Things (IoT) as well as the increasing use of mobile, cloud, and virtual platforms. These technologies were developed primarily to increase efficiency and interoperability, rather than privacy and security, and they increase both the size and complexity of cyber-attack surfaces. NATO faces unique challenges in that the security of the Alliance depends on the cyber capabilities of individual Allies and partner organizations. Cyber resilience and experience solving these hard challenges varies widely between member Nations and across NATO sites and operational facilities.

As the scope and scale of cyber-attacks continue to increase, a primary problem that numerous organizations around the World are presently facing is the inability to make astute decisions from the vast amount of data they are collecting in the cyber domain. The market is crowded with cyber defence tools, such as intrusion and malware detection, which can be time-consuming and expensive to deploy and maintain. These tools generate extremely large, complex, and disjoint datasets for analysts to sift through, making it easier for a persistent threat to eventually slip through. Given the sheer volume, variety, veracity, and velocity of information, computer analytics are necessary to make sense of it. The complexity of implementing these tools can also result in silos of information that are handled by separate teams, thus inhibiting effective coordinated responses. The goal of a cyber COP is not only to gather all data critical to imminent decisions, but also to provide appropriate and consistent context for interpreting the data. We examine how novel correlation techniques and visual analytics enable users to distil large amounts of data to detect and quickly reveal the context of malicious activities and support strategic decision-making on how to counter, mitigate, and/or defend against discovered attacks.

2.0 VISUALISING THE CYBER DOMAIN

The development of tools for CSA requires an understanding of the technical and cognitive tasks that support perception and sense-making. A critical component of CSA tools is scalable visualizations of cyber data sets and analytical results [1]. For cyber analysts, this critical information can consist of a wide variety of host-based and network-based events such as malware identification alerts, intrusion detection alerts, network flow records, firewall and proxy logs, and health and status messages. These data sources may be only semi-structured and extracting meaningful information through data correlation can be challenging. Visual analytics plays a valuable role in summarizing, correlating, and contextualizing these extensive, complex datasets common to computer network operations. They should address how data, related analytics, and appropriate mission contexts are visualized and presented effectively to stakeholders to support the decision-making process. Visual representations of this domain should be customizable to meet the needs of specific users and stakeholders, but must also facilitate sharing information and analytic results across multiple organizations [2].

Another major challenge faced by those users seeking broad cyber situational awareness is the need to visualize cyber data across a multitude of network enclaves. In many cases, these enclaves are not solely 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.

2.1 GEOSPATIAL MAPPING

Many approaches to CSA visualization rely on geospatial imagery because they have both a meaningful and universally familiar frames of reference (Figure 1A). To use the geospatial domain, however, network elements must first be correlated with their physical geographic location. The process of geolocation of network elements, typically referred to as Internet Protocol (IP) geolocation, is subject to many difficulties [3]. First among these difficulties is that IP geolocation eliminates the uniqueness of an IP address in favour of an estimate of the hardware's latitude and longitude. Even if one were willing to give up traceability to a unique element, today's Internet does not allow for wide-spread, reliable, high-resolution geolocation. For western European countries, commercial IP geolocation databases are typically ~40-60% accurate at 10km resolution, and ~50-70% accurate at 25km resolution [4]. With this level of geo-resolution, network elements across an entire city are grouped together regardless of their nature, function, or ownership.

For virtual machines (VM), network elements behind a network address translation (NAT) device, and cyber activities using anonymizing networks like TOR, IP geolocation is nearly impossible without owner- supplied databases, human intelligence (HUMINT) efforts, or active geolocation methodologies. Target- assisted geolocation that exploits global positioning systems (GPS), wireless networking, and Bluetooth technologies deployed in the target's local infrastructure requires source code to be running on the target asset, which is not feasible for coverage of the global IP space for cyber defence. Measurement-based geolocation measure latency from known hosts or vantage points to a target asset and may use geolocation databases and multilateration techniques to estimate the target's location in correlation with the latency measurements. Research shows that the expected accuracy of measurement-based geolocation may only be ~100km, which accounts for several limitations on the probing rate. Since network paths may be stable for only a couple days, this technique must complete all probing requests within that period of time for global IP-space. The probing rate must also be capped to avoid significantly inhibiting network traffic at important vantage points [5].

Geospatial mapping also has limited relevance in the cyber domain, where the physical location of users, systems and data paths may not be informative for developing cyber situational awareness. Two assets that are physically collocated may have no correlation regarding their function or mission context. Modern computing and networking architectures allow system assets and functions to be broadly distributed and easily relocated. In geospatial data visualizations, these assets would not appear in a consistent location, limiting the ability of an analyst to recognize patterns and detect anomalies. While critical assets such as command and control systems and industrial control systems are unlikely to implemented in the cloud or on virtual hardware, assets at the network edge in many sectors use these technologies and may have a more vulnerable cybersecurity posture. Since geopolitical boundaries are not represented in the cyber domain and cyber-attacks certainly do not respect them, geospatial mapping of cyber information is not sufficient for holistic CSA. The limitations of IP geolocation are also a major challenge in the attribution of international cyberattacks to state or non-state actors.

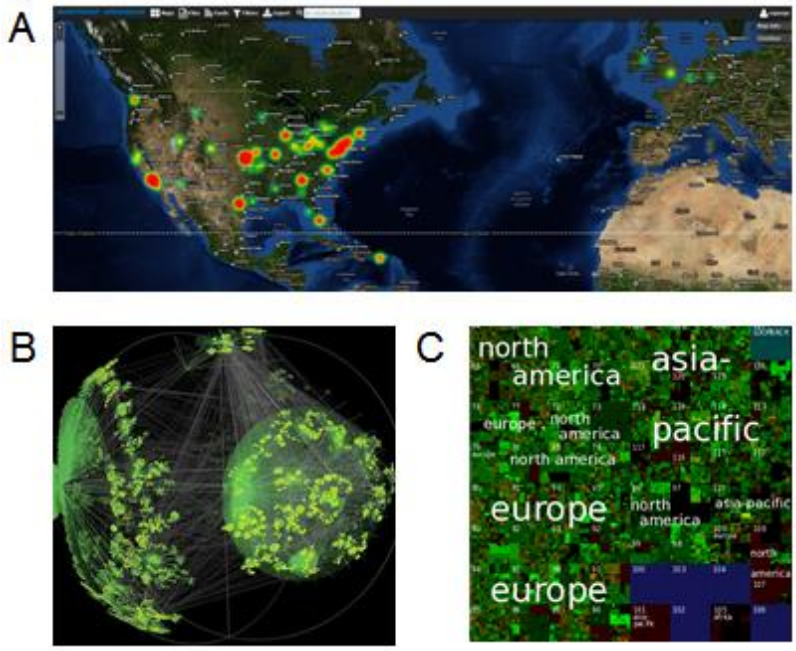

Figure 1: Three common methods for correlating and visualizing information in the cyber domain. A) Geospatial heat map of IP addresses targeted by Anonymous hacking group's OpUSA distributed denial of service (DDoS) attacks on May 7, 2013 [12]. The same dataset is visualized using hierarchical IP-space maps in Figure 3. B) A network graph representing just 0.1% of global IP-space. C) A section of the Hilbert map visualizing the USC/LANDER Internet address census taken November 2010 with partitions labelled with geographic location [10] .

2.2 NETWORK GRAPHS

Another traditional approach to visualizing the cyber domain is a connection-based network graph in which each node represents a network asset and the edges connecting each node describe the actual network connectivity [6]. This is an intuitive representation because every element on a network must have a unique Internet Protocol (IP) address within a subnet, just as every physical object must have a unique position in physical space. A class of algorithms known as graph analytics has long been applied to network graphs to understand the structure and activity of the computing network, detect changes and anomalies in network structure and activity, conduct forensic investigation of cyber-attacks, and orchestrate cyber defence strategies [7]. The correlation of cyber events, such as intrusion detection, or attributes, such as software patch status, with nodes in the network can be used to identify vulnerability or attack paths [8].

Furthermore, by visualizing the network as a network graph, there exists no absolute location for any network element. The elements are dynamically arranged to minimize visual clutter and/or to focus the user's attention on a particular part of the network. The user can often select from a variety of options for generating the layout of the network topology. As nodes and edges are added or removed from these graphs, the relative position of network elements changes significantly. The lack of a stable frame of reference forces users to orient themselves every time they change focus to a different part of the network or to a different network element. There is no "at-a-glance" understanding of the space when displayed in this manner. For global data sets that may contain hundreds of thousands of nodes and connections, the visual clutter is prohibitive to developing actionable

insights and CSA (Figure 1B).

2.3 IP-SPACE MAPS

While IP-space is a natural domain for network-centric data, the challenge remains of consistently organizing IP-space such that it 1) preserves element uniqueness, overcoming a limitation of IP geolocation-based visualizations, and 2) preserves clarity, overcoming a limitation of network graph visualizations. One of the most popular representations uses the Hilbert Curve, a space-filling curve that can be used to preserve a one-dimensional feature of a dataset as much as possible in a two-dimensional layout [9]. For IP-space described by a 32-bit IPv4 address, the map is arranged such that numerically sequential IP addresses are next to each other (Figure 1C). Similar to heat maps and network graphs, blocks in the Hilbert map can be visually color-coded to represent metrics such as activity, and larger partitions can be further labelled with other identifying information such as a common geographic location or ownership by a single organization [10].

Since these IP-space mapping techniques are based solely on the numeric representation of the address, the contextual value of the data visualization is limited. In the Hilbert map, it is possible for more than one partition to exist for a single label (Figure 1C). This could be due to multiple factors: address assignment delegation to Internet Service Providers (ISP), IP address recovery and reassignment, increase in total IP-space assigned to an organization as it grows over time, etc. Many organizations own multiple non-contiguous portions of global IP-space, which dilutes the power of the Hilbert IP-space map to reveal hotspots within an organization. Furthermore, the visual distance between partitions and nodes is one of the most immediate factors for an analyst looking for patterns or anomalies. The numerical adjacency of IP address encodes very little information in itself, and the Hilbert map's constraint on the use of visual space prevents it from efficiently correlating other available metadata that could be useful for cyber situational awareness.

2.4 HIERARCHICAL TREE MAPS OF IP-SPACE

We present hierarchical tree maps of IP-space that overcome the data visualization challenges of previous approaches and support greater cyber situational awareness by incorporating user-customizable metadata. Our approach generalizes the concept of hierarchical network maps to organize one-dimensional IP-space into hierarchical tiers represented as a two-dimensional map [11]. Hierarchical network maps are tree maps where leaf nodes are characterized by the IP addresses in a subnet and the levels of the tree represent a user-defined hierarchy. The hierarchy centers on the concept of "ownership" within the IP-space. Each IP address or block of addresses, known as a subnet, is "owned" by a series of progressively larger, more encompassing entities.

Using specialized IP-space maps in conjunction with maps of the public IP-space improves cyber situational awareness by providing network analysts with a more complete contextual understanding of who is the source or destination of particular network events and how that event correlates with other events within a given contextual view. 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. 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. 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

detailed network topology data using the IP-space map concept.

A hierarchical IP-space map with a geospatial context could have tree levels that represent continent, country, ASN, and IP prefixes. However, this adds little value that is not already available with traditional geospatial maps. A hierarchy with an organizational or administrative context could have tree levels that represent individual sub-organizations, agencies, buildings, or departments according to some hierarchical structure (Figure 2). Network administrators already commonly allocate portions of IP-space so that they correspond to organizational structures for convenience. The hierarchy may also span both geographical and non-geographical contexts. For example, an IP address may be owned by a server, that server is used by a division within a company, that company received the IP address from an Internet service provider (ISP), and that Internet service provider received it as part of a country-wide allocation by a regional internet registry. Thus, that IP address is progressively owned by an individual server, a division within a company, a company, an ISP, and a country. Geographically dispersed IP addresses belonging to the same division within a company will be readily apparent in a hierarchical view but may not be apparent in a geospatial view.

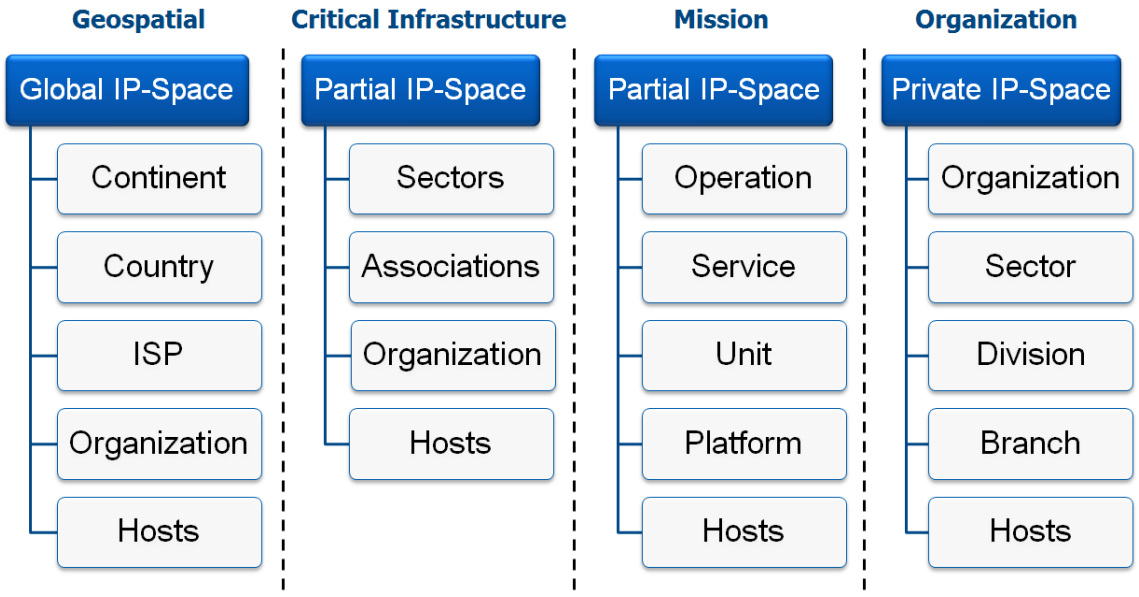

Figure 2: IP-space map hierarchies are user-defined and result in different visualizations that provide more strategic or operational context to the data and can improve cyber situational awareness.

Similarly, the user-defined hierarchy can represent mission areas or capability focuses. For the Anonymous hacking group's OpUSA distributed denial of service (DDoS) attacks on May 7, 2013 [12], we analysed the target list of domain URLs that was published online by the organization. First, the target URLs were translated to IP addresses using public domain name services. Then, the IP addresses were visualized on an IP-space map representing the critical infrastructure sectors of the United States, as identified by the U.S. Department of Homeland Security (Figure 3). Visual analysis of this IP-space map immediately highlights clusters of activity in the Banking and Finance and the Information Technology sectors. Targets in the Information Technology sector are located primarily in companies that provide web-hosting services. Zooming in on the IT sector further showed a concentration on Akamai IP-space, a web-hosting service that many of the targeted banks were using and which was indirectly affected by the attacks. 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.

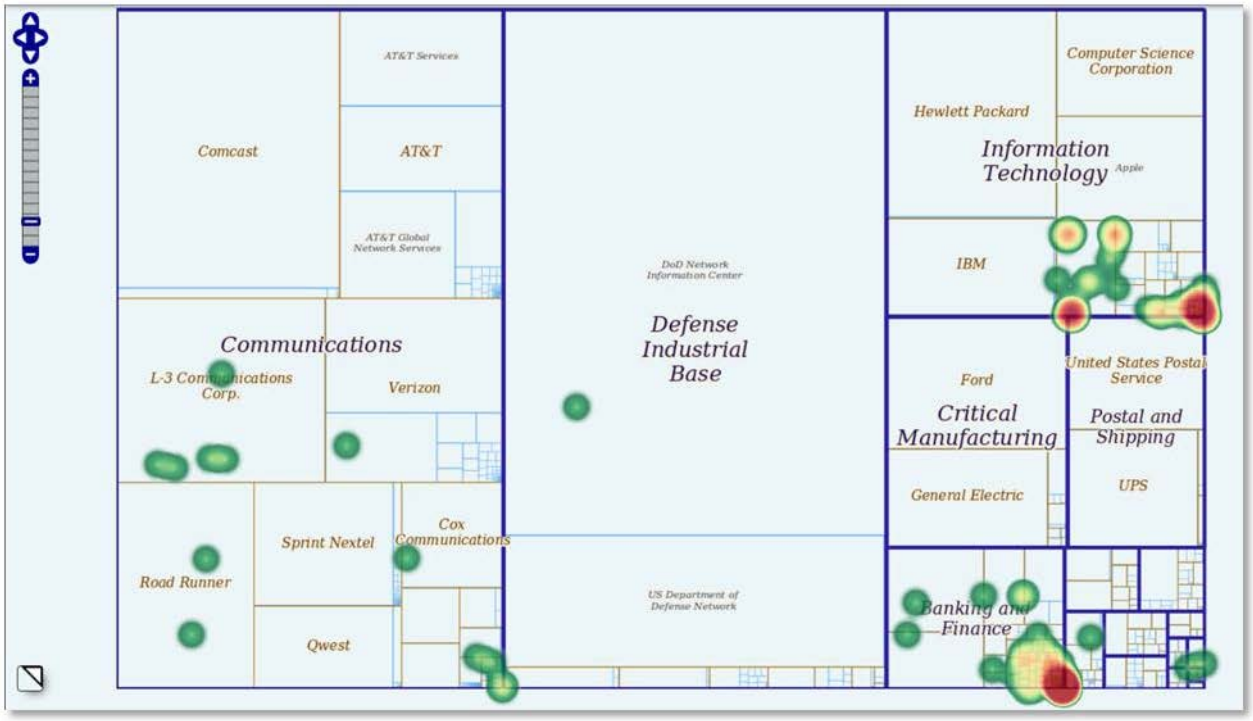

Figure 3: Data point density heat map of Anonymous hacking group's OpUSA distributed denial of service (DDoS) attacks on May 7, 2013. Target IP addresses plotted over an IP-space map representing sectors and corporations that make up the U.S. critical infrastructure. [12] The same dataset is visualized geospatially in Figure 1A.

Our approach to cyber data correlation also provides additional opportunities to combine automated predictive techniques, such as machine learning, with visual analytics to improve CSA. Even though all IP addresses within a given tier are, in a sense, equidistant from each other in the cyber domain, the IP-space map necessarily displays related IP addresses at some distance from each other. Since the Euclidean distance in the image is the primary means by which analysts visually identify clusters, this is a limitation of the approach. To address this issue, we explored several automated cluster detection algorithms using cyber-specific distance metrics, such as the number of network "hops" between IP addresses or network hierarchy separation. Results showed that the hierarchy tier-based distance measures caused cluster groups to be formed within second- or third-level network hierarchy elements. Euclidean distance and network hop measures produced clusters across multiple hierarchy boundaries. While changing the IP-space map hierarchy alters the tree map layout of the visualization, changing the distance metric and cluster algorithm only alters the colour or styling of individual points in distinct layers of the visualization. This allows users to interactively explore distinct combinations of IP-space map hierarchies and distance metrics to visually reveal meaningful clusters that might otherwise remain unidentified.

By visualizing a domain rather than a particular dataset, IP-space maps support the visualization of disparate data sources from multiple perspectives as overlays on the maps, akin to weather maps that simultaneously display area radar information and point-based observations such as lightning. Traditional tree maps are limited to

visualizing two simultaneous attributes at most, and traditional hierarchical network maps are limited to only one. IP-space maps, however, can use a rich combination of dots, lines, polygons, and icons, each with potential variations in size, line width, and colour for each independent data source, as shown in Figure 4. These additional elements can be correlated and visualized at any layer of the user-defined hierarchy, allowing analysts to make full use of any available metadata that could improve their cyber situational awareness. For information security continuous monitoring (ISCM) and cyber incident response use cases, these visual elements can be updated in real-time to indicate assets with current security risks, alert a network operations team to anomalous network activity or user activity, and view the interconnectivity of network assets. Cybersecurity teams with different geographical or functional areas of responsibility may use different hierarchical maps to provide context that is specific to their team's mission while sharing a single, authoritative cyber COP. For example, a team that is focused on external threats to an organization may configure a map to focus on indicators of network intrusion. A team that is focused on insider threats may configure a map to focus on indicators of unauthorised software and user access violations. A key feature of this approach is that the underlying heterogeneous data sources that provide input to the IP-space maps can remain the systems of record for specific tasks such as hardware asset management, software asset management, and vulnerability management.

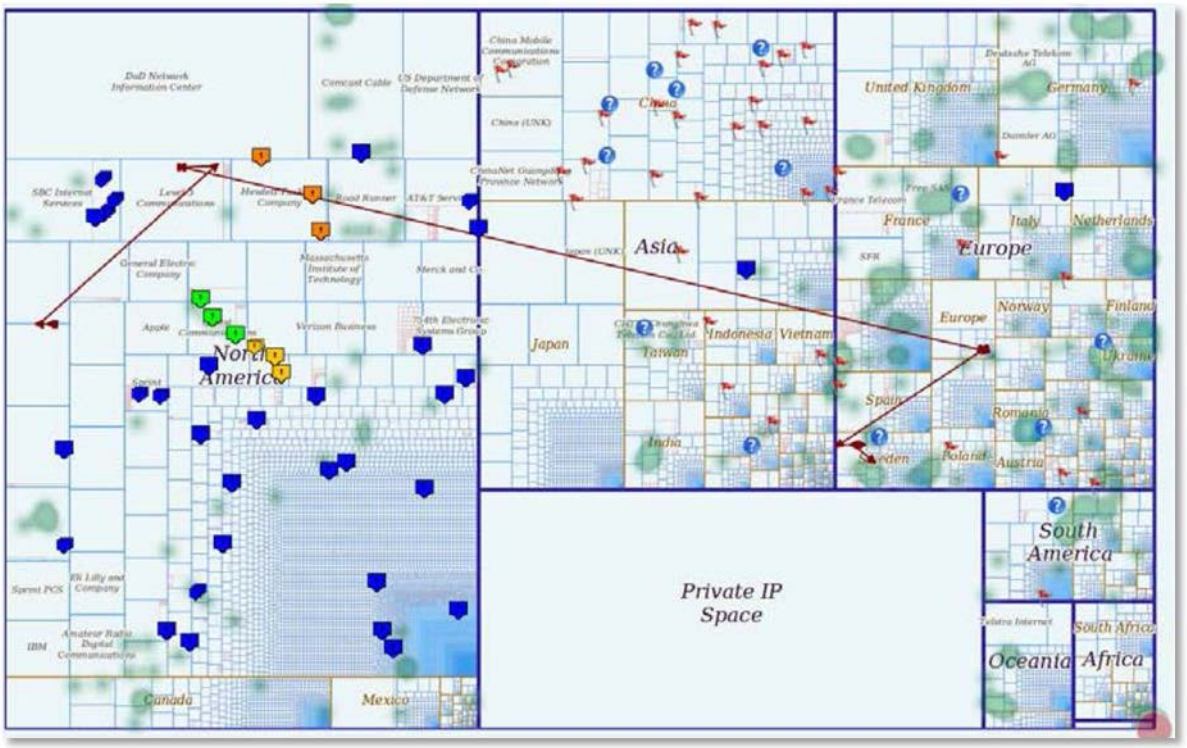

Figure 4: Global IP-space map with a hierarchy containing geospatial and non-geospatial entities and multiple data overlays that demonstrate how heterogeneous data sources can be correlated within a single data visualization to provide cyber situational awareness. Individual hosts are represented with square place marks, which can be stylized based on metadata provided by underlying data sources or analytical algorithms such as clustering. Other icons can be used to locate cyber events, such as network intrusion, or to indicate the status of specific network assets. Lines can be used to represent network connections between assets or the network path of a cyberattack.

For time-varying datasets, a tree map 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. Time variant geometries on an IP-space map always appear in the same location, which allows the user to develop familiarity with the visual layout and reduces the time required to visually analyse the data. This feature is critical to the use case of forensic analysis following a successful cyber-attack, which may include reconnaissance and preparatory activities carried out by the adversary over a period of months. A flexible, customizable display that can provide a stable frame of reference over a long period of time can be used to reconstruct a narrative of a cyberattack or a hypothesis about a situation. This can improve an organization's understanding of how the structure and usage of its systems have evolved over time and how to prioritize actions to mitigate its cybersecurity risks.

3.0 CONCLUSION

In this paper, we discuss several traditional approaches to correlating and visualizing cyber information and present a recently developed approach that we believe is a valuable tool for establishing a cyber common operating picture and supporting heightened cyber situational awareness. Hierarchical IP-space maps maintain the scalability and data agnostic features inspired by geospatial mapping yet incorporate the power of logical views to correlate extensive metadata within a consistent framework. It can also be used in combination with familiar geospatial and network graph visualizations. Furthermore, these IP-space maps provide users the flexibility to define multiple hierarchical structures through which their data can be visualized. This approach addresses many of the challenges identified for visual analytics tools in cyber defence [2]. As adoption of Internet Protocol version 6 (IPv6) spreads, this technique will need to be extended to support the new address format, the increase in address space, and the current transition period while both protocol versions are being used. IPv6 will result in a different distribution of IP addresses compared to IPv4 since it introduces stateless address autoconfiguration (SLAAC) in place of dynamic host configuration protocol (DHCP) and largely discontinues the use of NAT. It is possible that IPv6 can even improve the accuracy of geolocation since an IPv6 address is partially derived from the MAC address of the device [3].

Our implementation of the hierarchical IP-space map technique uses a modular, web-based cloud analytic framework and familiar user interface components that allow users to perform interactive analytical steps similar to other big data tools. These include the ability to "pivot" a data set between different user-defined hierarchies and "drill down" to critical information using graphical zoom capabilities. This enables users to quickly switch between different perspectives, which can be used to explore different hypotheses or changes in the data over time. We also designed an open data format called Cyber Markup Language (CML, similar to Google's KML for geospatial datasets) to specify the dataset and parameters for visualization, such as colour, position, and style. This allows users to determine how much detail they want to share with other organizations based on applicable legal or policy restrictions, such as when distributing information about incidents, threats and vulnerabilities amongst national Computer Security Incident Response Teams (CSIRTs) [13]. The CML data format also ensures consistent visualization of the data independent of any specific IP-space map implementation. CML can be used as a common data format for visual analytics use cases by integrating metadata from any data source containing IP address information. Individual organizations can easily add their own data to the visualization via additional layers without extensive data fusion processing.

While there is an increasing emphasis on automated techniques in cybersecurity for continuous monitoring and vulnerability management, we believe that effective visual analytics will continue to play a critical role in the human decision-making process. Humans still outperform machines at many visual tasks involving pattern recognition and anomaly detection and it is important to consider the role of collaborative interfaces and

interactive experiences for analytical tools and decision support systems. Visual representations are a key part of the decision process and a mechanism that can change with new input and help provide insight specifically for partially or "ill-defined questions" [14]. This is of great interest in the cyber domain and one of our approaches, the hierarchical global IP-space map, can be implemented to continually add new information to help bound and identify the cyber event being analysed.

REFERENCES

[1] U. Franke and J. Brynielsson, "Cyber situational awareness–a systematic review of the literature," Computers & Security, vol. 46, pp. 18-31, 2014.

[2] D. M. Best, A. Endert and D. Kidwell, "7 key challenges for visualization in cyber network defense," in Proceedings of the Eleventh Workshop on Visualization for Cyber Security, 2014.

[3] R. Koch, M. Golling and G. D. Rodosek, "Geolocation and Verification of IP-Addresses with Specific Focus on IPv6," Journal of Communication and Computer, vol. 11, pp. 1381-1395, 2013.

[4] MaxMind, "GeoIP2 City Accuracy," 2016. [Online]. Available: https://www.maxmind.com/en/geoip2-city-database-accuracy.

[5] Z. Hu and J. Heidemann, "Towards Geolocation of Millions of IP Addresses," in Proceedings of the ACM Internet Measurement Conference, Boston, MA, USA, 2012.

[6] I. Herman, G. Melancon and M. Marshall, "Graph visualization and navigation in information visualization: A survey," IEEE Transactions on Visualization and Computer Graphics, vol. 6, no. 1, pp. 24-43, 2000.

[7] C. Phillips and L. P. Swiler, "A graph-based system for network-vulnerability analysis," in Proceedings of the 1998 Workshop on New Security Paradigms, 1998.

[8] S. Abraham and S. Nair, "A Predictive Framework for Cyber Security Analytics using Attack Graphs," International Journal of Computer Networks & Communications (IJCNC), vol. 7, no. 1, 2015.

[9] R. Munroe, "Map of the Internet," December 2006. [Online]. Available: http://xkcd.com/c195.html.

[10] "ANT Censuses of the Internet Address Space," 2010. [Online]. Available: https://ant.isi.edu/address/index.html.

[11] F. Mansmann and S. Vinnik, "Interactive Exploration of Data Traffic with Hierarchical Network Maps," IEEE, vol. 12, no. 6, pp. 1440-1449, 2006.

[12] C. Thompson, "Anonymous to US: 'We Will Wipe You Off the Cybermap'," CNBC.com, 6 May 2013. [Online]. Available: http://www.cnbc.com/id/100712145.

[13] R. M. Ruefle and M. Murray, "CSIRT Requirements for Situational Awareness," Carnegie Mellon

Software Engineering Institute, 2014.

[14] J. C. Roberts, D. Keim, T. Hanratty, R. R. Rowlingson, R. Walker, M. Hall, Z. Jacobson, V. Lavigne, C. Rooney and M. Varga, "From Ill-defined problems to informed decisions," in EuroVis Workshop on Visual Analytics, 2014.