Designing a Pragmatic Graphical Grammar

Leonard Eusebi, Sean Guarino
Human Effectiveness Division
Charles River Analytics
Cambridge, MA

***Abstract*— Modern adversaries have become more proficient in conducting cyber attacks against our military's command and control (C2) infrastructure. To maintain security against these threats, operators perform a range of high-fidelity security assessments of existing and evolving software systems. This is just one example of the many settings in which massive amounts of data—*Big Data*—can prove difficult to understand in a timely manner for taking actions and responding to threats. To support such real-time analysis, situation awareness tools must reduce the cognitive load of monitoring multiple large, simultaneous data streams. This paper seeks to provide a Pragmatic Graphical Grammar (PGG) that evolves the concept of graphical grammars into an observability-focused method of data presentation.**

Keywords—cyber; graphical grammar; big data; situation awareness

I. Introduction

Cyber attacks produce massive amounts of data with subtle patterns and effects that can be difficult or impossible to interpret even in a post-experimental manner, let alone in real time or across multiple sessions [1]. This is just one example of the many settings in which massive amounts of data—Big Data—can prove difficult to understand in a timely manner for taking actions and responding to threats. To effectively bring this data to users, researchers have investigated graphical grammars [2]–[6] which map data to visualization approaches based purely on the type of data (e.g., using scatterplots to view floating point pairs; using network maps to view network nodes). However, while graphical grammars are useful tools for linking data to possible visualizations, they do nothing to ensure that human users can effectively recognize critical information patterns found within that data. Merely being able to view the data does not guarantee that cyber defenders (and users in other domains with big data) can find the key patterns that support the decision-making process.

In this paper, we describe an extension to graphical grammars, the Pragmatic Graphical Grammar (PGG). The PGG evolves the concept of basic graphical grammars into an observability-focused method of data presentation, providing an internal layer that identifies critical information patterns that are essential to visualize within the data, before selecting appropriate graphical approaches to highlight those information patterns. In the remainder of this paper, we provide background research on both graphical grammars and the cyber analysis domain that we used to guide our development of the PGG, followed by a detailed description of our PGG concept and its advantages over a standard graphical grammar. We conclude with a discussion of broader applications and future work.

A. Graphical Grammars

In any domain where massive data is collected, it can be difficult to effectively visualize and understand that data. Common cyber defense approaches use basic presentations of the data (e.g., raw or numerical forms or basic charts) that make the data available [7], but do not assist the analyst in understanding the data and finding critical patterns within it—that is, in observing the data [8], [9]. To be effective, cyber evaluation must make analysis results and impacts readily observable.

One specific approach that has had some success is the concept of graphical grammars. Graphical grammars [2], [3] provide an ontological map between data types and basic characteristics, and graphical visualization methods that can be used to view data of that type. In standard graphical grammars, data are characterized by their types (e.g., an array of temporally related floating point values), and the primary concern is to identify a method that is able to show that data, regardless of what patterns may be found within it (e.g., a time series chart that can show an array of temporally related floating point values). These graphical methods can be thought of as being orthogonal to one another, in the sense that every possible combination of methods is potentially representative of a meaningful graphic. Graphical grammars are highly expressive modes of representing visual information, featuring relatively few graphical features, compared to the space of possible graphics that they can generate [3].

B. Cyber Analysis Metrics

The development of cyber analysis tools begins with a suite of appropriate performance metrics to analyze cyber defense success or failure. Prior research has focused on different types of cyber metrics, including a range of general measures of effectiveness (MOEs) for cyber defenses and attacks [10], as well as a number of more specific approaches (e.g., comparison of ideal to actual outcomes [11]; goal-oriented metrics; quality of protection (QoP) metrics; and adversary-based metrics) [12]. However, many of these metrics lack the representational richness to incorporate critical contextual information, such as the attacker's level of authorization to the underlying system, relations to policy definitions, and impacts across different layers of the System Under Test SUT (e.g., network, node, and application layers). Furthermore, all of this previous research has focused on post-experimental performance assessment, not on real-time measurement of

system performance. An effective cyber evaluation capability must provide robust, contextualized metrics to assess the SUT in real time.

II. PRACTICE INNOVATION

While standard graphical grammars make data available to the user, they fail to make the data observable—that is, they fail to support the user in rapidly understanding the data and observing critical patterns within it [8], [9]. In cyber defense, observability is needed to understand the successes and failures of particular attacks, and the potential implications on the safety of the system and the adaptation of future attacks. The Pragmatic Graphical Grammar incorporates a dynamic pragmatics layer between metric results and graphical visualization methods, capturing potential types of information patterns that may be recognized within the metric results, and linking those patterns to the visualization views and adaptations that can effectively stress those patterns for the analyst. We then link this grammar to effective cyber visualization tools [13], [14], including network, timeline, statistical, and geospatial displays for network information. The PGG defines appropriate visualization methods for illustrating critical data patterns, applies adaptive features to highlight aspects of those visualizations to truly stress those patterns (e.g., making threatened nodes in a network display more salient) and to combine methods to better illustrate those patterns (e.g., using the appearance of a border to highlight threatened nodes in a network display).

Our key focus in developing the PGG was to link metrics to visualizations based on underlying information patterns that are useful for analysts, rather than data type. This basic structure is shown in Figure 1. In our PGG, metrics are mapped to an intermediate set of critical information patterns that can be found within the underlying data structures (e.g., threat patterns, benefit patterns, and activity patterns). These patterns are then linked to specific visualization methods where they identify where in the visualization to show the data and how to adapt that illustration based on the value of the data.

These patterns move beyond an understanding of the structure of the data to an understanding of the content, features, and context that the analyst needs to observe. For example, a standard graphical grammar might determine that an MOE assessment of bandwidth values, which are provided as a time-series array of numerical values, might be best represented in an x-y scatterplot, where x is time and y is the bandwidth assessment. Our internal layer, however, can identify the need to stress the pattern of information loss in the network, suggesting that the data instead be mapped to links in a temporally updated network representation. Furthermore, because this information loss is considered threatening, it can be mapped to a graphical salience feature that naturally (and culturally) draws human users to recognize the negative implications, such as making the links shift to red as the information loss increases. This visualization allows analysts to understand which aspects of the network are directly impacted by a threat and when, much faster than if they were viewing a series of charts for each link.

Fig. 1. Pragmatic Graphical Grammar with examples

III. FINDINGS

We developed an initial set of information patterns, based on discussions with cybersecurity experts about typical ways to analyze data, and then trimmed to provide a useful set of building block. These patterns are summarized in Table 1.

In our current implementation, we focus on linking these patterns to effective network visualization methods, identifying specific rendering patterns available in the network that can address the requirements of understanding the patterns. Each pattern is then directly linked to a range of actual visualization options, as discussed under Application, below. We linked available metrics from our generated data set to one or more of these information patterns to provide an initial pragmatic graphical grammar, as shown in Table 2.

TABLE I. IMPLEMENTED INFORMATION PATTERNS

Information Pattern	Display Method	Description
Importance with Increase	Network; Gantt Chart	Node importance increases with value of metric. High values should be linked to high-salience colors (e.g., yellow spectrum) or symbols.
Threat with Increase (Decrease)	Network; Gantt Chart	Node threat increases (decreases) with value of metric. High (low) values should be linked to high-threat colors (e.g., red spectrum) or symbols.
Benefit when Maxed	Network; Gantt Chart	Node is in a beneficial or good state when high, and bad state when low. High values should use beneficial colors (e.g., green spectrum) and low values should use high-threat colors.
Active Threat (Important)	Network; Gantt Chart	Boolean value that is threatening (important) when true and should then be mapped to high-threat (salience) colors / symbols when true and to nothing when false.
Inactive Benefit (Threat)	Network; Gantt Chart	Boolean value that is beneficial (threatening) when false, and should be mapped to beneficial (high-threat) colors or symbols in that case (note, this can be used in conjunction with other impacts when true).
Temporal	Timeline	Data is well-depicted in by time series visualization.
Categorical	Frequency Plot	Categorical mapping of some class of object, to value, that is effectively displayed in a bar-chart. Usually focused on post-analysis.

TABLE II. IMPLEMENTED PRAGMATIC GRAPHICAL GRAMMAR

Metric	Information Patterns	Metric	Information Patterns
Essential CPU Use	Temporal Importance with Increase	Compromise Status	Active Threat Inactive Benefit
Non-essential CPU Use	Temporal Threat with Increase	Detection Status	Active Important
Essential Memory Use	Temporal Importance with Increase	Criticality	Active Important
Non-essential Memory Use	Temporal Threat with Increase	Activity	Active Threat
Essential CPU Provided	Temporal Threat with Decrease Benefit when Maxed	Mean Time to Compromise/ Detection/ Recovery	Temporal Categorical
Essential Memory Provided	Temporal Threat with Decrease Benefit when Maxed	Successful Defenses	Importance with Increase
Latency	Temporal Threat with Increase	AAR Detection/ Recovery Rate	Categorical

Currently, this grammar is defined in a loaded preferences file that can be easily edited and updated to include new patterns.

We also identified a number of key visualization methods from prior work [13], [14] and potential associated information patterns. These are provided here in Table 3.

IV. DISCUSSION

A. Application

We created an application that uses the Pragmatic Graphical Grammar (PGG) to control network visualization views and ran it on synthetic data developed for an example scenario. Figure 2 shows a Network Map and a linked Gantt chart with a number of rendering strategies active and linked to the metrics through the pragmatic grammar. Specific active rendering strategies include:

- Node Color is mapped to Essential CPU Provided / Benefit when Maxed
- Border is mapped to Critical Threat State / Status
- Warning Symbol is mapped to Node Attack State / Status

These mappings provide a simple ability to rapidly observe the attack surface for the ongoing attack, recognizing that hosts c-host1 through c-host4 and WWW are currently compromised. Using the Metrics Table in the upper left, analysts can rapidly switch what metrics are linked to which attributes. For example, Figure 30 shows the network view with Node Color linked instead to CPU Use / Threat with Increase. Here, rather than show unaffected nodes in green, all nodes are some shade of red, based on the level of CPU use. The difference between these two charts highlights the strength of contextualized displays (0Figure 2), which incorporates context on CPU need, and only highlights those nodes that cannot fulfill that need) over non-contextualized displays (Figure 3, which merely links to absolute CPU use, which is not truly threatening until it overloads the CPU).

B. Conclusions

PGG is a powerful tool for rapid observation of complex data. Real-time analysis of multiple high-density data sources can be broken down into observable patterns that reduce the effort an analyst must spend to interpret that data. This frees up mental cycles that are better used for high-level data synthesis and decision-making.

Just as graphical grammars can be applied to any domain in which the graphical display of data is useful, the information patterns of the PGG can be applied to any form of data source and linked to available methods of visualization. This general capability represents an advance in display customization methods that provides a set of building blocks from which users can define the meaning they hope to extract from their data before choosing where and how to display that meaning. This rapid linking of meaning and visualization will allow cybersecurity operators to explore many more ways to detect and verify potential threats and vulnerabilities during real-time or after-action analysis.

TABLE III. INITIAL PRAGMATIC GRAMMAR VISUALIZATIONS

Display	Description	Shape of Metric(s)
Frequency Plot	Histogram of a particular metric using ranges or categories	A single variable or classification that falls in a set range or category (e.g., Number of defenses)
Timeline	Plot of values versus time for a given time window	A single variable whose behavior over time is relevant (e.g., Mean time to compromise)
Network Map	A map of the SUT with connections between nodes, arranged to group connected nodes closer to each other	Variables or properties which exist in multiple locations across the network or affect the connections in the network (e.g., compromised systems, disabled systems, available bandwidth on links)
Gantt Chart	A time-based display of when events occur or properties change and their durations	Any event or property change that occurs at specific times or periods of time (e.g., Privilege escalation events; attack plans and sequences)
Attack Surface View	An overlay on the network map that shows compromised nodes and the possible attack paths that originate from those nodes	Properties of attack paths or defenses against them (e.g., mission adjusted risk)
Alert View	Ordered list of prioritized alerts, showing basic alert info and allowing navigation to the relevant data	Events or measurements that indicate a need for action (e.g., attack and defense events)
Index View	Detailed, sortable data on nodes, events, services, etc. Arranged like a spreadsheet	Any data that might need sorting, comparison, or simply access to raw values (e.g., Bandwidth use)

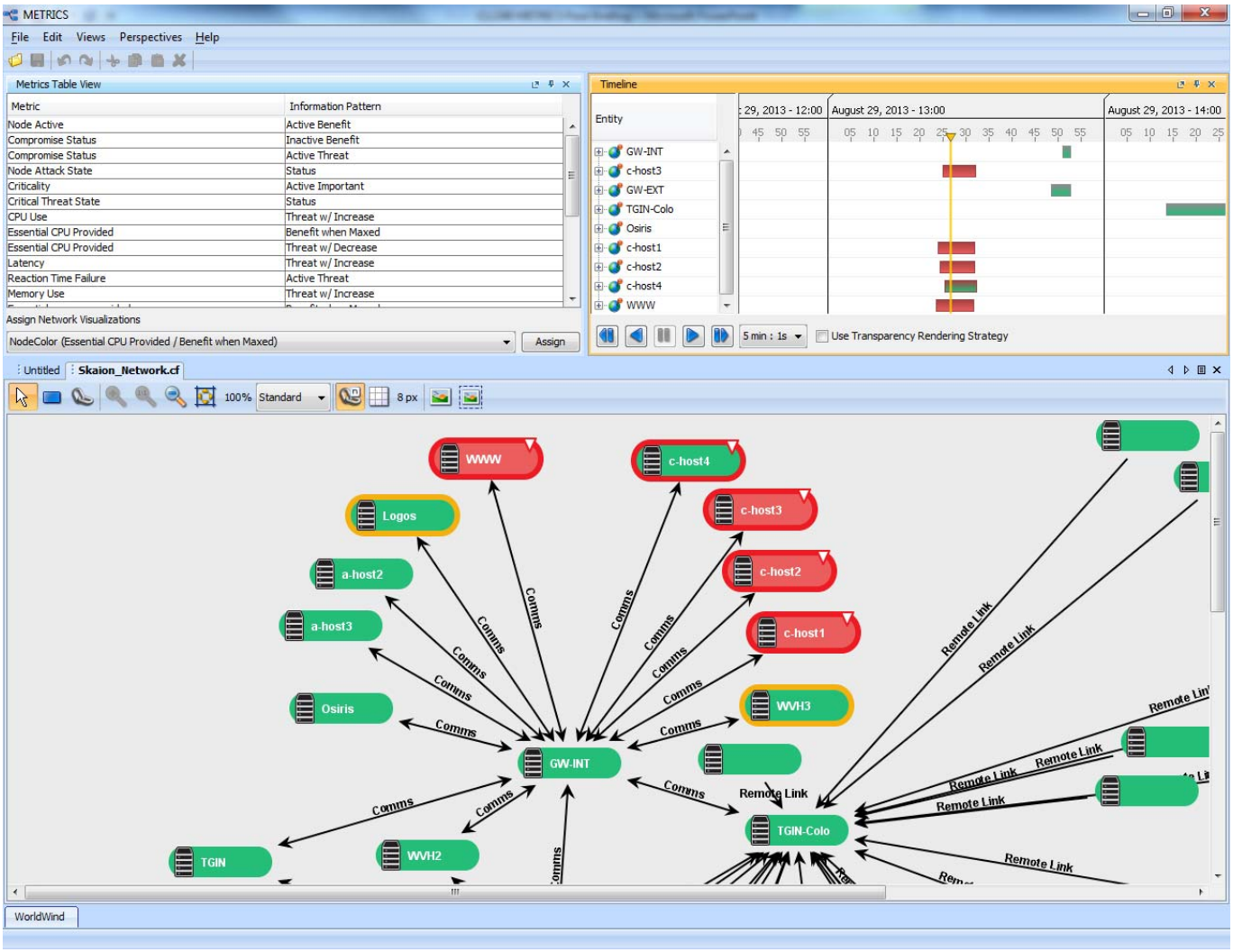

Fig. 2. METRICS network visualization with metrics linked to contextualized node characteristics shows how the context can be used to highlight important information (such as the threat from the red nodes)

C. *Future Work*

In ongoing and future work, we are refining this grammar in several ways. Initially, the focus will be on expansion and extension of the grammar to cover additional domains. While many of the information patterns identified are generally applicable, further work can be done to identify additional patterns that make the grammar more robust. Similarly, extension of the grammar to support additional visualization methods will provide new options to display data of different shapes. Finally, a user-driven capability to identify and craft new information patterns will make the grammar highly adaptable to emerging and evolving domains.

ACKNOWLEDGMENT

This material is based upon work supported by the U.S. Air Force Research Laboratory under Federal Contract Number FA8750-13-C-0080. The authors thank Dr. Asher Sinclair, Mr. Robert Grant, Mr. Tim West, and Mr. Lou Gianelli for their support and guidance in this effort.

DISTRIBUTION

REFERENCES

[1] R. Bhatti, R. LaSalle, R. Bird, T. Grance, and E. Bertino, “Emerging trends around big data analytics and security: Panel,” in Proc. of the 17th ACM symposium on Access Control Models and Technologies, 2012, pp. 67–68.

[2] E. Thomsen, "OLAP solutions: building multidimensional information systems," New York, NY: John Wiley & Sons, 2002.

[3] L. Wilkinson, "The grammar of graphics," New York, NY: Springer Science & Business Media, 2006.

[4] M. Bostock and J. Heer, “Protovis: A Graphical Toolkit for Visualization,” IEEE Trans. Vis. Comput. Graph., vol. 15, no. 6, pp. 1121–1128, Nov. 2009.

[5] H. Wickham, “A layered grammar of graphics,” J. Comput. Graph. Stat., vol. 19, no. 1, pp. 3–28, 2010.

[6] H. Wickham and W. Chang, “An implementation of the Grammar of Graphics,” R Package Version, 2013.

[7] S. Mahoney, E. Roth, K. Steinke, J. Pfautz, C. Wu, and M. Farry, “A cognitive task analysis for cyber situational awareness,” in Proc. of the HFES Annual Meeting, vol. 54, pp. 279–283, 2010.

[8] E. S. Patterson, E. M. Roth, and D. D. Woods, “Predicting vulnerabilities in computer-supported inferential analysis under data overload,” Cogn. Technol. Work, vol. 3, no. 4, pp. 224–237, 2001.
[9] D. D. Woods, E. S. Patterson, E. M. Roth, and K. Christoffersen, “Can we ever escape from data overload? A cognitive systems diagnosis,” in Proc. of the HFES Annual Meeting, 1999, vol. 43, pp. 174–178.
[10] W. R. Simpson and R. N. Meeson, “National Comparative Risk Assessment Pilot Project, Cyber Intrusion Analysis-Process Control System,” DTIC Document, 2007.
[11] M. McQueen, W. Boyer, S. McBride, M. Farrar, and Z. Tudor, “Measurable Control System Security through Ideal Driven Technical Metrics,” in Proc. S4: SCADA Security Scientific Symposium, 2008.
[12] Department of Homeland Security, “A roadmap for Cybersecurity research,” 2009. [Online]. Available: https://www.dhs.gov/sites/default/files/publications/CSD-DHS-Cybersecurity-Roadmap.pdf. [Accessed: 27-Feb-2017].
[13] S. Dudzic, A. Godwin, and R. Kilgore, “Visual strategies for enhancing user perception of task relationships in emergency operations centers,” in Proc. of the SPIE Defense, Security, and Sensing, 2010, p. 77090W–77090W.
[14] J. A. Godwin and R. M. Kilgore, “Conveying network features in geospatial battlespace displays,” in Proc. of the IEEE Symposium on Visual Analytics Science and Technology (VAST), 2010, pp. 221–222.

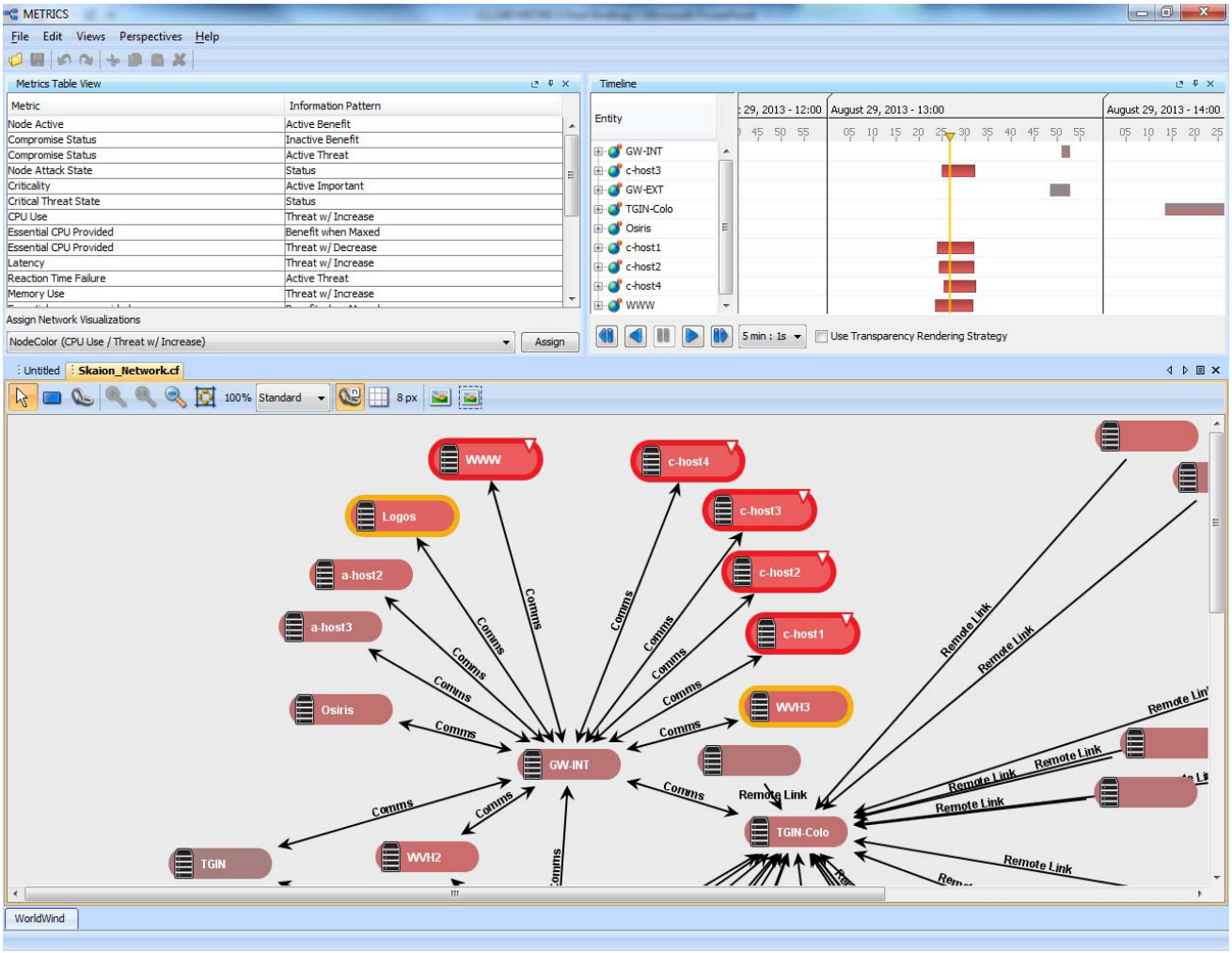

Fig. 3. METRICS network visualization with CPU Use mapped to Node Color shows how dififcult it can be to distinguish threatening levels of use from normal, non-threatening ones when viewing raw data