View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Smith College: Smith ScholarWorks

Masthead Logo

Smith ScholarWorks

Computer Science: Faculty Publications

Computer Science

2014

Visualization Evaluation for Cyber Security: Trends and Future Directions

Diane Staheli
Massachusetts Institute of Technology Lincoln Laboratory

Tamara Yu
Massachusetts Institute of Technology Lincoln Laboratory

R. Jordan Crouser
Massachusetts Institute of Technology Lincoln Laboratory, jcrouser@smith.edu

Suresh Damodaran
Massachusetts Institute of Technology Lincoln Laboratory

Kevin Nam
Massachusetts Institute of Technology Lincoln Laboratory

See next page for additional authors

Follow this and additional works at: https://scholarworks.smith.edu/csc_facpubs

Part of the Computer Sciences Commons

Recommended Citation

Staheli, Diane; Yu, Tamara; Crouser, R. Jordan; Damodaran, Suresh; Nam, Kevin; O'Gwynn, David; Harrison, Lane; and McKenna, Sean, "Visualization Evaluation for Cyber Security: Trends and Future Directions" (2014). Computer Science: Faculty Publications, Smith College, Northampton, MA.
https://scholarworks.smith.edu/csc_facpubs/94

This Conference Proceeding has been accepted for inclusion in Computer Science: Faculty Publications by an authorized administrator of Smith ScholarWorks. For more information, please contact scholarworks@smith.edu

Authors
Diane Staheli, Tamara Yu, R. Jordan Crouser, Suresh Damodaran, Kevin Nam, David O'Gwynn, Lane Harrison, and Sean McKenna

This conference proceeding is available at Smith ScholarWorks: https://scholarworks.smith.edu/csc_facpubs/94

Visualization Evaluation for Cyber Security: Trends and Future Directions

Diane Staheli
MIT Lincoln Laboratory

Tamara Yu
MIT Lincoln Laboratory

R. Jordan Crouser
MIT Lincoln Laboratory

Suresh Damodaran
MIT Lincoln Laboratory

Kevin Nam
MIT Lincoln Laboratory

David O'Gwynn
MIT Lincoln Laboratory

Lane Harrison
Tufts University

Sean McKenna
University of Utah

ABSTRACT

The Visualization for Cyber Security research community (VizSec) addresses longstanding challenges in cyber security by adapting and evaluating information visualization techniques with application to the cyber security domain. This research effort has created many tools and techniques that could be applied to improve cyber security, yet the community has not yet established unified standards for evaluating these approaches to predict their operational validity. In this paper, we survey and categorize the evaluation metrics, components and techniques that have been utilized in the past decade of VizSec research literature. We also discuss existing methodological gaps in evaluating visualization in cyber security, and suggest potential avenues for future research in order to help establish an agenda for advancing the state-of-the-art in evaluating cyber security visualization.

Keywords

Cyber security; information visualization; evaluation

1. INTRODUCTION

In cyber security, organizations rely on skilled analysts to make critical decisions regarding threats, vulnerabilities, and overall network health and performance. The fields of information visualization and visual analytics strive to leverage the unique perceptual capabilities of humans in concert with algorithmic support in order to better understand complex data. In recent years, visualization has emerged as a promising technique to better equip analysts to operate effectively in an evolving digital threat landscape.

Towards this goal, a research community that focuses on visualization for cyber security, called VizSec, was founded in 2004. The past 10 years of research in the VizSec community have led to numerous systems and techniques for analyzing security data in novel ways. However, novel cyber visualizations we have observed to date are either too complex or too basic for the intended users, or too rigid to adapt to different workflows and missions. There is little research on what makes a cyber visualization "good".

User evaluation provides a means to obtain actionable evidence of the measurable benefits of cyber visualization systems and gauge the impact of visualization tools on mission effectiveness. Iterative evaluation spirals help researchers to understand what visual support is needed, elicit user requirements, determine the efficiency, effectiveness, and utility of a visualization tool, predict end user adoption, and provide recommendations for improvement. To date, little attention has been given to comprehensive, human-in-the-loop evaluation for cyber visualization.

Evaluation not only provides measures of effectiveness and performance, but also an improved understanding of domain specific concerns (network operations, forensics, threat monitoring), tasking (data analysis, decision making, communication), work style (individual or collaborative, peer-to-peer or hierarchical), user cognition (experience, mental models, biases) and the work environment (24/7 Ops centers, contested terrains). Specific quantitative and qualitative evaluable dimensions include user experience and preference, usability and learnability, feature set utility, effect on collaboration, cognitive workload, task performance, physical demand, algorithmic efficiency, component interoperability, and insight generation.

Following previously-established methodologies from recent research in information visualization, we conducted a survey of evaluation approaches used in VizSec papers, identify gaps in the current state of the practice in evaluation, and make recommendations for future research directions.

2. PREVIOUS WORK

A number of established publication venues, such as the annual ACM Conference on Human Factors in Computing Systems (CHI), heavily emphasize the evaluation of a proposed system or visualization as part of a successful submission. Although some in the community argue that the use of usability evaluation may be ineffective in some cases [16], the importance of validating submitted work is explicitly stated in a recent year's guideline for reviewing publication submissions [8]. The guideline asks submitters to "assess the validity of the results you are presenting" through appropriate means, such as analyses, data, or evaluations, and furthermore explain why a particular method used for val-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
VIZSEC '14 Paris, France
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...$15.00.

idation is appropriate for the kind of contribution the submission claims to make. It is generally accepted by the CHI community that evaluation is a requirement for papers on system design and the absence of proving validity is often a reason for rejection [16].

The emphasis on evaluation grew over time as the CHI conference matured. In an analysis of the evaluation trend in representative papers spaning 24 years [4], the authors found that "evaluations" in the 1980s was mostly "describing the system or conceptual model in detail" that would be similar to use cases often described in VizSec papers. The authors noted that in the 1990s, usability evaluation started to gain emphasis, and more diverse evaluation techniques were used as the field expanded to include other related research fields. The use of qualitative and quantitative empirical evaluations gradually increased over the years, and by 2006 nearly all papers included some form of formal evaluation [4]. Interestingly, the authors found that evaluation techniques that are widely adopted in industry, such as Heuristic Evaluation and Cognitive Walkthrough, were not being used in the research community. They speculate that the techniques "evolve to accommodate demands of business" and may not be as readily used in academia. However, many of the visualizations and systems described in VizSec papers are designed for real users in operational settings, in which case the techniques may be appropriate and provide valuable information.

2.1 Evaluation for Visualization Research

Evaluation plays a key role in both visualization design and research. Many visualization designers, practitioners, and researchers have studied the role of evaluation for visualization, ranging from in-depth studies to formal theories and models for the design and evaluation of visualization techniques and systems. With respect to visualization for cyber security, there are many potentially useful and interesting components of these studies and models to consider for the evaluation of visualization in this field. The research on visualization evaluation can be broken down into: paper types, evaluation methodologies, models and theories, and in-depth studies on evaluation.

An important aspect to visualization research and evaluation is the review process, where reviewers must assess the merits and qualities of visualization work before it can be published. This step is vital to ensure that the methods, techniques, and evaluation included in a publication match its respective contributions. For the field of visualization, it is commonly accepted that there are five main categories of papers: technique, system, application/design study, evaluation, and theory/model [9]. These categories were originally introduced by Munzner in 2008, where she describes these different paper types in detail to identify common pitfalls among rejected submissions and provide useful tips for the expectation of evaluation within each type [24]. For example, in a technique paper, a user study is by no means required, but a well-done study could strengthen the contributions of the paper. This paper categorization scheme has since been adopted at several major visualization venues: IEEE VIS [9] and EuroVis.

Researchers have also explored, analyzed, and developed different evaluation methodologies for visualization. Tory and Muller conducted an in-depth analysis of human factors research and how visualization can support human cognition, accompanied by specific evaluation techniques that can be employed such as usability inspection and user studies [13]. In discussing the challenges of visualization evaluation, Plaisant promotes the exploration of evaluation methodologies beyond just usability studies or controlled experiments, instead emphasizing the role of data/task repositories, toolkits, and formalized case studies [33]. The evaluation methodology of Multi-Dimensional In-depth Long-term Case studies (MILCs) stresses ethnographic observation for visualization evaluation, working with domain experts for several years on larger research projects [40]. Further emphasizing the need for varied evaluation, Perer and Shneiderman refined the MILCs evaluation methodology through a series of four detailed case studies [30]. In the nine-stage framework for a methodology of design studies, Sedlmair et al. highlight that evaluation plays a critical role across all stages of visualization design, both internally for a project and externally within a community [39].

In addition to different methodologies for evaluation, visualization research has also identified new models and theories with respect to the evaluation of visualization techniques and tools. For visual analytics, Scholtz identifies and defines five key areas of visual analytic environments which she claims should be the target of varied evaluation methods and techniques: situation awareness, collaboration, interaction, creativity, and utility [37]. North creates a visualization model for measuring and capturing insight within a visualization tool, where insight has the following defining characteristics: complex, deep, qualitative, unexpected, and relevant; insight would be the step towards creating more complex benchmark tests that can be used for evaluation [27]. Carpendale distinguishes between two main kinds of evaluation in her model: quantitative versus qualitative, discussing key methods within both types and further characterizing these methodologies with respect to precision, generalizability, and realism [7]. Munzner's Nested Model highlights four different levels of visualization design in order to determine appropriate validation or evaluation methods for a visualization system [25]. To apply the approach of patterns to visualization evaluation, Elmqvist and Yi propose a naming scheme and definition for commonly-used evaluation techniques for visualization, and they categorize these patterns into five main types: exploration, control, generalization, validation, and presentation [11]. Gates and Engle recently identified potential future directions for evaluation in cyber security visualization; wherein, they promote evaluation of the data analysis process particularly through the use of case studies [13].

Lastly, two empirical studies have been conducted a systematic review of evaluation in visualization literature, which is close to our work. The original approach by Lam et al. consists of both an extensive literature review and open coding exercise on over 800 visualization publications from InfoVis, EuroVis, IVS, and VAST; the result of this review was the identification of seven main classes of evaluation scenarios [21]. They further characterize these scenarios with concise descriptions and example techniques. Overall, their work highlights a rise in the number of reported evaluations in visualization papers across the years. Isenberg et al. built upon this work by further coding all papers from the IEEE VIS conference, both information and scientific visualization. They adapted the original model for a total of eight evaluation scenarios and grouped the scenarios into those that target data analysis and those that target visualization

systems and algorithms [18]. While these reviews are thorough, neither of them addressed visualization specifically for cyber security.

2.2 Evaluation for System Design

Motivation for conducting evaluation can vary, but is generally accepted as falling into one of two categories:

- "**Informative** [or formative] motivations for assessment are to improve the quality of the system (or inform future designs) and, therefore are part of the design process.
- **Summative** motivations for assessment are to understand the qualities of the system" [14].

Formative evaluations are used by system designers and developers to test out various features and aspects of the visualization, compare alternatives, gauge user reactions, and generally shape the final direction of a product. Summative evaluation implies a level of finality and assumes a certain level of product maturity at the time a product is evaluated. As Gleicher points out, this distinction between formative and summative is not clean; a summative evaluation may still provide the basis for future design revisions. As such, summative evaluation serves as a snapshot or benchmark of the state of a visualization tool at a particular stage of development, and establishes a baseline measurement for future iterations [36].

As a formative research activity, evaluation helps us to understand what visual support is needed in the cyber domain. By evaluating cyber visualizations in the formative stages, we have the opportunity to ask questions that shed light on how visualization can best support user tasks, how data analysis is conducted, how individual work styles and experiences can impact visualization consumption, and predict how a tool will be used in a particular environment.

As a summative activity, evaluation provides the evidence for measurable operational utility of a visualization tool in a real-world environment, and gives decision-makers objective justification for investing in the development and productization of a particular artifact. Evaluation can also help us to understand the process by which research visualization transitions into operational tools [33]– it is still a long road from the lab to the watch floor. Summative evaluation of research visualization help us to understand which aspects of novel visualizations are useful to users, and how such aspects should be iterated upon until they are operations-ready.

3. WHAT CONSTITUTES EVALUATION?

While there is little disagreement in the security visualization community about the importance of evaluation, there is no general consensus on what constitutes an evaluation. For an evaluation to be useful, one must consider its purpose and scope, select the appropriate metrics and correctly apply assessment techniques. As previously noted, much work has been done on evaluation in other visualization fields. This section collects and systematizes ideas and taxonomies from prior research. Section 3.1 discusses the different dimensions one might consider evaluating. Section 3.2 covers the different components in a visualization system one might consider instrumenting and evaluating. Note that we treats a visualization system broadly to include the computer, the human and any environmental factors that impact the interaction between a visualization and its intended user. Section 3.3 enumerates a range of techniques. It is our goal to provide a "lay of the land" for evaluation. While evaluations need not to be complete or even always necessary, we hope this helps security visualization practitioners to construct evaluations that get at the aspects most important for their situations.

3.1 Dimensions Evaluated

The following section identifies dimensions of a visualization system that may be useful to evaluate, ranging from human performance to system performance. These dimensions were compiled from a variety of existing work in evaluating human-machine collaborative systems [7, 16, 18, 25].

- **User experience / preference:** The overall experience of using an interface, especially in terms of how easy or pleasing it is to use
- **Usability / learnability:** The ease of use and learning curve of an interface
- **Effect on collaboration:** Does an interface encourage more collaboration (measured in terms of increased communication, shared control of resources, etc.)?
- **Insight generation:** Does using this system enable more "aha!" moments? Note: this is notoriously difficult to measure; usually we ask the person to self-report their insights and count / otherwise aggregate them
- **Cognitive workload:** From a cogsci perspective: how effectively does the system utilize a person's working memory? More heuristically: how hard does the person have to think to accomplish their tasks while using the system?
- **Task performance:** How well does a person / team perform on a predefined task using this system?
- **Physical demand:** Physiological measures: how hard does a person's body have to work to use the system effectively?
- **Feature set utility:** How useful or advantageous are the set of features available? Are there features that get used more heavily, others than never get used at all?
- **Algorithmic efficiency:** Traditional algorithmic / empirical system performance measures
- **Component interoperability:** How well do the pieces of the system fit together? Are they independent, do they interact with one another? Can they be rearranged?

3.2 Components

Adapting a systems approach, a visualization system encompasses the machine, the human and the interaction of the two. Pike et al. [32] introduce a model that maps a taxonomy of the user's goals and tasks to a taxonomy of the visualization application's representation and interactive controls. This is an excellent basis for constructing summative evaluations to answer the question of how effective a visualization is in helping the user accomplish her

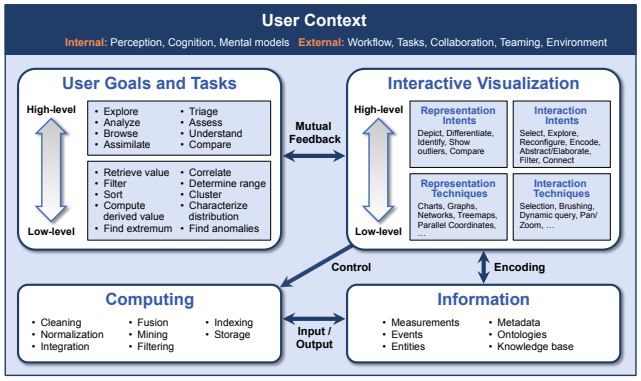

Figure 1: Components of a visualization system that can be evaluated. This model extends existing work by Pike et al. [32] and Sedig et al. [38] to include components of human internal processes and external factors that influence human behavior.

tasks. Sedig et al. [38] further broaden the definition of interactivity in terms of a linear transfer of information from *information space, computing space, representation space, interaction space,* to *mental space*. We prefer to define a visualization system as a system where a user interacts with a visualization application that encodes information, within the context of the user's tasks and goals as well as her internal and external states. The information can be collected data, knowledge base or transformed data resulted from computational processes. The transformation may or may not be directly dictated by the application. We observe that the most effective users are cognizant of the nuances of the data's provenance as well as the data as presented by the application, so that they understand the limitations and biases and form their judgements accordingly. There are other background forces that influence how a user perceives and interacts with a visualization. These may be internal, such as cognition and mental models, or external, such as physical or organizational environment and workflows. When conducting summative evaluations, one ought to keep these contexts in mind, whether to assess their effects or leverage them to interpret results. Finally, there are great opportunities to conduct formative evaluations on these context elements, as they help inform the needs that potential visualization applications could meet and constraints such applications must respect to be effective.

3.3 Techniques

In this section, we outline some commonly-used techniques for evaluating visualization. As we survey the VizSec literature, we will use these definitions to categorize the evaluations performed.

- **Critique:** Holistic assessment of a human reading and interpretation of the visualization via "meticulous group discussion centered on how well particular aspects or details of a visualization support the intended goal" [19].

- **Co-Creation:** Participatory design methodology where users are actively involved in the process of creating a system, and support iterative evaluation throughout the project lifecycle [23], thus being positioned as system co-creators. The Living Laboratory framework is a variation in which users are co-located with researchers for the purpose of ethnographic study, iterative experimentation and evaluation [22].

- **Inspection:** Set of informal methods in which experts review an interface to determine adherence to a set of best practices or user interface guidelines for a given domain. Example methods include heuristic evaluation, cognitive walkthrough, formal usability inspection, pluralistic walkthrough, feature inspection, consistency inspection, and standards inspection [26].

- **Interview:** User is asked a series of structured or semi-structured questions to elicit knowledge regarding a particular topic, domain, or workplace [6, 43].

- **Usability Testing:** Measurement of "the extent to which the product can be used with effectiveness, efficiency and satisfaction in a particular context" [2]. Data collected can include completion rates, errors, tasks times, usability problems and qualitative feedback [36].

- **Surveys:** Compilations of questions, consisting of quantitative or qualitative rating or open-ended free response questions, aimed at extracting information from a representative sample of a target population, usually for the purpose of user evaluations, opinions, or demographics [28].

- **Longitudinal Studies:** Research conducted over a period of time to measure user performance, efficiency, and utility of a particular tool from initial user training through proficiency. This method also captures process or behavioral changes as a result of introducing a new product into an environment. Multiple methodologies may be used [40].

- **Simulation:** Controlled experiments performed with representative users in a laboratory environment that incorporate multiple variables and realistic scenarios and conditions.

- **Interface Instrumentation:** "Application instrumentation, collection of usage data, and data analysis" to extract user interaction information from software tools [5]. Examples include log file analysis or click-tracking.

- **Psychophysiological Measurement:** Physiological measurements taken during a user's interaction with a product that are indicative of a cognitive state, such as attention, cognition, or emotional response. Examples include eye tracking, pupil dilation, heart rate, respiration, skin conductance, muscle activity, and brain measurements [35].

- **Automated Image Analysis:** Computer-generated analysis of a digital image for visual characteristics such as consistency of rendering, visual density, complexity, element alignment, ratio of text-to-background or graphics-to-text, balance, and symmetry, as a proxy for human evaluation [42, 44].

- **Application Performance Testing:** Automated or computer-generated analysis of system load or response times under particular use conditions, based on predefined scenarios.

4. SURVEY METHODOLOGY

We surveyed 130 papers from the past 10 years of VizSec proceedings, and from this corpus, we have identified 49 papers that included some form of evaluation according to the criteria and definitions outlined in Section 3. In addition to this categorization, we also surveyed whether or not users were involved in the evaluation process, whether or not they were expert users, and at what point during the development process they were involved. The raw results of this analysis (by year) is outlined in Figure 2. Our analysis has uncovered key patterns of evaluation within the VizSec community, as well as several interesting methodological gaps, which will be discussed in detail in Section 5. The results of this analysis provide a common framework for describing and understanding the current state of the practice in evaluation in cyber security, which we hope will motivate future work in the field.

5. SURVEY FINDINGS

Evaluation involves several multi-faceted choices, each with significant tradeoffs made by the evaluators. Yet throughout our analysis we encountered cases where key details of evaluations were missing. The purpose of this section, therefore, is to support future evaluation by outlining several common choices and tradeoffs.

5.1 Users: Experts or Everyone Else?

Our analyis indicates that the choice of users in security visualization evaluation is often all-or-nothing: either expert users are recruited (32% of evaluations), or no users are recruited at all (46%). Only rarely are non-expert users involved (10%), and in some cases, user details were omitted entirely (12%).

As the target users of a system, expert users provide valuable feedback for many of the metrics listed in Section 3.1, such as insight generation and usability. However, our analysis indicates that previous studies involving expert users focus overwhelmingly on feature set utility (73%) and usability (42%), whereas other metrics are rarely visited, such as insight generation (23%), component interoperability (11%), and cognitive workload (3%). Future studies can make valuable contributions by examining these under-explored metrics with expert users.

The under-utilized category of non-expert users (10% of evaluations), may yield many benefits for the future of VizSec. For example, Ball et al. recruited non-expert users for a thorough usability- and performance-focused evaluation of their system, which used a set of network administration tasks informed by their previous collaborations with expert users [3]. Similarly, future research in security visualization might focus on distilling common analytical tasks into more their more basic perceptual, cognitive, and motor substrates, which will make it possible to conduct empirical comparative evaluations of visualization and interaction techniques with non-expert users.

5.2 Data: Real or Repeatable?

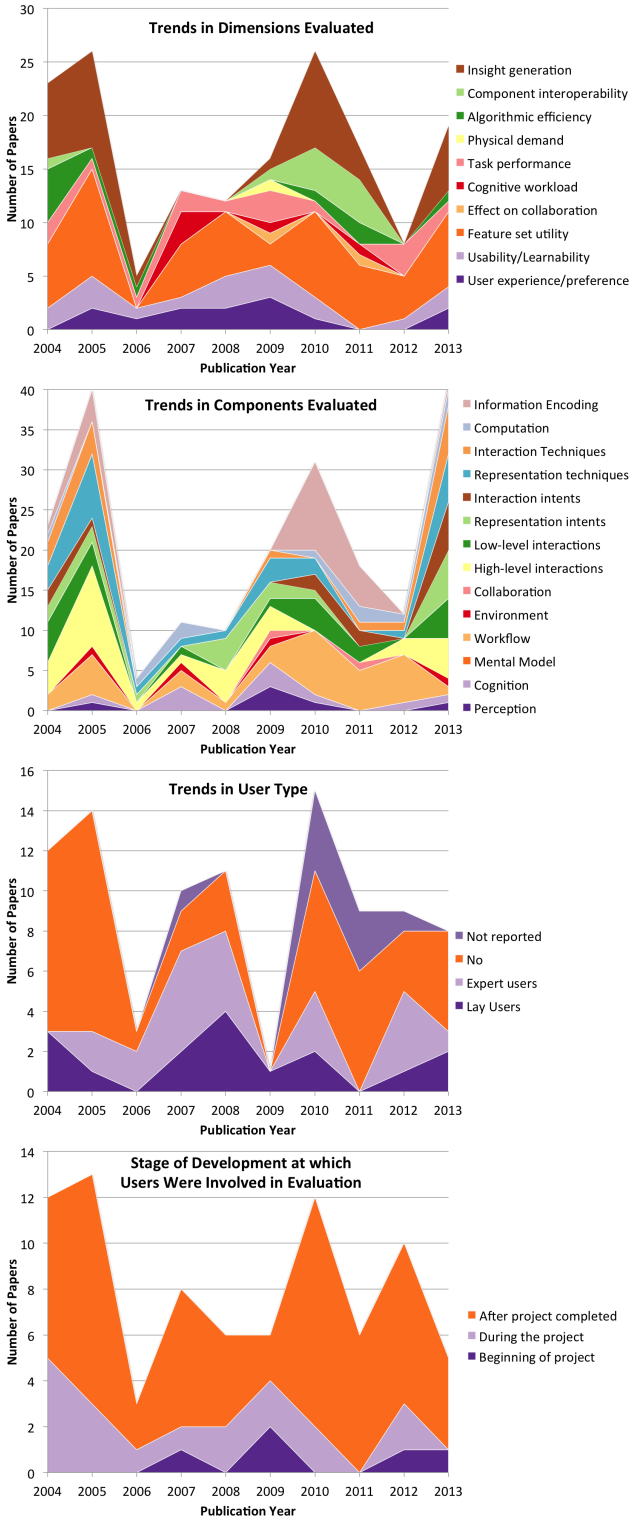

Figure 2: Annual trends in the evaluation of various dimensions of visualization systems over the past decade of VizSec.

Datasets used in system and technique evaluations should reflect real-world scenarios as much as possible, while simultaneously lending themselves to reproducibility and com-

parison in future research. Balancing these goals is difficult and sometimes impossible, making it necessary to understand the tradeoffs between using different types of datasets for evaluation.

Datasets that are obtained through collaborations with real-world users often make compelling case studies, since they can reveal previously undiscovered insights within an organization. Yet real-world datasets are rarely published, even in anonymized form, making meta-analysis in follow-up research difficult. The size and complexity of these datasets are also uncontrolled, and cannot be assumed to be equivalent to datasets in other organizations. The use of real-world datasets in evaluation, therefore, should include an adequate description of its characteristics to support followup research.

As an alternative to real-world datasets, several organizations have crafted open datasets in order to facilitate comparative evaluations between security visualization tools. The VAST Challenge program, for instance, has provided system logs, NetFlow data, packet-capture data, vulnerability scans, along with complex scenarios and ground truth data. For some visualizations, however, even these datasets may be limited in terms of data complexity and size. One way to directly control the size and complexity of data is to use simulation tools and environments. By controlling the size and complexity of data, it is possible to directly test the limits of the chosen visualization techniques, encodings, and interactions.

5.3 Evaluation using Case Studies

Case studies ground the evaluation of visualization tools into realistic settings [33]; however, many VizSec papers utilize the term "case studies" when actually a more apt term would be a "usage scenario". As characterized in a systematic study of evaluation by Isenberg et al., case studies can be classified into four main types [18]:

1. how a domain expert or analyst used a tool
2. how a tool is developed based on collaboration between the visualization researcher and domain expert
3. how a visualization researcher used a tool to solve a real problem
4. documentation or demonstration of how a tool may be utilized

In their systematic study, the authors argue that this fourth type is not a formal case study. To classify as a more formal case study, the study must involve both real data and real users [39], otherwise it is a usage scenario. While the first three categories clearly involve evaluation of a tool, the last category is not as strong and merely a demonstration of the tool. In our analysis of VizSec papers from 2004 through 2013, we found all of these cases present, with a clear dominance of usage scenarios over case studies.

In our review of VizSec papers, the majority contained usage scenarios over case studies. For example, since 2006, there have been a total of 44 papers containing any one of these four types of studies, but only six of these evaluated using a more formal case study (see Fig. 3). In fact, for each year since 2006, there has not been more than one formal case study presented at the conference. Many authors of the many VizSec papers are not consistent in their use of

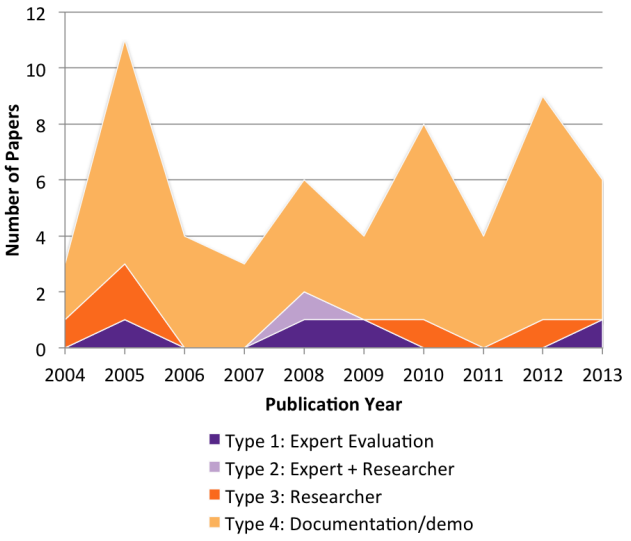

Figure 3: Annual trends in the utilization of four types of Case Studies. Note that Type 4 (documentation or demonstration of how a tool may be utilized) is the most heavily utilized.

these terms. Most usage scenarios are reported in a section called "case study," and others have reported their formal case studies as other names, such as an "example analysis session" [17]. Usage scenarios can serve a purpose, but it is important to note that the lack of connection back to real users or real data (or both) may question the validity and utility of the evaluated tool. The preponderance of usage scenarios as a type of evaluation method in these papers seems to favor new techniques and tools even if there are not enough user-based studies that suggest the tool or technique will be useful in the field.

5.4 Technique: Tying it Together

The choice of evaluation technique depends on the overall goal of evaluation, as well as the available users and datasets. Although our survey found that most VizSec evaluations make use of only the *usage scenario* technique (48%), there are several notable exceptions that employ less common techniques and metrics, or combinations of techniques. For example, Fink et al, involved target users throughout the system design and evaluation through interviews and co-creation, before evaluating the performance of their system using more controlled task-based experiment techniques [12]. Similarly, Rasmussen et al use a combination of structured feedback protocols and surveys to assess how their system met analyst's needs for defensible insights and recommendations [34]. These examples underscore the need for future security visualization research to explore the utility of currently unexplored evaluation techniques.

6. FUTURE DIRECTIONS

The results of our survey uncovered several gaps and trends in the evaluation of security visualizations. We discuss these gaps in this section, along with several possible directions for future research.

6.1 Common Framework

In many cases, identifying evaluation methodologies used was difficult for the purposes of categorization, as limited detail was provided by the authors. This limitation impacted our ability to generate meaningful findings beyond discussion of the techniques utilized. We can speculate as to reasons why this may be the case. Evaluation is not required of a VizSec paper and therefore may be deprioritized by the authors. Cyber visualization is a niche practice; an individual researcher may have a highly specialized skill set that may not include standard HCI experimentation practices or vocabulary. While the scientific method for visualization does not necessarily require exact reproducibility of results, exposing further detail on techniques and metrics used would benefit the visualization community by setting expectations for similar results using similar techniques. To help facilitate continued dialogue on visualization evaluation in the VizSec community, we recommend the adoption of a common framework for discussion of evaluation such as (but not limited to) the one outlined in Section 3.

6.2 Psychophysiological Methods

Our analysis found that no papers used physiological methods for evaluating security visualizations (see Fig. 4). Yet given the sustained focus in the security community on topics such as situational awareness and information overload, existing research in physiological techniques from visualization and human-computer interaction present valuable new dimensions for security visualization evaluation. For example, recent research in brain-sensing from Peck et al. [29] measured changes in cognitive workload to evaluate basic visualization tasks. These changes in workload were reliably measured in real-time, even when traditional metrics like participant accuracy and user preference showed little change. Similarly Afergan et al. have used brain-sensing not only for detecting cognitive overload in a visual interface, but also for adapting to the user by decreasing or increasing the amount of information operators must analyze [1]. Physiological metrics include more than just brain-sensing, however. Examples of other well-researched physiological methods in visualization and human-computer interaction include eye-tracking [41], as well as galvanic skin-response, heart-rate, and posture/orientation [31].

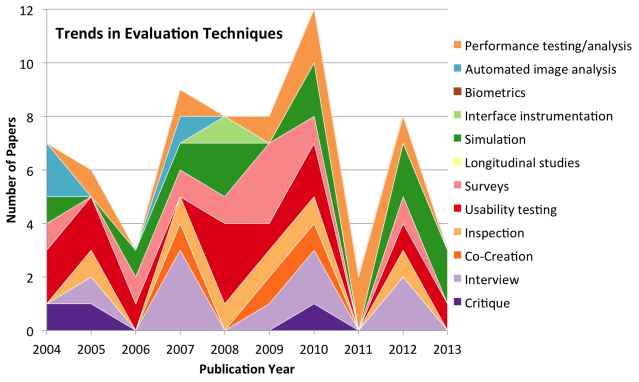

Figure 4: Annual trends in the utilization of various evaluation techniques.

6.3 Interface Instrumentation

While several papers in our analysis used observational protocols to analyze how participants interact with a visualization system, only one paper used interface instrumentation, where system interactions are logged and analyzed as part of the evaluation [20]. Logging and analyzing interactions presents several new directions for the evaluation and design of security visualization systems. Research in visual analytics has demonstrated that low-level interaction logs can be mined to infer information about a user's strategies, methods, and findings [10]. Similarly, Gotz et al. showed that low-level interactions can be mapped into existing visualization task taxonomies to evaluate and compare how well tools supported common data analysis tasks [15]. Adapting methods like these will lead to more quantitative, scalable, and repeatable approaches for security visualization evaluation.

6.4 Longitudinal Studies

No papers in our study utilized longitudinal study as an evaluation method. The method, as defined by Shneiderman and Plaisant [40], combines ethnographic observation in the normal user environment, automated activity logging, and intense engagement with researchers over a long period of time. This gap in utilization in VizSec is not surprising considering the challenges inherent in the cyber security user research. Firstly, access to environments where cyber security analysis activities are taking place is often tightly controlled to protect the security and privacy of the organization; it is difficult for researchers to gain access to analysts or watch floors for direct observation for even short periods of time. Audio or video recording, activity logging, or sharing of meaningful findings with the research community is often restricted, if not completely prohibited as a matter of policy. Finally, it is difficult for management to justify analyst time spent working with closely researchers as opposed to their daily job responsibilities, as there has been little evidence of return on investment for an organization's participation in this type of study. However, the "living laboratory" concept [22] – pairing researchers with analysts in a hybrid operations-research environment – is starting to gain traction with research universities and national laboratories; we look forward to future results and lessons learned from these collaborations.

7. CONCLUSION

In this work, we have outlined the "lay of the land" for visualization evaluation, as well as surveyed and categorized the evaluation metrics, components and techniques that have been utilized in the past decade of VizSec research literature. We have identified existing methodological gaps in evaluating visualization in cyber security, and suggested potential avenues for future research. It is our hope that this study will help establish an agenda for advancing the state-of-the-art in evaluating cyber security visualization, as well as encourage future dialogue on evaluation for operational utility.

8. ACKNOWLEDGMENTS

The authors would like to thank Andrea Brennen, Matt Daggett, Jeffrey Gottschalk, and Andy Vidan for their early contributions to this study.

This work is sponsored by the Assistant Secretary of Defense for Research & Engineering under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government.

9. REFERENCES

[1] D. Afergan, E. M. Peck, E. T. Solovey, a. Jenkins, S. W. Hincks, E. T. Brown, R. Chang, and R. J. Jacob. Dynamic difficulty using brain metrics of workload. In *Proc. 32nd annual ACM Conf. on Human Factors in Computing Systems*, pages 3797–3806. ACM, 2014.

[2] W. Albert and T. Tullis. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Newnes, 2013.

[3] R. Ball, G. A. Fink, and C. North. Home-centric visualization of network traffic for security administration. In *Proc. 2004 ACM Workshop on Visualization and Data Mining for Computer Security*, pages 55–64. ACM, 2004.

[4] L. Barkhuus and J. A. Rode. From mice to men-24 years of evaluation in chi. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*. ACM, 2007.

[5] S. Bateman, C. Gutwin, N. Osgood, and G. McCalla. Interactive usability instrumentation. In *Proc. 1st ACM SIGCHI Symposium on Engineering Interactive Computing Systems*, pages 45–54. ACM, 2009.

[6] H. Beyer and K. Holtzblatt. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[7] S. Carpendale. Evaluating information visualizations. In *Information Visualization*, pages 19–45. Springer, 2008.

[8] Guide to a successful archive submission. Online, July 2014.

[9] V. P. Committee. Paper submission guidelines: Paper types. Online, jul 2014.

[10] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang. Recovering reasoning process from user interactions. *IEEE Computer Graphics & Applications*, 2009.

[11] N. Elmqvist and J. S. Yi. Patterns for visualization evaluation. *Information Visualization*, page 1473871613513228, 2013.

[12] G. A. Fink, C. L. North, A. Endert, and S. Rose. Visualizing cyber security: Usable workspaces. In *Visualization for Cyber Security, 6th Int'l Workshop on*, pages 45–56. IEEE, 2009.

[13] C. Gates and S. Engle. Reflecting on visualization for cyber security. In *Intelligence and Security Informatics (ISI), 2013 IEEE Int'l Conf. on*, pages 275–277. IEEE, 2013.

[14] M. Gleicher. Why ask why?: considering motivation in visualization evaluation. In *Proc. 2012 BELIV Workshop: Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, page 10. ACM, 2012.

[15] D. Gotz and M. X. Zhou. Characterizing users' visual analytic activity for insight provenance. *Information Visualization*, 8(1):42–55, 2009.

[16] S. Greenberg and B. Buxton. Usability evaluation considered harmful (some of the time). In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pages 111–120. ACM, 2008.

[17] L. Hao, C. G. Healey, and S. E. Hutchinson. Flexible web visualization for alert-based network security analytics. In *Proc. 10th Workshop on Visualization for Cyber Security*, pages 33–40. ACM, 2013.

[18] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Moller. A systematic review on the practice of evaluating visualization. *Visualization and Computer Graphics, IEEE Trans. on*, 19(12):2818–2827, 2013.

[19] B. Jackson, D. Coffey, L. Thorson, D. Schroeder, A. M. Ellingson, D. J. Nuckley, and D. F. Keefe. Toward mixed method evaluations of scientific visualizations and design process as an evaluation tool. In *Proc. 2012 BELIV Workshop: Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, page 4. ACM, 2012.

[20] T. Jankun-Kelly, J. Franck, D. Wilson, J. Carver, D. Dampier, and J. E. Swan Ii. Show me how you see: Lessons from studying computer forensics experts for visualization. In *Visualization for Computer Security*, pages 80–86. Springer, 2008.

[21] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *Visualization and Computer Graphics, IEEE Trans. on*, 18(9):1520–1536, 2012.

[22] M. D. McNeese, K. Perusich, and J. R. Rentsch. Advancing socio-technical systems design via the living laboratory. In *Proc. Human Factors and Ergonomics Society Annual Meeting*, volume 44, pages 2–610. SAGE Publications, 2000.

[23] M. J. Muller. Participatory design: the third space in hci. *Human-Computer Interaction: Development Process*, pages 165–185, 2003.

[24] T. Munzner. Process and pitfalls in writing information visualization research papers. In *Information Visualization*, pages 134–153. Springer, 2008.

[25] T. Munzner. A nested model for visualization design and validation. *Visualization and Computer Graphics, IEEE Trans. on*, 15(6):921–928, 2009.

[26] J. Nielsen. Usability inspection methods. In *Conf. Companion on Human Factors in Computing Systems*, pages 413–414. ACM, 1994.

[27] C. North. Toward measuring visualization insight. *Computer Graphics and Applications, IEEE*, 26(3):6–9, 2006.

[28] A. A. Ozok. Survey design and implementation in hci. *Human-Computer Interaction: Development Process*, page 253, 2009.

[29] E. M. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pages 473–482. ACM, 2013.

[30] A. Perer and B. Shneiderman. Integrating statistics and visualization: case studies of gaining clarity during exploratory data analysis. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pages 265–274. ACM, 2008.

[31] R. W. Picard and S. B. Daily. Evaluating affective interactions: Alternatives to asking what users feel. In *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches*, pages 2119–2122, 2005.

[32] W. A. Pike, J. Stasko, R. Chang, and T. A. O'Connell. The science of interaction. *Information Visualization*, 8(4):263–274, 2009.

[33] C. Plaisant. The challenge of information visualization evaluation. In *Proc. Working Conf. on Advanced Visual Interfaces*, pages 109–116. ACM, 2004.

[34] J. Rasmussen, K. Ehrlich, S. Ross, S. Kirk, D. Gruen, and J. Patterson. Nimble cybersecurity incident management through visualization and defensible recommendations. In *Proc. Seventh Int'l Symposium on Visualization for Cyber Security*, pages 102–113. ACM, 2010.

[35] N. Riche et al. Beyond system logging: human logging for evaluating information visualization. In *Proc. BELIV 2010 Workshop: Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, 2010.

[36] J. Sauro and J. R. Lewis. *Quantifying the user experience: Practical statistics for user research*. Elsevier, 2012.

[37] J. Scholtz. Beyond usability: Evaluation aspects of visual analytic environments. In *Visual Analytics Science and Technology, 2006 IEEE Symposium On*, pages 145–150. IEEE, 2006.

[38] K. Sedig, P. Parsons, and A. Babanski. Towards a characterization of interactivity in visual analytics. *JMPT*, 3(1):12–28, 2012.

[39] M. Sedlmair, M. Meyer, and T. Munzner. Design study methodology: Reflections from the trenches and the stacks. *Visualization and Computer Graphics, IEEE Trans. on*, 18(12):2431–2440, 2012.

[40] B. Shneiderman and C. Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proc. BELIV 2006 Workshop: Beyond Time and Errors: Novel Evaluation Methods for Information Visualization*, pages 1–7. ACM, 2006.

[41] B. Steichen, G. Carenini, and C. Conati. User-adaptive information visualization: using eye gaze data to infer visualization tasks and user cognitive abilities. In *Proc. 2013 Int Conf. on Intelligent User Interfaces*, pages 317–328. ACM, 2013.

[42] M. Wattenberg and D. Fisher. Analyzing perceptual organization in information graphics. *Information Visualization*, 3(2):123–133, 2004.

[43] L. E. Wood. Semi-structured interviewing for user-centered design. *Interactions*, 4(2):48–61, 1997.

[44] X. S. Zheng, I. Chakraborty, J. J.-W. Lin, and R. Rauschenberger. Correlating low-level image statistics with users-rapid aesthetic and affective judgments of web pages. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*, pages 1–10. ACM, 2009.