The Battle for New York: A Case Study of Applied Digital Threat Modeling at the Enterprise Level

Rock Stevens,* Daniel Votipka, Elissa M. Redmiles,† and Michelle L. Mazurek
University of Maryland
rstevens,dvotipka,eredmiles,mmazurek@cs.umd.edu

Colin Ahern
NYC Cyber Command
colin@cyber.nyc.gov

Patrick Sweeney
Wake Forest University
sweenepj@wfu.edu

Abstract

Digital security professionals use threat modeling to assess and improve the security posture of an organization or product. However, no threat-modeling techniques have been systematically evaluated in a real-world, enterprise environment. In this case study, we introduce formalized threat modeling to New York City Cyber Command: the primary digital defense organization for the most populous city in the United States.

We find that threat modeling improved self-efficacy; 20 of 25 participants regularly incorporated it within their daily duties 30 days after training, without further prompting. After 120 days, implemented participant-designed threat mitigation strategies provided tangible security benefits for NYC, including blocking 541 unique intrusion attempts, preventing the hijacking of five privileged user accounts, and addressing three public-facing server vulnerabilities. Overall, these results suggest that the introduction of threat modeling can provide valuable benefits in an enterprise setting.

1 Introduction

Threat modeling — a structured process for assessing digital risks and developing mitigation strategies — originated more than 30 years ago and is commonly recommended in industry and academia as a useful tool for mitigating risk in software, systems, and enterprises [57]. While a number of threat-modeling approaches have been proposed, few provide efficacy metrics, and essentially none have been systematically evaluated in an enterprise environment [9, 14, 15, 20, 24, 25, 28, 34, 35, 37, 38, 42, 46, 53]. As a result, it can be difficult to quantify the benefit of threat modeling in practice.

In this paper, we present the first case study of threat modeling in a large, high-risk enterprise environment: New York City Cyber Command (NYC3).[1] NYC3 is responsible for defending the most populous city in the United States from cyber attacks, including a digital infrastructure that supports 60 million visitors and 300,000 government employees each year.

Similar to many other enterprise organizations, prior to our study, NYC3 did not use threat modeling but protected its assets primarily via vendor technologies meeting city-specific and industry guidelines. As part of a unique cooperative opportunity, we introduced 25 NYC3 personnel to an exemplar threat-modeling approach through group training sessions. We then tracked the impact of this threat modeling training on NYC3's security posture quantitatively, through analysis of 120 days of log data, and qualitatively, via pre-, post-, and 30-day-post-training surveys with participants. To our knowledge, this represents the largest-scale real-world evaluation of threat modeling efficacy to date.

Our results suggest that threat modeling may provide valuable benefits in an enterprise setting. Participants' perceptions of threat modeling were very positive: after 30 days, 23 participants agreed that it was useful in their daily work and 20 reported that they have adopted its concepts in their daily routine. Collectively, participants developed 147 unique mitigation strategies, of which 64% were new and unimplemented within NYC3. Additionally, participants identified new threats in eight distinct areas within their environment, such as physical access-control weaknesses and human configuration errors. Within one week of developing these plans, NYC3 employees started implementing participant-designed plans to mitigate these eight newly-identified threat categories. In the 120 days following our study, NYC3 implemented participant-designed defensive strategies that prevented five privileged account hijackings, mitigated 541 unique intrusion attempts, and remedied three previously unknown web-server vulnera-

*We would like to thank the leadership and strategic communications personnel of NYC Cyber Command for making this study possible. Additionally, we would like to thank Lujo Bauer of Carnegie Mellon University for his advice and expertise in shaping this study.

†Elissa Redmiles acknowledges support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1322106 and a Facebook Fellowship.

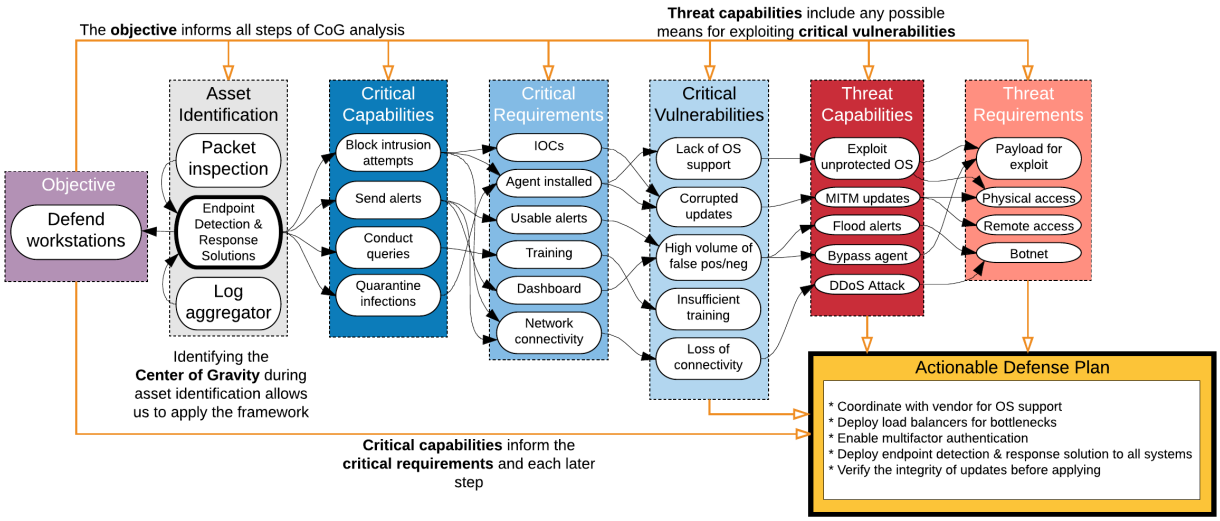

Figure 1: Step-by-step process for threat modeling with CoG, using participant P17's responses as an example.

bilities.

While our findings are drawn from a single enterprise environment, NYC3 shares many similarities to many U.S. enterprises today, such as the use of widely-mandated compliance standards [29, 44, 45], use of pervasive vendor technologies, and the mission to protect a spectrum of organizations ranging from the financial sector to law enforcement [13].[2] Consequently, our observations and metrics provide a scaffolding for future work on threat modeling and enterprise-employee security training.

2 Background

In this section, we describe threat modeling, detail the specific threat-modeling approach we used in this study, and briefly review prior empirical studies of threat modeling.

2.1 Threat-modeling frameworks

Threat modeling is a structured approach to assessing risks and developing plans to mitigate those risks. Many threat-modeling frameworks aim to improve practitioners' situational awareness and provide them with a decision-making process for complex problems [15, 25]. Some frameworks focus on thinking like an adversary, helping practitioners identify and block essential tasks that would lead to a successful attack [9,14,28,43]. Other frameworks help users automatically or manually identify likely threats to a particular system based on past data and ongoing trends [38, 39, 53, 54, 57].

2.2 The Center of Gravity framework

In this study, we introduced NYC3 employees to the Center of Gravity (CoG) framework, which originated in the 19th century as a military strategy [64]. As a military concept, a center of gravity is the "primary entity that possesses the inherent capability to achieve the objective [17]." As a threat modeling approach, CoG focuses on identifying and defending this central resource. This approach is applicable within any contested domain [60] and is synonymous with centrality, which appears in network theory for social groups [30] and network theory in the digital domain [62]. CoG supports planning offensive cyberspace operations [8] and prioritizing digital defenses [11].

The constraints of our partnership with NYC3 — in particular, the requirement to minimize employees' time away from their duties — only allowed us to introduce and examine one threat modeling framework. We selected CoG because it incorporates many key characteristics from across more pervasive frameworks: CoG provides practitioners with a top-down approach to identifying internal points of vulnerability, similar to STRIDE [38, 39], and it assists with assessing vulnerabilities from an adversarial perspective, similar to attack trees, security cards, persona non grata, and cyber kill chain [9, 14, 28, 54]. Uniquely among popular threat modeling approaches, it allows organizations to prioritize defensive efforts based on risk priority.

We next briefly describe the process of applying the CoG approach. Figure 1 illustrates these steps using an example provided by one participant.

To begin using CoG, analysts must start by codifying the long-term organizational objective, or "end state," of defensive measures. An end state provides the *why* for implementing defenses and allows an individual practitioner to understand their own specific security objective with respect to the organization.

Once the practitioner understands their objective, the next step is to identify all of the assets currently in use that support accomplishing the objective. In this context,

an asset can be a system, a service, a tool, or anything relevant to accomplishing the objective (not just security-specific assets). The practitioner then identifies the CoG as the pivotal asset on which all other assets depend for accomplishing the objective.

Once the practitioner identifies the CoG, they can deconstruct it into three components: critical capabilities, critical requirements, and critical vulnerabilities [17]. *Critical capabilities* (CC) are distinguished by two key features: they support the practitioner's objectives, and the CoG would cease to operate without them [21]. For each CC, there are one or more *critical requirements* (CR), defined as supporting resources that enable the CC to function [21]. Eikmeier distinguishes between capabilities and requirements using a "does/uses" litmus test [17]: If the CoG does something, that something is a capability, and if it uses something, that something is a requirement. *Critical vulnerabilities* (CV) are directly related to critical requirements; CVs are thresholds of diminished CRs that make the CoG inoperable [55]. Practitioners identify CVs by asking the following question for each CR: what would cause this requirement to no longer function as intended? Some CVs are binary, such as the complete loss of a CR, but others may cause a reduced functionality beyond some threshold, preventing the CoG from accomplishing the objective.

Building a thorough list of critical vulnerabilities allows the practitioner to understand how their objectives can be threatened. The practitioner should consider both malicious and accidental threats to collectively describe the worst-case situation for their organization and objectives. The CoG approach models all threats with a singular, unified motivation: exploiting critical vulnerabilities. This allows practitioners to develop a list of threats that can encompass nation-state hackers, insiders, poorly trained users, and others. The practitioner iterates over the list of critical vulnerabilities to develop a corresponding list of *threat capabilities* (TC). For each CV, they ask: what could take advantage of this vulnerable condition? From the list of TCs, they enumerate all of the threat requirements (TR) needed to support each capability.

The final step in the CoG analysis process is building an *actionable defense plan* (ADP) that can neutralize identified threat capabilities and requirements, mitigate critical vulnerabilities, and protect the identified CoG. Each component of an ADP, designed to dampen or eliminate one or more potential risks, is referred to as a *mitigation strategy*.

2.3 Empirically evaluating threat models

A limited number of threat-modeling frameworks have been empirically evaluated, and none have been assessed at the enterprise level. Sindre and Opdahl [50, 58] compared the effectiveness of *attack trees* against *misuse cases* and Labunets et al. [32] compared *CORAS* [34] against *SREP* [37]. In both of these empirical studies, researchers measured the effectiveness of each framework by quantitatively comparing output from student groups. Additionally, these studies measured the perceived effectiveness of the frameworks through post-study questionnaires based on the Technology Acceptance Model [12]. Massacci et al. [35] used small groups of industry practitioners and students to compare the performance of four threat models [20,24,34,42] against fictional scenarios in a classroom environment, largely based on participants' perception of the frameworks.

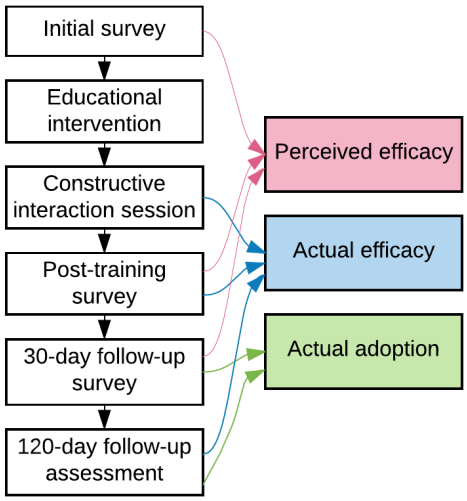

Figure 2: Our six-part study protocol and metrics.

In our study, we do not compare different frameworks to each other. Instead, we use one particular approach as a case study to examine the introduction of threat-modeling within an enterprise environment, using participants with a direct, vested interest in improving their job performance and the security posture of their environment. We utilize qualitative research methods based on studies from Sindre, Opdahl, Labunets, and Moody [32, 41, 50] while aggregating quantitative data to determine how well threat modeling protects digital systems.

3 Case study: Threat modeling at NYC3

To evaluate the impact of introducing threat modeling to an organization that had not previously used it, we partnered with NYC3 to introduce a specific threat-modeling framework (CoG) and observe the effects. NYC3 is responsible for protecting the most populous city in the U.S. and its government from cyber attacks. The Government of the City of New York (GoNYC) includes 143 separate departments, agencies, and offices with more than 300,000 employees that support 8.6 million residents and 60 million yearly visitors [48]. It maintains nearly 200,000 external IP addresses and has its own In-

ternet Service Provider, with hundreds of miles of fiber-optic cable and dozens of major points of presence. Further, the city is responsible for maintaining industrial control and mainframe systems. We drew our participant pool from the civil servants and private-sector contractors who work directly with NYC3.

Throughout this study we focus on the *efficacy* of threat modeling, which in this context we define as the ability to achieve a desired outcome. Both *effectiveness*, the ability to successfully achieve an outcome, and *efficiency*, the ability to reduce effort to achieve an outcome, comprise efficacy.

Because we introduced threat modeling in NYC3's operational environment, we were not able to conduct a comparative experiment; instead, we designed a primarily observational study to obtain as much insight as possible — both qualitative and quantitative — into the effects of introducing threat modeling within an enterprise environment. Our study includes six components (as shown in Figure 2), that occurred from June through November 2017, and was approved by the University of Maryland Institutional Review Board. Due to the study's sensitive nature, we generalized some details about defenses and vulnerabilities to protect NYC. Additionally, we redacted sensitive information when quoting participants and generalized job descriptions so as to not deanonymize participants.

3.1 Recruitment

NYC3 leadership sent all of its employees an email that outlined the voluntary nature of our study as well as our motivation and goals. The email informed NYC3 employees that they would be introduced to new techniques that could potentially streamline their daily duties, and that the findings from the study would be directly applied to defending NYC3 systems and networks. We conducted the study during participants' regularly scheduled work hours and did not provide them with any additional monetary incentives for participating.

3.2 Study protocol

We designed a multi-part study protocol, as follows.

Protocol pilot. Prior to deploying our protocol with participants, we conducted three iterations of the study using non-NYC3 employees (two security practitioners and one large-organization chief information security officer) to pre-test for relevance, clarity, and validity. We updated the study protocol based on pilot feedback and overall study flow. After three iterations, we arrived at the final protocol described below.

Baseline survey. Establishing a baseline for NYC3 defensive practices allows us to compare the security posture before and after our training intervention. We asked participants about their specific work role, responsibilities, and demographics; their understanding of organizational mission statements; which assets they use to accomplish their daily duties; their sentiment towards NYC3's current security posture; and their perceived self-efficacy for performing digital security tasks.

We used a combination of open-ended, close-ended, and Likert-scale questions in our 29-question online survey (App. B). We based all self-efficacy questions on best-practices and question-creation guides from established educational psychology studies [5]. We used an identical structure for the post-training survey and 30-day follow-up survey. Capturing self-efficacy before, immediately after, and 30 days after receiving the educational intervention allowed us to measure how each participant perceived the model's efficacy. We were interested in measuring efficacy perceptions, as self-efficacy has been shown to be an important component of individual success at performing job duties in enterprise settings [4]; one key component of self-efficacy is belief in the efficacy of the tools you use to complete tasks.

Educational intervention. After completing the initial survey, we provided groups of participants with in-person instruction on the history of CoG, its application as a threat modeling technique, the CoG process outlined in Section 2.2, and two examples of applying the framework. We scheduled three independent sessions and allowed participants to choose the session most convenient to their work schedule.

We based our 60-minute educational intervention on fundamentals from adult learning research and the experiential learning theory (ELT) [31]. Kolb and Kolb found that adults learn new concepts better through ELT by (1) integrating new concepts into existing ones, (2) accommodating existing concepts to account for the new concepts, and (3) "experiencing, reflecting, and acting" to reinforce the new concepts [31]. Social learning theory (SLT) further supports this process, indicating that adults learn new patterns of behavior best through direct experience [6]. Thus, our class was designed to reinforce each concept with a hands-on exercise using scenarios relevant to the audience and their domain knowledge.

During the class, the instructor introduced participants to tabular and graph-based methods performing CoG analysis [59]; we include examples of both in App. D. The tabular tool allows users to record their responses to each subtask of the CoG framework; each section supports data in follow-on sections. The graph-based method provides users with an alternative, complementary method for eliciting the same data. Previous research indicates that various learning styles benefit from multiple forms of data elicitation [31].

During the first classroom example, the instructor guided participants through a scenario drawn from the Star Wars movie franchise to determine the CoG for the Galactic Empire. The instructor provided step-by-

step instructions for using the tabular and graphical tools throughout. In the second example, the participants worked together without instructor guidance to apply CoG and framework tools to a fictional e-commerce scenario. We describe both fictitious scenarios in App. A.

Prior to providing the intervention, the instructor observed NYC3 employees at work for four days to better understand their operating environment. The instructor developed the fictitious scenarios so that they did not reflect any specific conditions within NYC3. We chose these scenarios in lieu of NYC3-specific scenarios to reduce bias during training that would inadvertently coach participants towards providing "approved solutions."

To control for variations in instruction, each group had the same instructor. The instructor is a member of the research team with extensive subject-matter knowledge and experience, including six months of formal university training on threat modeling. The instructor communicated this experience prior to each class to establish a baseline of credibility with the group. During each class, participants could ask questions at any time, and the instructor maintained a running log of these questions. To maintain consistency across class sessions, the instructor incorporated answers to these questions at relevant points in future sessions, and emailed the answers to participants who had attended previous sessions.

Performance evaluation session. After all participants finished the educational intervention training, they each completed a 60-minute individual session where they applied CoG to their daily duties. For example, P17 used the framework in his role as a security analyst to develop plans for better defending NYC endpoint workstations (See App. A.3). This phase of the study provided hands-on reinforcement learning, as recommended by ELT and SLT [6, 31].

We audio recorded each session, provided participants with clean worksheets and whiteboards for brainstorming (App. D), and allowed participants to bring in any notes from the previous educational intervention training. Without notifying the participants, we logged task completion times for each step, in an effort to measure the efficiency of the framework without putting undue pressure on participants.

The interviewer used the constructive interaction method for communicating with the participants, asking them to openly communicate throughout each subtask in Section 2.2 [40]. During each step, the instructor re-stated participants' previous verbal comments or documented responses to assist with data elicitation but did not introduce any new concepts to prevent data bias. For consistency, the same interviewer completed all performance evaluation sessions.

At the completion of each session, we retained a copy of the completed worksheets, photographed the whiteboards, and returned the original worksheets to the participant to help guide their responses for the second online survey. The aggregated worksheets and time logs support measurements for the actual efficacy of the CoG framework (See Section 4.3.2).

The performance evaluation interviewer transcribed responses to the open-ended questions after each session using the audio recordings. Two researchers jointly analyzed all open-ended survey questions and each transcription using iterative open-coding [61]. In alignment with this process, we coded each research artifact and built upon the codebook incrementally. We resolved all disagreements by establishing a mutually agreed upon definition for coded terms. From here, we re-coded previously coded items using the updated codebook and repeated this process until we coded all responses, resolved all disagreements, and the codebook was stable.

Post-training survey. In this 27-question online survey (App. B), conducted immediately after the performance evaluation session, we collected responses measuring the framework's actual and perceived efficacy. We asked participants to re-apply CoG to their daily duties, which allowed them to account for any new details they might have considered since the previous session. Additionally, we asked them to re-evaluate their perception of the NYC3 baseline security posture and their ability to complete digital security tasks. Using this information, we can measure changes in how participants view the organization and their own abilities [19]. Further, we asked participants to evaluate their ability to complete digital security tasks solely using the CoG framework and and to answer comprehension questions measuring their current understanding of the framework.

Follow-up survey. The 13-question follow-up survey (App. B) allowed us to measure framework adoption, knowledge retention, and perceived efficacy 30 days after researchers departed. To measure the extent to which participants adopted CoG analysis without instructor stimulus, we asked them to describe whether and how they used the information derived from CoG analysis or the framework itself within their daily duties. These questions allow us to understand participants' ability to apply output from the framework, measure their adoption rates at work, and measure their internalization of CoG concepts. We also continued to use self-efficacy questions supplemented with survey questions from the technology acceptance model (TAM) [12].

Long-term evaluation. After 120 days, we evaluated the efficacy of adopted defense plans for protecting NYC3 systems. We used a combination of NYC3 incident reports and system logs extracted solely from defensive measures that participants recommended and implemented because of their use of CoG threat modeling.

NYC3 deployed these new defensive measures in "blind spots," so each verified intrusion attempt or vulnerability clearly links an improved security posture to these new defensive measures.

3.3 Limitations

All field studies and qualitative research should be interpreted in the context of their limitations.

We opted to measure only one threat-modeling framework: although our sample represents 37% of the NYC3 workforce, 25 participants (in many cases with no overlap in work roles) would not have been sufficient to thoroughly compare multiple approaches. Testing multiple models within participants was impractical due to the strong potential for learning effects and the need to limit participants' time away from their job duties. As such, it is possible that other threat-modeling or training approaches would be equally or more effective. We believe, however, that our results still provide insight as to how threat modeling in general can benefit a large enterprise.

As we will describe in Section 4.3.2 below, we used two NYC3 leaders to jointly evaluate the defense plans produced by our participants. More, and more independent, evaluators would be ideal, but was infeasible given confidentiality requirements and time constraints on NYC3 leadership.

Our results may be affected by demand characteristics, in which participants are more likely to respond positively due to close interaction with researchers [27, 51, 63]. We mitigated this through (1) anonymous online surveys that facilitated open-ended, candid feedback, (2) removing researchers from the environment for 30 days before the follow-up survey, and (3) collecting actual adoption metrics. Further, because we explained the purpose of the study during recruitment, there may be selection bias in which those NYC3 personnel most interested in the topic or framework were more likely to participate; we mitigated this by asking NYC3 leaders to reinforce that (non-)participation in the study would have no impact on performance evaluations and by recruiting a large portion of the NYC3 workforce.

NYC3's mission, its use of pervasive defensive technologies, and its adherence to common compliance standards indicate that NYC3 is similar to other large organizations [29, 44, 45]; however, there may be specific organizational characteristics of NYC3 that are especially well (or poorly) suited to threat modeling. Nonetheless, our results suggest many directions for future work and provide novel insights into the use of threat modeling in an enterprise setting.

TAM has been criticized (e.g., by Legris et al. [33]) for insufficient use coverage. Additionally, the positive framing of TAM questions may lead to social desirability biases [16]. To address coverage, we use TAM in conjunction with the Bandura self-efficacy scales for a more complete picture. Moreover, reusing validated survey items and scales in this study is a best-practice in survey design that has been shown to reduce bias and improve construct validity [18, 22]. Lastly, we elicited participant feedback with a negative framing explicitly after each performance evaluation session, and implicitly when assessing threat modeling adoption at the 30-day evaluation. Eliciting feedback through negatively-framed mechanisms allowed participants to provide their perceptions from both perspectives.

For each qualitative finding, we provide a participant count, to indicate prevalence. However, participants who did not mention a specific concept during an open-ended question may simply have failed to state it, rather than explicitly disagreeing. We therefore do not use statistical hypothesis tests for these questions.

4 Results

Below we present the results of our case study evaluating threat modeling in an enterprise environment, drawing from transcripts and artifacts from performance evaluation sessions, survey answers, and logged security metrics. We report participant demographics, baseline metrics, immediate post-training observations, 30-day observations, and observations after 120 days.

We organize our findings within the established framework of perceived efficacy, actual efficacy, and actual adoption [32,41,50]. Participants' perceived efficacy and belief that they will achieve their desired outcomes directly shape their motivation for adopting threat modeling in the future [3]. Actual efficacy confirms the validity of perceptions and further shapes the likelihood of adoption. Lastly, regardless of perceived or actual efficacy, a framework must be adopted in order to demonstrate true efficacy within an environment. Through these three measurements, we provide security practitioners with the first structured evaluation of threat modeling within a large-scale enterprise environment.

4.1 Participants

Qualitative research best practices recommend interviewing 12-20 participants for achieving data saturation in thematic analysis [23]. To account for employees who might need to withdraw from the study due to pressing work duties, we recruited 28 participants for our study. Of these, 25 participants completed the study (Table 1), above qualitative recommendations, and we also reached saturation in our performance evaluation sessions. For the rest of this paper, all results refer to the 25 participants who completed the study. This sample represents 37% of the NYC3 employees as of August 8, 2017.

Technicians such as network administrators and security engineers account for 18 of the participants; the remainder fulfill supporting roles within NYC3 (e.g., lead-

ID	Duty Position	IT Exp (yrs)	Trng. (yrs)	Educ.[1]
P01	Leadership	16-20	6-10	SC
P02	Data Engr.	16-20	6-10	G
P03	Sec Analyst	11-15	0-5	SC
P04	Sec Engineer	11-15	0-5	BS
P05	Governance	16-20	6-10	SC
P06	Sec Engineer	6-10	11-15	P
P07	Sec Engineer	0-5	6-10	G
P08	Net Admin	21-25	6-10	G
P09	Sec Engineer	11-15	0-5	SC
P10	Sec Engineer	11-15	6-10	BS
P11	Net Admin	16-20	6-10	BS
P12	Sec Engineer	25+	6-10	G
P13	Sec Analyst	0-5	0-5	BS
P14	Sec Engineer	11-15	0-5	BS
P15	Sec Engineer	16-20	25+	SC
P16	Support Staff	6-10	0-5	BS
P17	Sec Analyst	16-20	16-20	G
P18	Sec Engineer	21-25	16-20	G
P19	Sec Analyst	21-25	6-10	SC
P20	Leadership	11-15	6-10	G
P21	Sec Analyst	0-5	6-10	G
P22	Leadership	11-15	6-10	G
P23	Sec Analyst	16-20	6-10	BS
P24	Leadership	0-5	0-5	BS
P25	Leadership	0-5	0-5	G

[1] SC: Some College, BS: Bachelor's, G: Graduate degree, P: Prefer not to answer

Table 1: Participant demographics

ership, policy compliance, and administrative support). This composition is similar to the actual work role distribution across NYC3, with 50 of 67 employees serving as technicians. Prior to this study, one participant had a high-level understanding of the military applications of CoG, and none of the participants had any applied experience using any threat-modeling framework.

All participants had at least some college education, with ten holding a graduate degree and eight holding a bachelor's. Additionally, 15 possessed at least one industry certification. Participants had an average of 14.7 years of information technology and security experience in large organizations, with a mean of 8.5 years of formal or on-the-job training.

4.2 Pre-intervention baseline

To measure the impact of threat modeling within NYC3 systems, we first established a baseline of how participants deployed defensive strategies prior to our training. Most commonly, they prioritized defending high-impact service-based systems such as NYC.gov (n=7) and adhering to compliance frameworks (n=7), followed by applying risk management strategies (n=6) and assessing which systems are most susceptible to attack (n=3). Participants reported using the following guidelines and programs for assessing NYC's digital security posture: city-specific policies and executive orders such as the NYC remote access policy [49] (n=6), NIST Cybersecurity Framework [44] (n=4), and NYC3's one-time accreditation process for adding new technologies to their network (n=2). Of these guidelines, participants stated that none of the programs were applied frequently enough, with P5 stating that "compliance is only as good as your last assessment." With too much lapsed time between audits, defenders cannot establish an accurate assessment of the environment's security posture over time. The remainder of respondents (n=13) said they were unsure about which programs or policies were applicable.

4.3 Immediate observations

In contrast to the baseline survey, performance evaluation session observations and post-training surveys indicate that threat modeling provided participants with a better understanding of their security environment, that participants felt more confident in their ability to protect NYC, and that participants could successfully apply threat modeling relatively quickly with accurate results.

4.3.1 Perceived efficacy

We observe participants' initial threat modeling perceptions in the context of new insights, framework usefulness, and changes in self-efficacy.

New understanding. Overall, 12 of 25 participants reported that threat modeling allowed them to understand new critical capabilities, requirements, or vulnerabilities that they had never previously considered. In particular, four participants had never previously mapped threats to vulnerabilities. P16, a non-technical administrative support staffer, used threat modeling to understand the implications of wide-open security permissions on a wiki and networked share drive.

Threat modeling provided two participants with self-derived examples of why crisis continuity plans exist for large organizations. P04 stated that this new understanding would further assist him with planning for crises, allowing him to recommend to "senior management the plan of action for what should be done first."

Of the 13 participants who did not report discovering anything new, seven stated threat modeling was simply a restructured approach to current defensive concepts like defense-in-depth [36]. Four stated threat modeling did not help them discover anything new but added additional emphasis to areas they should be concerned with.

Four participants identified an over-reliance on personal relationships (rather than codified policies) as a critical vulnerability for organizational success, which conceptually is something none of them had ever before considered. During his performance evaluation session, P24 discussed how changes in the political environment from the local to federal level can affect established trust across the GoNYC; a large turnover in personnel could halt some progress and potentially kill some initiatives. P25 stated "I had not really considered... the impact that some sort of major, non-cyber event could have

on our ability to be successful," discussing how a major terrorist event within NYC could decrease NYC3's ability to sustain critical requirements and capabilities. Thus, both participants recommended codifying existing relationship-based agreements into legislation capable of withstanding non-digital security threats to their daily responsibilities. An example of this includes establishing a formal memorandum of understanding (MoU) with law enforcement agencies in NYC to facilitate the exchange of threat indicators.

Perceived framework usefulness. After completing the performance evaluation session, 23 participants agreed that threat modeling was useful to them in their daily work. For example, ten said the framework allowed them to prioritize their efforts. P24 developed a new litmus test for adding any defensive efforts, stating that "If the adversary doesn't care, then it's all just fluff [inconsequential]." P21 used threat modeling to show "what we're lacking or what we need to concentrate [on]," such as standard cyber hygiene.

Eight participants expressed that threat modeling added much-needed structure and perspective to difficult problems. P11 feels empowered by its structure and believes it allows him to "accept the things you cannot change, change the things you can, and have the wisdom to know the difference. I feel [CoG is] along those lines; this is your world, this is what you control." He believes threat modeling makes a positive difference with available resources, while helping to prioritize requests for future capabilities and support.

Five participants reported that threat modeling allowed them to plan defensive strategies more effectively. P05 stated that threat modeling helps him "plan effectively, document, track, monitor progress, and essentially understand our security posture."

Threat modeling allowed four participants to comprehend how threats can affect systems within their environment; these technicians previously relied upon best security practices without fully considering threats. While applying the framework, P10 declared that "insider threats overcome the hard shell, soft core" within most enterprise networks and that threat modeling helped him identify new ways to neutralize the impact of insiders bypassing perimeter defenses and exploiting trusted internal systems.

Four participants stated that purposefully considering their asset inventory during threat modeling allowed them to fully understand their responsibilities. Three participants stated that threat modeling provides them with a new appreciation for their position within NYC3. P14 said, "When I did my job, I didn't think about what the purpose of our group is [within NYC3]...[threat modeling] aligns what we're thinking with what I think my role is in this organization."

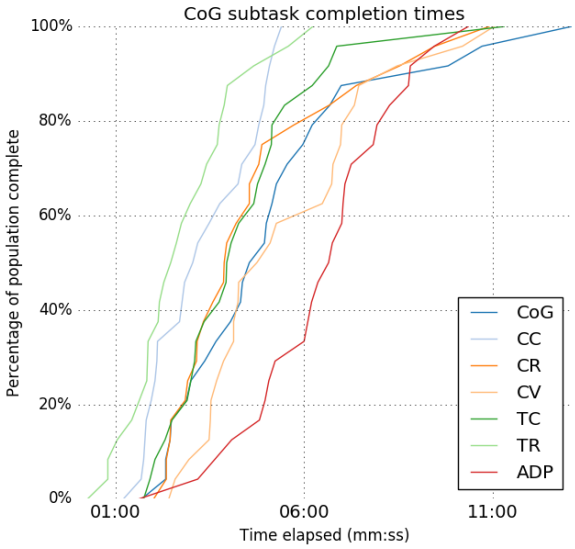

Figure 3: A cumulative distribution function (CDF) for participant subtask completion times.

Interestingly, both of the participants who did not find threat modeling useful felt that cybersecurity is too nebulous of a realm for a well-structured approach like CoG. P12, when asked to clarify his difficulties with the framework, stated that cloud environments present unique problems for defenders: we care about "the center keep of your castle, well there's this other castle somewhere out there, we don't know where, [and it is] part of our CoG." However, these two participants did successfully use threat modeling to discover critical vulnerabilities within their daily work that they had not previously considered.

Changes in self-efficacy. When comparing responses from the post-training survey to baseline responses, 10 participants reported a perceived increase in their ability to monitor critical assets, 17 reported an increase in their ability to identify threats, 16 reported an increase in their ability to mitigate threats, 15 participants reported an increase in their ability to respond to incidents. Respectively, averages increased by 8.8%, 19.3%, 29.8%, and 20.0%. Using the Wilcoxon signed-rank test [65], we found significant increases in participants' perceived ability to identify threats (W=61.0, p=0.031), mitigate threats (W=47.0, p=0.010), and respond to incidents (W=59.0, p=0.027).

4.3.2 Actual efficacy

We measure the actual efficacy of threat modeling using several metrics: the accuracy of participants' output, task completion times, similarities between participants' identified CoGs, and the contents of their actionable defense plans.

Output accuracy. Simply completing CoG tasks is insufficient to demonstrate success; the resulting output must also be valid and meaningful. Thus, we assess the

accuracy of participants' results via an expert evaluation from two NYC3 senior leaders. Both of these leaders received in-person training on CoG and are uniquely qualified to assess the accuracy of the provided responses given their intimate knowledge of the NYC3 environment and cybersecurity expertise. We provided the evaluators with an anonymized set of the study results and asked them to jointly qualify the accuracy of the identified centers of gravity, critical vulnerabilities, threat capabilities/requirements, and ideal defense plans using a 6-point Likert scale ranging from zero to five with zero being "extremely unlikely (UL)" and five being "extremely likely (EL)" (See App. C). Additionally, we asked the leaders to indicate whether each ADP was sufficiently detailed to implement. We included one fictitious participant entry as an attention check and validity control, which both panel members identified and rejected.

The panel concluded that: 22 of 25 identified centers of gravity were accurate with respect to a participant's responsibilities ('EL'=3, 'Likely [L]'=9,'Somewhat likely [SL]'=10); all critical vulnerabilities were accurate for the identified centers of gravity (EL=6, L=7, SL=12); 23 of 25 threat capability and requirement profiles were accurate (EL=6, L=7, SL=10), and 24 of 25 actionable defense plans would accurately address the identified threats (EL=5, L=11, SL=8).

We used a logistic regression, appropriate for ordinal Likert data, to estimate the effect of work roles, experience in IT, and educational background on the accuracy of the panel results. We included a mixed-model random effect [26] that groups results by work roles to account for correlation between individuals who fill similar positions. Our initial model for the regression included each demographic category. To prevent overfitting, we tested all possible combinations of these inputs and selected the model with minimum Akaike Information Criterion [1]. The final selected model is given in Appendix E. Based on this regression, we found that no particular work role, amount of education, IT experience, or combination thereof enjoyed a statistically significant advantage when using threat modeling. These high success rates across our demographics support findings by Sindre and Opdahl that indicate threat modeling is a natural adaptation to standard IT practices [58].

Time requirements. We use the time required to apply CoG analysis to measure efficiency, which is a component of efficacy. On average, participants used the framework and developed actionable defense plans in 36 minutes, 46 seconds ($\sigma = 9:01$). Figure 3 shows subtask completion times as a cumulative distribution function (CDF). Participants spent the greatest amount of time describing critical vulnerabilities and developing actionable defense plans, with these tasks averaging 5:27 and 6:25 respectively. Three out of five participants in a leadership role affirmed without prompting that threat modeling provided them with a tool for quickly framing difficult problems, with P24 stating "within an hour, [CoG] helped me think about some items, challenge some things, and re-surface some things, and that is very useful for me given my busy schedule." P22 applied the framework in 22 minutes and commented during his closing performance evaluation session that he would "need much more time to fully develop" his ideas; however, he also said the session served as a catalyst for initiating a necessary dialogue for handling vulnerabilities.

CoG consistency. Analysis of the performance evaluation session results reveals that participants with similar work role classifications produced similar output. For example, 16 of 18 technicians indicated that a digital security tool was their CoG (e.g., firewalls, servers) whereas four of six participants in support roles identified a "soft" CoG (e.g., relationships, funding, and policies). Participants produced actionable defense plans averaging 5.9 mitigation strategies per plan and ranging from a minimum of three strategies to a maximum of 14.

Actionable defense plans. We use the contents of participants' actionable defense plans to further evaluate success. Participants identified real issues present within their environment and developed means for reducing risk. Within the 25 actionable defense plans, participants cumulatively developed 147 mitigation strategies; we provide detailed examples in Section 4.5. Participants indicated that 33% of the mitigation strategies they developed using threat modeling were new plans that would immediately improve the security posture of their environment if implemented. Additionally, participants stated that 31% of the mitigation strategies would improve upon existing NYC3 defensive measures and more adequately defend against identified threats. Participants felt that the remaining 36% of their described mitigation strategies were already sufficiently implemented across the NYC3 enterprise.

The NYC3 leadership panel indicated a majority of the actionable defense plans were sufficiently detailed for immediate implementation ('Yes'= 16). This shows that, even with limited framework exposure, many participants were able to develop sufficient action plans. We illustrate an ADP with insufficient detail using a security analyst's plan. After identifying his CoG as an Endpoint Detection and Response (EDR) system[3] and applying the framework, his ADP consisted of three mitigation strategies: "Make sure there is a fail-over setup and test it. Better change control. Better roll back procedures." While all of these address critical vulnerabilities, they provide no implementation details. In cases such as this, individuals require additional time to improve the fidelity of their responses or may benefit from expert assistance in

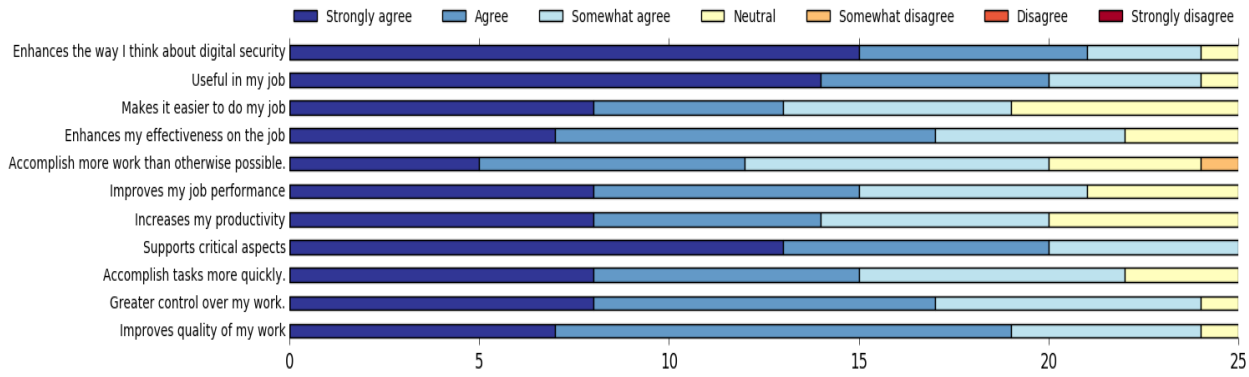

Figure 4: Perceived efficacy after using threat modeling for 30 days.

transforming their ideas into fully developed plans.

4.4 Observations after 30 days

After 30 days, we observed that participants still had a favorable opinion of threat modeling, most participants actually implemented defensive plans that they developed through our study, and that NYC3 institutionalized threat modeling within their routine practices.

4.4.1 Perceived efficacy

Thirty days after learning about CoG, there was a slight decrease in the perceived efficacy of the framework when compared to participant perceptions immediately after training: a 1.47% decrease for monitoring critical assets (W=81.0, p=0.57), 3.22% decrease for identifying threats (W=131.0, p=0.83), 3.58% decrease for mitigating threats (W=94.0, p=0.18), and 1.67% decrease for responding to incidents (W=100.0, p=0.59); none of these decreases were statistically significant. When comparing these 30-day metrics to the baseline, however, participants' perceived ability to monitor critical assets increased 7.4%, perceived ability to identify threats increased 16.1%, perceived ability to mitigate threats increased 26.3%, and perceived ability to respond to threats increased 18.3%. Participants' perceived ability to mitigate threats is a statistically significantly increase from the baseline (W=73.5, p=0.049).

Figure 4 shows participants' evaluations of the efficacy of CoG analysis after 30 days. Overall, all participants agreed ("Strongly"= 13) that threat modeling supports critical aspects of their job. Additionally, 24 participants agreed ("Strongly"= 15) that threat modeling enhances the way they think about digital security. Despite the aforementioned decrease in perceived efficacy over the 30-day period, the number of participants who found the framework useful to their jobs increased from 23 to 24, as NYC3's adoption of ADPs within their environment caused one participant to believe in the framework's usefulness. Lastly, 245 of 275 responses to our 11 TAM questions indicated threat modeling is valuable for digital security.

4.4.2 Actual efficacy

We measure actual efficacy after 30 days using participants' knowledge retention. Measuring knowledge retention allows us to evaluate the longevity of organizational impacts from integrating the framework. After 30 days, participants averaged 78% accuracy on four comprehension questions. This is an increase from 69% immediately after learning the framework, suggesting threat modeling may become more memorable after additional applied experience. Each comprehension question required participants to pinpoint the best answer out of three viable responses; this allowed us to measure if participants understood critical relationships. In the 30-day follow-up, all participants accurately answered our critical vulnerability question, 23 correctly identified a CoG visually, 17 correctly identified a critical requirement for a capability, and 13 correctly identified a critical capability for a notional CoG.

4.4.3 Actual adoption

After 30 days, 21 participants reported that they implemented at least one mitigation strategy that they developed using threat modeling. In addition, 20 participants reported after 30 days that they integrated concepts from threat modeling within their daily work routines. For example, seven participants now use the framework for continually assessing risk; this is in contrast to the baseline results, where participants typically assessed risk only during audits and initial accreditation. Five participants stated that they now use threat modeling to prioritize their daily and mid-range efforts. Participants who did not adopt said they were too busy with urgent tasks (n=4) or needed more applied training (n=1).

NYC3 started to institutionalize threat modeling after participants had discussed their results with one another and realized the important implications of their findings. One week after completing their performance evaluation sessions, six participants transformed a wall within their primary meeting room into an "urgent priorities" board (Figure 5) for implementing defensive actions that address critical vulnerabilities identified during this study.

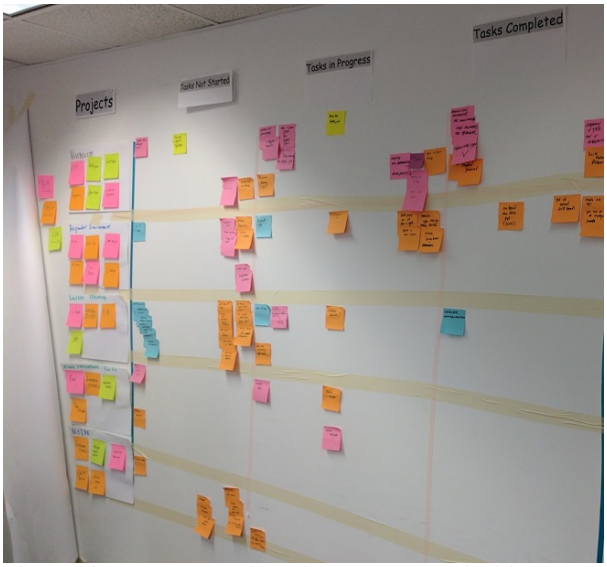

Figure 5: NYC3 developed an "urgent priorities" task tracker to address problems identified in this study.

Their board facilitates two-week action periods and improves how the organization communicates the impact of their progress to senior leaders. NYC3 leaders have since formalized this board using project management software and other practices such as "demo days" to demonstrate the viability of their defensive efforts.

4.5 Observations after 120 days

Observing NYC3's environment 120 days after our study concluded allows us to understand the longer-term impact of threat modeling within live work environments. In total, we find that NYC3 implemented eight new categories of controls directly based on the ADPs developed by participants in this study. Additionally, NYC3 provided us with access to server logs, their alert dashboard, and vulnerability reports so that we could measure the actual efficacy of three of these new controls.

4.5.1 Actual adoption

Below we provide a sample set of ADPs that participants derived using threat modeling. NYC3 leaders monitored the implementation of these ADPs using their priorities board, and all mitigation strategies persist within the NYC environment 120 days after the study. We only provide high-level details about the ADPs below to avoid placing NYC3 systems at risk.

Testing readiness. Nine participants cited resilient systems as critical requirements within their environment, and two identified untested disaster recovery plans as critical vulnerabilities. To dampen the impact of a cyber attack, natural disaster, or terrorist attack, they recommended frequently using multiple "fail-over" sites to validate functionality. Accordingly, NYC3 has begun testing fail-over servers within their local domain and plans to implement periodic, mandatory readiness tests across all NYC networks.

Securing accounts. Several participants identified user account permissions – a fundamental security control in any networked environment – as insufficiently well managed. Three participants stated that it is common for employees to migrate across the organization and retain permissions to data shares and assets they no longer need. NYC3 now directs monthly audits and re-certification of user access to narrow the impact of insider threats or stolen credentials. Seven participants recommended implementing multi-factor authentication. As a proof of concept, NYC3 implemented multi-factor authentication for 80 user accounts within a monitored subdomain.

Protecting physical network assets. Seven participants determined that if control measures restricting physical access to networking infrastructure were weak, it would create critical vulnerabilities. All expressed concern with insider threats causing damage or stealing data, but they all indicated that the most likely threat stems from accidental damage. Three participants discussed concerns with inadvertent, wide-scale power outages or power surges to networking infrastructure that could cause some issues to persist for an extended duration. These three participants recommended security escorts for all personnel, in addition to multi-factor access control near all networking infrastructure. Since the performance evaluation sessions, NYC3 has been working with federal, state, and private-sector entities on issues related to this topic.

Crowdsourcing assessments. Two participants reported that automated vulnerability assessment tools might not detect all vulnerabilities and that manual testing is needed for identifying more complex issues. Thus, P21 recommended that NYC establish a bug bounty program for public-facing services to benefit from the collective security community. Because of his recommendation, NYC3 partnered with a bug bounty service provider to conduct a 24-hour proof-of-concept assessment against one of its web services.

Sensor coverage. Ten participants acknowledged that the NYC environment is far too vast for manual monitoring and that automated sensors play a critical role in defense. In this situation, a gap in sensor coverage can lead to unprotected systems or the successful exploitation of known vulnerabilities. Four participants recommended deploying additional EDRs on systems in specific subdomains within which NYC3 had limited visibility. Within 30 days after the threat modeling training, NYC3 technicians deployed 1331 new EDR sensors within these subdomains.

Protecting legacy systems. Three participants stated that legacy systems significantly impact their ability to

secure systems; some were installed five decades ago and were never intended to be networked. Thus, they recommended segmenting non-critical legacy systems until they are replaced/upgraded. NYC3 is now working closely with partners to protect segmented systems and those that must remain online.

Protecting against data corruption. Participants P02 and P17 identified data corruption as risks to NYC3 systems. NYC3 technicians now verify the integrity of each software and indicator of compromise (IOC) update provided by third-party vendors to prevent the exploitation of update mechanisms, as seen in the 2017 NotPetya malware outbreak [56].

Reducing human error. Human error was another common theme across the threat landscape. Six participants stated that a simple typo in a configuration script, like the one that caused the 2017 Amazon S3 outage [2], could have significant impacts across multiple systems or networks. Three defenders recommended two-person change control when updating configuration files on firewalls and EDR systems. Such controls require one person to propose a change and another to review and implement the change to reduce the likelihood of human error. NYC3 now enforces two-person change control on all modifications to access control lists.

4.5.2 Actual efficacy

Quantitative metrics captured in the 120 days after threat modeling training empirically support the efficacy of threat modeling. A NYC3 security analyst verified every intrusion, incident, and vulnerability within these data records. To protect the operational security of NYC3, we do not report on specific threats that would enable a malicious actor to re-target their systems.

Securing accounts. User account logs allow us to analyze account hijacking attempts based on the geographic origin of attempts, time frequency between attempts, and why the attempt failed (e.g., wrong password or invalid token). Over 120 days, NYC3 recorded 3749 failed login attempts; based on frequency and subsequent successful logins, we associate 3731 of these attempts with employees forgetting their password. Among the remaining failed logins, NYC3 successfully blocked hijacking attempts that originated from a foreign nation against seven *privileged* user accounts. Of these seven accounts, the attacker failed at the multi-factor login step for five accounts and failed due to password lockout on the other two accounts. Prior to this study, this subdomain did not have multi-factor verification enabled; these five privileged accounts were protected by mechanisms implemented solely because of the introduction of threat modeling.

Crowdsourcing assessments. The 24-hour bug-bounty trial program yielded immediate results. Overall, 17 security researchers participated in the trial program and disclosed three previously unknown vulnerabilities in a public webserver protected by NYC3, verified through proof-of-concept examples. NYC3 validated these vulnerabilities and patched the production systems in accordance with policy and service-level objectives. After the success of this trial, NYC3 has authorized an enduring public program that will focus on improving the security posture of web applications under NYC3's purview. Such a program is a first for the City of New York and NYC3, created as a direct result of introducing threat modeling.

Sensor coverage. EDR reports allow us to uniquely identify which IOCs appeared in which systems, their severity level, and frequency of attempts. NYC3 deployed 1331 new sensors to endpoints that were previously unmonitored and were able to verify and respond to 541 unique intrusion attempts identified by these new sensors. Of these 541 intrusion attempts, 59 were labeled critical and 135 were labeled high severity; NYC3's partnered vendor security service manually validated each of these intrusions and verified their severity levels as true positives. One important aspect to note: if any systems had been infected prior to sensor deployment, our study would have captured both new intrusion attempts and any re-infection attempts that occurred after NYC3 deployed the sensors for the first time. According to the lead NYC3 EDR engineer, all 541 of these events could have led to successful attacks or loss of system availability if technicians had not deployed the sensors to areas identified during threat modeling.[4]

5 Discussion and conclusions

We provide the first structured evaluation of introducing threat modeling to a large-scale enterprise environment. Overall, our findings suggest that threat modeling, in this case the CoG framework, was an effective and efficient mechanism for developing actionable defense plans for the NYC3 enterprise. Defense plans created using CoG led to measurable, positive results. These results suggest that even a relatively small amount of focused threat modeling performed by IT personnel with no previous threat-modeling experience can quickly produce useful improvements.

Immediately after completing the performance evaluation sessions, 23 participants reported that they found the framework useful; after 30 days of use, 24 participants reported finding the framework useful and 20 participants reported regularly using concepts from threat modeling in their daily processes. In less than 37 minutes on average, our 25 participants developed 147 unique mitigation strategies for threats to their organization. NYC3 adopted many of these recommendations, improving their security posture in eight key areas. After

120 days, participant-designed ADPs blocked account hijackings of five privileged user accounts, blocked 541 unique intrusion attempts, and discovered (and remedied) three vulnerabilities in public-facing web servers, all of which support that introducing threat modeling made NYC3 more secure.

We note that many of the ADPs that NYC3 employees developed and implemented (Section 4.5) contain straightforward recommendations, such as applying multi-factor authentication. We believe that this in itself constitutes an important finding: despite adhering to applicable federal, state, and local compliance standards and "best practices," these measures were not already in use. Threat modeling offered our participants the agility to identify and implement defensive measures not (yet) prescribed in these standards. In this case, threat modeling helped the organization gain new perspective on their security gaps and proactively mitigate issues.

Many organizations are currently making significant investments in digital-security tools and capabilities [10]. Our case study of threat modeling, in contrast, shows promising results that can be achieved by leveraging existing resources, without the need for new technologies or personnel. Further, our approach included only two hours of employee training, which we expect would be palatable for many organizations.

5.1 Lessons learned

Based on our case study, we make several observations about the process of adopting threat modeling in a large organization.

Hands-on learning. Our participants indicated that our hands-on approach to teaching threat modeling worked well. After the performance evaluation sessions, without prompting, 24 of 25 participants said that the personalized, hands-on application allowed them to understand the framework better than the educational intervention classes alone. Our logistic regression analysis on participants' CoG accuracy revealed a relatively level understanding of the framework across educational backgrounds, experience levels, and work roles. This suggests that many different practitioners can potentially benefit from this hands-on approach, supporting findings from Kolb & Kolb [31] and Bandura [6].

Mentoring and peer partnering. Multiple participants mentioned a desire for social and organizational support to facilitate the adoption of threat modeling. In their 30-day follow-up surveys, P18 and P24 stated that NYC3 would need organizational programs in place to aid wide-scale adoption of threat modeling, such as pairing junior personnel with mentors and facilitating peer-to-peer partnerships. During their performance evaluation sessions, P09 and P19 both mentioned that threat modeling would also be useful for integrating new personnel into NYC3. We hypothesize that pairing experienced employees with junior personnel could permit mentors to orient their mentee to the environment and provide context to ongoing defensive initiatives, all while reinforcing their own understanding of threat modeling.

Further, the NYC3 leadership panel results indicated that 9 of 25 actionable defense plans were insufficiently detailed for immediate implementation. Peering would allow small teams to challenge one another and elicit details until results are adequately robust. This accords with prior studies of threat-modeling techniques, as well as peer partnering examples from other domains, that demonstrate the benefits of peer collaboration [9, 14, 15, 20, 24, 25, 28, 34, 35, 37, 38, 42, 46, 53].

Communication with leadership. After threat-modeling training, participants reported that they were better able to communicate the importance of various threats to NYC3 leadership. This was reflected in the immediate deployment of mitigation strategies, as discussed in Section 4.5. We hypothesize that use of a single threat modeling framework — in this case CoG — across administrative boundaries may help to facilitate a shared language within the organization for communicating about threats. It would be particularly interesting to explicitly evaluate whether training executive-level leadership along with on-the-ground practitioners might yield useful communication benefits.

Shortcomings. Knowledge retention results show that participants struggled with framework-specific terminology; only 17 of 25 participants correctly identified critical requirements after 30 days. When institutionalizing threat modeling, it may be helpful to provide learners with quick-reference guides containing relatable examples to help clarify essential terminology.

5.2 Future work

In this work we took advantage of a unique cooperative opportunity to evaluate the introduction of an exemplar threat-modeling approach into an enterprise environment. In future work, comparative evaluation — ideally also in real-world environments — is necessary to understand the relative effectiveness of different threat-modeling approaches and may also help to clarify in what situations and environments different threat-modeling approaches are likely to be most effective.

To this end, we suggest that threat modeling should be tested in multiple environments, to understand when and why these frameworks should be applied. Future evaluations may be able to consider how organization size, experience level and typical workload of staff members, organizational culture, and existing threat-modeling and/or security-analysis processes affect the efficacy of threat modeling. Future work should also explore less tangible organizational characteristics, such as employees' under-

standing of organizational objectives, hierarchical structure, lines of communication within and across groups, and the empowerment given to mid-level leaders.

In summary, our results indicate that introducing threat modeling — in this case, CoG — was useful for helping a large enterprise organization utilize existing resources more effectively to mitigate security threats. These findings underscore the importance of future evaluations exploring when and why this result generalizes to other real-world environments.

Notes

[1] NYC3 was formerly known as the Department of Information Technology & Telecommunications Citywide Cybersecurity Division, which was subsumed by NYC3 midway through this study [13]. For convenience, we only refer to the organization as NYC3.

[2] Due to operational security risks, we do not name specific vendor solutions.

[3] Endpoint Detection and Response (EDR) describes a suite of tools focused on detecting and investigating suspicious activities, intrusions, and other problems on endpoint systems.

[4] NYC3 deployed additional defensive capabilities based on ADPs that also assisted with detection, but are not described here in order to protect operational security concerns.

References

[1] AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control 19*, 6 (1974), 716–723.

[2] AMAZON. Summary of the Amazon S3 Service Disruption in the Northern Virginia (US-EAST-1) Region.

[3] ATKINSON, J. W. Motivational determinants of risk-taking behavior. *Psychological review 64*, 6p1 (1957), 359.

[4] BANDURA, A. Perceived self-efficacy in cognitive development and functioning. *Educational psychologist 28*, 2 (1993), 117–148.

[5] BANDURA, A. Guide for constructing self-efficacy scales. *Self-efficacy beliefs of adolescents 5*, 307-337 (2006).

[6] BANDURA, A., AND WALTERS, R. H. *Social learning theory*. Prentice-Hall Englewood Cliffs, NJ, 1977.

[7] CHUVAKIN, A. Named: Endpoint Threat Detection & Response, 2013.

[8] CLEARY, C. DEF CON 19: Operational Use of Offensive Cyber.

[9] CLELAND-HUANG, J. How well do you know your personae non gratae? *IEEE software 31*, 4 (2014), 28–31.

[10] COLWILL, C. Human factors in information security: The insider threat–who can you trust these days? *Information security technical report 14*, 4 (2009), 186–196.

[11] CONTI, G., AND RAYMOND, D. *On Cyber: Towards an Operational Art for Cyber Conflict*. Kopidion Press, 2017.

[12] DAVIS, F. D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[13] DE BLASIO, B. Executive Order 28: New York City Cyber Command, 2017.

[14] DENNING, T., FRIEDMAN, B., AND KOHNO, T. The Security Cards: A Security Threat Brainstorming Toolkit.

[15] DYKSTRA, J. A., AND ORR, S. R. Acting in the unknown: the cynefin framework for managing cybersecurity risk in dynamic decision making. In *Proceedings of the 8th International Conference on Cyber Conflict* (2016), CyCon US '16, IEEE, pp. 1–6.

[16] EDWARDS, A. L. The social desirability variable in personality assessment and research.

[17] EIKMEIER, D. C. Center of gravity analysis. *Military Review 84*, 4 (2004), 2–5.

[18] FLOYD, J., AND FOWLER, J. Survey research methods. *Survey Research Methods (4th ed.). SAGE Publications, Inc. Thousand Oaks, CA: SAGE Publications, Inc* (2009).

[19] FORSYTH, D. R. Self-serving bias. In *International Encyclopedia of the Social Sciences*, W. A. Darity, Ed., vol. 7. Macmillan Reference USA, Detroit, 2008.

[20] GIORGINI, P., MASSACCI, F., MYLOPOULOS, J., AND ZANNONE, N. Modeling security requirements through ownership, permission and delegation. In *Proceedings. 13th IEEE International Conference on Requirements Engineering* (2005), RE '05, IEEE, pp. 167–176.

[21] GORTNEY, W. E. Department of defense dictionary of military and associated terms. Tech. rep., Joint Chiefs of Staff, Washington, United States, 2016.

[22] GROVES, R. M., FOWLER, F. J., COUPER, M. P., LEPKOWSKI, J. M., SINGER, E., TOURANGEAU, R., ET AL. Survey methodology.

[23] GUEST, G., BUNCE, A., AND JOHNSON, L. How many interviews are enough? an experiment with data saturation and variability. *Field methods 18*, 1 (2006), 59–82.

[24] HALEY, C., LANEY, R., MOFFETT, J., AND NUSEIBEH, B. Security requirements engineering: A framework for representation and analysis. *IEEE Transactions on Software Engineering 34*, 1 (2008), 133–153.

[25] HARDY, G. Beyond continuous monitoring: Threat modeling for real-time response. *SANS Institute* (2012).

[26] HEDEKER, D. Multilevel models for ordinal and nominal variables. In *Handbook of multilevel analysis*. Springer, 2008, pp. 237–274.

[27] HOLBROOK, A. L., GREEN, M. C., AND KROSNICK, J. A. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public opinion quarterly 67*, 1 (2003), 79–125.

[28] HUTCHINS, E. M., CLOPPERT, M. J., AND AMIN, R. M. Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains. *Leading Issues in Information Warfare & Security Research 1*, 1 (2011), 80.

[29] INTERNAL REVENUE SERVICE. Publication 1075: Tax Information Security Guidelines For Federal, State and Local Agencies, 2016.

[30] KATZ, N., LAZER, D., ARROW, H., AND CONTRACTOR, N. Network theory and small groups. *Small group research 35*, 3 (2004), 307–332.

[31] KOLB, A. Y., AND KOLB, D. A. Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of management learning & education 4*, 2 (2005), 193–212.

[32] LABUNETS, K., MASSACCI, F., PACI, F., ET AL. An experimental comparison of two risk-based security methods. In *Proceedings of the 7th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement* (2013), ESEM '13, IEEE, pp. 163–172.

[33] LEGRIS, P., INGHAM, J., AND COLLERETTE, P. Why do people use information technology? a critical review of the technology acceptance model. *Information & management 40*, 3 (2003), 191–204.

[34] LUND, M. S., SOLHAUG, B., AND STØLEN, K. *Model-driven risk analysis: the CORAS approach*. Springer Science & Business Media, 2010.

[35] MASSACCI, F., AND PACI, F. How to select a security requirements method? a comparative study with students and practitioners. *Secure IT Systems* (2012), 89–104.

[36] MAY, C. J., HAMMERSTEIN, J., MATTSON, J., AND RUSH, K. Defense in depth: Foundations for secure and resilient it enterprises, 2006.

[37] MELLADO, D., FERNÁNDEZ-MEDINA, E., AND PIATTINI, M. Applying a security requirements engineering process. *Computer Security–ESORICS 2006* (2006), 192–206.

[38] MICROSOFT CORPORATION. The STRIDE Threat Model. Tech. rep., Microsoft Corporation, 2005.

[39] MICROSOFT CORPORATION. Microsoft Threat Modeling Tool 2016. Tech. rep., Microsoft Corporation, 2016.

[40] MIYAKE, N. Constructive interaction and the iterative process of understanding. *Cognitive science 10*, 2 (1986), 151–177.

[41] MOODY, D. L. The method evaluation model: a theoretical model for validating information systems design methods. *Proceedings of the 11th European Conference on Information Systems* (2003), 1327–1336.

[42] MOURATIDIS, H., GIORGINI, P., AND MANSON, G. Integrating security and systems engineering: Towards the modelling of secure information systems. In *Proceedings of the 15th International Conference on Advanced Information Systems Engineering* (2003), CAISE '03, Springer, pp. 63–78.

[43] MUCKIN, M., AND FITCH, S. C. A threat-driven approach to cyber security. *Lockheed Martin Corporation* (2014).

[44] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. NIST Cybersecurity Framework, 2014.

[45] NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. NIST Special Publication 800-53, 2017.

[46] NATIONAL SECURITY AGENCY INFORMATION ASSURANCE DIRECTORATE. NSA Methodology for Adversary Obstruction, 2015.

[47] NIELSEN, J. Usability metrics, July 2001. Accessed: 2017-09-01.

[48] NYC DOITT. CityNet, 2017.

[49] NYC DOITT. Cybersecurity Requirements for Vendors & Contractors, 2017.

[50] OPDAHL, A. L., AND SINDRE, G. Experimental comparison of attack trees and misuse cases for security threat identification. *Information and Software Technology 51*, 5 (2009), 916–932.

[51] ORNE, M. T. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist 17*, 11 (1962), 776.

[52] SABOTTKE, C., SUCIU, O., AND DUMITRAS, T. Vulnerability disclosure in the age of social media: Exploiting twitter for predicting real-world exploits. In *Proceedings of the 24th USENIX Security Symposium* (2015), USENIX Security '15, pp. 1041–1056.

[53] SALTER, C., SAYDJARI, O. S., SCHNEIER, B., AND WALLNER, J. Toward a secure system engineering methodolgy. In *Proceedings of the 1998 Workshop on New Security Paradigms* (New York, NY, USA, 1998), NSPW '98, ACM, pp. 2–10.

[54] SCHNEIER, B. Attack trees. *Dr. Dobb'??s journal 24*, 12 (1999), 21–29.

[55] SCOTT, K. D. Joint planning. *Joint Publication 5-0* (2017).

[56] SHACKELFORD, S. Exploring the "shared responsibility" of cyber peace: Should cybersecurity be a human right? *Kelley School of Business Research paper* (2017), 17–55.

[57] SHOSTACK, A. *Threat modeling: Designing for security*. John Wiley & Sons, 2014.

[58] SINDRE, G., AND OPDAHL, A. L. Eliciting security requirements with misuse cases. *Requirements engineering 10*, 1 (2005), 34–44.

[59] STRANGE, J., AND IRON, R. *Understanding centers of gravity and critical vulnerabilities*. Department of War Studies, Swedish National Defence College, 2005.

[60] STRANGE, J., IRON, R., AND ARMY, U. Part 2: The cg-cc-cr-cv construct: A useful tool to understand and analyze the relationship between centers of gravity and their critical vulnerabilities. *Understanding Centers of Gravity and Critical Vulnerabilities* (2004).

[61] STRAUSS, A., CORBIN, J., ET AL. *Basics of qualitative research*, vol. 15. Newbury Park, CA: Sage, 1990.

[62] TANENBAUM, A. S., AND WETHERALL, D. J. *Computer networks*. Pearson, 2011.

[63] TOURANGEAU, R., AND YAN, T. Sensitive questions in surveys. *Psychological bulletin 133*, 5 (2007), 859.

[64] VON CLAUSEWITZ, C., AND GRAHAM, J. J. *On war*, vol. 1. London, N. Trübner & Company, 1873.

[65] WILCOXON, F. Individual comparisons by ranking methods. *Biometrics bulletin 1*, 6 (1945), 80–83.

A CoG examples

We used the following two scenarios during our educational intervention training to communicate CoG analysis concepts to participants.

A.1 Star Wars walkthrough

The educational intervention instructor guided participants through this scenario, explaining the CoG analysis for the Galactic Empire. The Galactic Empire's desired end state is to provide peace and stability throughout the galaxy. To do this, their objective is to eliminate rebel forces. The Empire has many assets available for destroying the rebel scum to include: TIE fighters, stormtroopers, Darth Vader, and the Death Star. Of these assets, we know that the most powerful means for destroying planets and eradicating sources of rebellion is the Death Star; thus, it is the CoG analysis for the Empire. Critical capabilities for the Death Star include the ability to destroy planets. Critical requirements for this capability include Kyber crystals, engineers, and the superlaser. A critical vulnerability against the superlaser is accessible via a thermal exhaust port with an exterior opening. Threat capabilities include the ability to fire weapons into the exhaust port and threat requirements include X-wing fighter aircraft. Given this scenario, an actionable defense plan for the Death Star would be concealing the thermal port or installing anti-aircraft turrets near the opening.

A.2 E-commerce scenario

In the second scenario, groups of participants applied CoG analysis without instructor assistance. The following examples are not exhaustive but include actual responses from the groups. This scenario was the first and only time participants completed CoG analysis analysis in a group setting.

We consider a small e-commerce business with the primary objective of maximizing profit and secondary objectives of customer satisfaction and website availability. We focus on defending assets that maximize our profits. The e-commerce business relies on a front-end webserver, a back-end database, redundant servers with load balancers, software developers, and a banking institution. Of the previously identified assets, the back-end database is the CoG analysis it conducts transactions with customers (the primary means for accomplishing our primary objective) and because of its interconnectedness with other assets. Critical capabilities for our business back-end database include (1) conducting atomic, consistent, isolated, and durable transactions, (2) permitting responsive queries from the front-end webserver, and (3) providing security safeguards for inventories and customer data. Critical requirements for providing security safeguards for inventories and customer data would be (1) encrypted communication between customers, the front-end webserver, and the database; (2) encrypted sensitive data within the database; and (3) compliance with regulatory guidelines for business transactions. Examples of critical vulnerabilities would be continued use of software without periodically checking for updates and patching, such as continued use of OpenSSL 1.0.1 which is vulnerable to Heartbleed [52]. Threat capabilities against a vulnerable version of OpenSSL include conducting reconnaissance and network scans of vulnerable systems. Threat requirements include a valid exploit and payload against OpenSSL. A simple actionable defense plan for our running example includes (1) upgrading OpenSSL to a version that is patched against Heartbleed and (2) validating system performance post-upgrade.

A.3 Participant P17 example

Understand the end state and objective. Participant P17 is a security analyst who works within the NYC Security Operations Center (SOC). The SOC's defensive end state is maintaining an environment that is resilient and responsive to known and unknown threats. Based on P17's work role in NYC3, his personal objective is to defend workstations and respond to threats against the NYC3 environment.

Identify assets. P17 relies on network traffic inspectors, endpoint detection and response (EDR) solutions, and log aggregators to accomplish his objective. EDRs are tools for investigating suspicious activities throughout networks, hosts, and other endpoints [7].

Identify the CoG. Of the previously identified P17 assets, the EDR is the CoG analysis because of its inherent ability to thoroughly protect systems across the enterprise, using input from network traffic inspectors and feeding log aggregators.

Identify critical capabilities (CC). P17's critical capabilities for EDR include blocking intrusion attempts, sending alerts, conducting queries, and quarantining infected systems.

Identify critical requirements (CR). CRs for P17 to block intrusion attempts include possessing updated indicators of compromise (IOCs) (i.e., threat signatures) and having the EDR agent installed on workstations.

Identify critical vulnerabilities (CV). P17 examples of critical vulnerabilities would be corrupted IOCs or workstation operating systems that are incompatible with a particular EDR application.

Enumerate threat capabilities (TC). With respect to our running example, representative TCs against corrupted updates include the ability to tamper with or man-in-the-middle IOC updates.

Enumerate threat requirements (TR). For P17, TRs include physical access or remote access to an update mechanism.

Develop an actionable defense plan (ADP). One mitigation strategy in P17's ADP verifies the integrity of updates from vendors before applying them to the EDR.

B Survey instruments

Full versions of the pre-intervention survey, post-intervention survey, and follow-up survey are viewable at `ter.ps/nycsurvey1`, `ter.ps/nycsurvey2`, and `ter.ps/nycsurvey3` respectively.

C NYC leadership panel questions

We asked our panel of NYC3 leaders to answer the following questions for each participants' post-training survey results.

1. How likely is the identified asset the critical enabler for the participant's responsibilities? Please use a scale from 0 to 5, with 0 being "extremely unlikely" and 5 being "extremely likely"

2. How likely would the identified vulnerabilities stop the participant from fulfilling their responsibilities? Please use a scale from 0 to 5, with 0 being "extremely unlikely" and 5 being "extremely likely"

3. How likely would the identified threats exploit the vulnerabilities and prevent mission fulfillment? Please use a scale from 0 to 5, with 0 being "extremely unlikely" and 5 being "extremely likely"

4. How likely would the plan of action mitigate threats from exploiting the critical vulnerabilities? Please use a scale from 0 to 5, with 0 being "extremely unlikely" and 5 being "extremely likely"
5. Is the proposed defense plan sufficiently detailed to implement? Please respond with yes, no, or unsure.

D Visualizing Center of Gravity

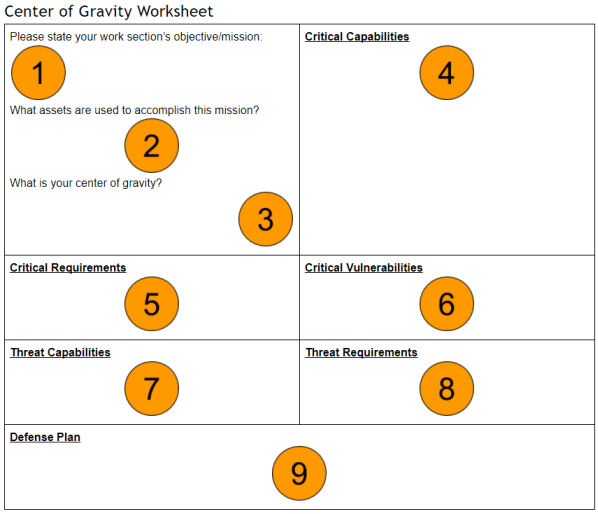

Figure 6: Depiction of CoG analysis tabular method.

Each participant received a printed version of the worksheet shown in Figure 6 to help guide them through CoG analysis. Numbers indicate the order in which participants completed the form, as described in Section 2.2. Additionally, we provided participants with a digital version of this worksheet during all online surveys. A more detailed version of the worksheet is available at: `https://goo.gl/icVMLX`.

Some participants opted to use a whiteboard to visually depict their thought processes and building heterogeneous, relational linkages between nodes. As shown in Figure 7, P18 began by writing his objective to "protect" networks. P18 then mapped how firewalls, EDRs, deep-packet inspection tools, and other defensive techniques support this objective. The commonality among all of these tools is that the defender uses cues from alerts to respond to incidents; thus, "alerts" are P18's CoG.

E CoG Identification Accuracy Regression

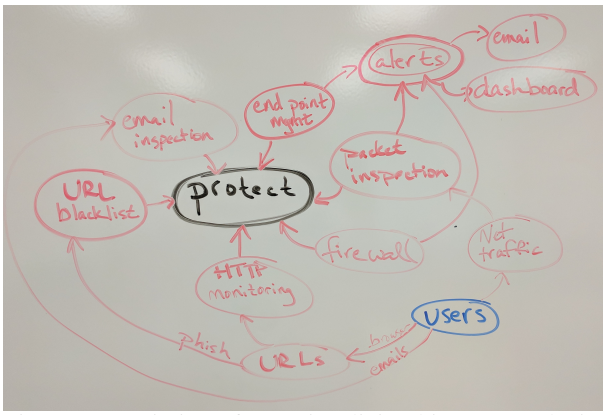

Figure 7: Depiction of P18 visualizing his CoG analysis.

Variable	Value	Odds Ratio	CI	p-value
IT Exp.	0-5 yrs	–	–	–
	6-10 yrs	0.17	[0, 11.36]	0.408
	11-15 yrs	3.82	[0.26, 55.28]	0.325
	16-20 yrs	0.74	[0.04, 12.16]	0.83
	21-25 yrs	0.39	[0.01, 20.26]	0.643
	26+ yrs	0.26	[0, 60.44]	0.626
Edu.	Some College	–	–	–
	Associates	3.02	[0.03, 289.4]	0.634
	Bachelors	3.51	[0.25, 49.43]	0.352
	Graduate	4.64	[0.21, 100.14]	0.327

*Significant effect – Base case (OR=1, by definition)

Table 2: Summary of regression over participants' accuracy at identifying centers of gravity with respect to their years of experience and education.