



CoSINT: Designing a Collaborative Capture the Flag Competition to Investigate Misinformation

Sukrit Venkatagiri*

sukritv@uw.edu

Center for an Informed Public,
University of Washington
Seattle, WA, USA

Department of Computer Science,
Virginia Tech
Arlington, VA, USA

Anirban Mukhopadhyay

Department of Computer Science,
Virginia Tech
Blacksburg, VA, USA
anirban@vt.edu

David Hicks

School of Education, Virginia Tech
Blacksburg, VA, USA
hicks@vt.edu

Aaron Brantly

Department of Political Science,
Virginia Tech
Blacksburg, VA, USA
abrantly@vt.edu

Kurt Luther

Department of Computer Science,
Virginia Tech
Arlington, VA, USA
kluther@vt.edu

ABSTRACT

Crowdsourced investigations shore up democratic institutions by debunking misinformation and uncovering human rights abuses. However, current crowdsourcing approaches rely on simplistic collaborative or competitive models and lack technological support, limiting their collective impact. Prior research has shown that blending elements of competition and collaboration can lead to greater performance and creativity, but crowdsourced investigations pose unique analytical and ethical challenges. In this paper, we employed a four-month-long Research through Design process to design and evaluate a novel interaction style called collaborative capture the flag competitions (CoCTFs). We instantiated this interaction style through CoSINT, a platform that enables a trained crowd to work with professional investigators to identify and investigate social media misinformation. Our mixed-methods evaluation showed that CoSINT leverages the complementary strengths of competition and collaboration, allowing a crowd to quickly identify and debunk misinformation. We also highlight tensions between competition versus collaboration and discuss implications for the design of crowdsourced investigations.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative and social computing systems and tools**; *Interaction techniques*; **Empirical studies in HCI**.

*The majority of this work was completed while the author was at Virginia Tech.



This work is licensed under a Creative Commons Attribution-NonCommercial International 4.0 License.

DIS '23, July 10–14, 2023, Pittsburgh, PA, USA

© 2023 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-9893-0/23/07.

<https://doi.org/10.1145/3563657.3595997>

KEYWORDS

crowdsourcing, CTF, capture the flag, competition, collaboration, communitition, design, research through design, misinformation, OSINT

ACM Reference Format:

Sukrit Venkatagiri, Anirban Mukhopadhyay, David Hicks, Aaron Brantly, and Kurt Luther. 2023. CoSINT: Designing a Collaborative Capture the Flag Competition to Investigate Misinformation. In *Designing Interactive Systems Conference (DIS '23)*, July 10–14, 2023, Pittsburgh, PA, USA. ACM, New York, NY, USA, 22 pages. <https://doi.org/10.1145/3563657.3595997>

1 INTRODUCTION

Vast amounts of publicly available data and powerful software tools have fueled the growth of crowdsourced investigations. These crowdsourced investigations have had significant real-world impact, from identifying mis- and disinformation during elections [108, 114, 126] to uncovering human rights abuses in war zones [42], among other examples [128, 145]. A small but growing subset of crowdsourced investigations have started to follow an ethos of analyzing only publicly available data, known as open source intelligence or OSINT [40].

Crowdsourced OSINT investigations follow two main interaction approaches: collaborative and competitive [12]. Examples of collaborative crowdsourced investigations include Amnesty International's Digital Verification Corps¹ and the University of California, Berkeley's Human Rights Center Investigations Lab.² The two organizations leverage a crowd of trained university students to collaboratively authenticate information on war crimes and investigate human rights abuses [103, 118]. Collaborative crowdsourced investigations have also helped to uncover misinformation [4, 64, 94, 114] as well as identify suspects involved in crimes [137, 139].

Competitive crowdsourced investigations often follow a capture the flag (CTF) model, a gamification concept borrowed from cybersecurity. In CTF competitions, teams compete against each other

¹<https://citizenzevidence.org/>

²<https://humanrights.berkeley.edu/home/>

to capture digital “flags” and score points to win the game. CTFs attract thousands of participants yearly, engaging in millions of hours of collective effort [26, 34]. However, most CTFs are designed to be theoretical in nature [66], with little emphasis on addressing real-world problems. For example, Hacktoria’s CTF seeks to introduce users to the investigative field and community of open source intelligence (OSINT) by helping them develop OSINT skills in a safe environment using simulated data [54]. Unique among CTFs, TraceLabs’ Search Party OSINT CTF seeks to address a real-life problem. TraceLabs’ members practice their OSINT skills by searching for information about missing persons, which is then submitted to law enforcement as tips [30].

The HCI and CSCW community has extensively studied collaborative [12, 24, 105] and competitive [12, 89, 132, 143] approaches for crowdsourcing complex work, with each having unique advantages and disadvantages. Collaborations benefit from increased communication frequency and a greater sense of group cohesion; but can suffer from overly rigid hierarchies and roles, as well as a need for increased articulation work [121]. Competitions benefit by introducing a greater sense of urgency and strongly motivating novices to participate [21, 66, 143], but can suffer from information silos and redundant effort, since each team may perform the same task and not communicate with one another [132].

Research on data science contests [131] and innovation contests [61] has found benefits to intentionally combining elements of collaboration and competition such as greater performance and creativity. Less work has studied collaboration and competition in crowdsourced OSINT investigations [e.g., 12] or sought to design hybrid approaches. Existing hybrid approaches cannot be directly applied to crowdsourced investigations because the latter pose different analytical and ethical challenges [44]. Data science contests require generating better performing models, while innovation contests require designing novel solutions. In contrast, investigations require synthesizing existing information into a coherent theory or conclusion [44, 112], with a greater focus on accuracy over performance and novelty. Further, improperly conducted crowdsourced investigations can result in immediate (versus delayed) harm to individuals through misidentification and harassment [e.g., 25, 76, 90, 102, 137].

In this work, we explore how to merge the complementary strengths of competition and collaboration in a new domain: crowdsourced investigations of online misinformation. We employed a four-month-long Research through Design (RtD) process [152] with 46 university students to design and evaluate a novel interaction style called *collaborative capture the flag competitions* (CoCTFs). We instantiated the CoCTF concept through CoSINT, a platform that enables a trained crowd to work with professional investigators to identify and investigate social media misinformation.

To ameliorate the disadvantages of competition, such as duplication of effort and information silos, CoSINT incentivizes information sharing and collaboration between competing teams. To reduce the disadvantages of collaboration, such as increased articulation work and inflexibility, CoSINT serves as a coordinative artifact [122] and gives the crowd greater agency to determine how to combine techniques and tools with scaffolding and rubrics from the field of OSINT. The competitive, gamified setting also provides additional motivation for the crowd. Finally, to augment

the crowds’ abilities and mitigate unwanted (unethical) behavior, they are provided with expert training and guidance.

Through our mixed-methods evaluation, we found that CoSINT enabled participants to quickly discover, archive, verify, and report on hundreds of pieces of potential misinformation on social media. Participants also said that they enjoyed using CoSINT and that it helped to better structure their workflows as they worked within their team and with other teams. Our RtD process and mixed-methods evaluation also highlighted tensions between competition versus collaboration, and in-depth versus broad investigations.

In summary, our work makes four contributions:

- (1) Our paper makes a conceptual contribution by introducing collaborative capture the flag competitions (CoCTFs) to support a rapid response to investigate misinformation online.
- (2) The CoSINT platform makes a system contribution by operationalizing CoCTFs, extending the small but growing number of Research through Design studies that address misinformation [8, 85, 98, 150].
- (3) Our semester-long mixed-methods evaluation with 46 students showed that CoSINT blended beneficial elements of competition and collaboration, and that participants enjoyed using CoSINT to structure their investigations.
- (4) Through reflection on our design process and system evaluation [152], we present design implications for supporting collaborative competitions in other high-stakes settings.

2 RELATED WORK

To situate our work, we review prior research on approaches to investigate mis- and disinformation; systems for supporting collaborative and crowdsourced sensemaking; and ways to introduce collaboration into competitive environments.

2.1 Open Source Intelligence Systems to Investigate Mis- and Disinformation

Prior work identified three approaches to addressing misinformation online: agent-, message-, and interpreter-oriented [142]. Agent-oriented approaches are concerned with the specific actors that generate and spread misinformation, while receiver-oriented approaches are focused on the targets of misinformation and how receivers are affected. Each approach can also be focused on individual or a collection of: agents, messages, and receivers [142].

Most closely related to our work are message-oriented approaches concerned with: a) identifying content that is potential misinformation [e.g., 109] and b) verifying or refuting claims associated with or made by the content [e.g., 50]. While the growth of information and communication technologies has made misinformation more prevalent, it has also democratized access to sensitive data and powerful tools to analyze it [49]. This has led to a well-established investigative field of open source intelligence (OSINT). OSINT has been widely used by media agencies [42], civil society organizations [134], governments [146], and online sleuths [7] to investigate misinformation.

Prior work [12, 146] identified four steps to any OSINT investigation, called the OSINT cycle: 1) discovering content; 2) verifying its provenance and determining the veracity of claims; 3) archiving the content to prevent information from being lost; and 4) reporting

on the results of the investigation. Researchers have also developed software systems to support the OSINT cycle's four steps. For example, CrowdTangle [47] and Algorithm Tips [36] support automated content discovery, Hoaxy [123] and DejaVu [94] support verification, the Web Archive Workbench [59] supports archiving, and Birdwatch [147] and Maltego [56] support reporting. Given the growing number of OSINT systems, Abdullah et al. [1] created OSINT Explorer, a system to help analysts determine which OSINT tool to use.

Prior systems largely focused on individual steps in the OSINT cycle, whereas CoSINT is designed to support all four steps. CAPER [3] is another tool that supports all four steps to help law enforcement agents prevent organized crime. Different from CAPER, CoSINT supports investigators in investigating misinformation on social media. Although the 'C' in CAPER stands for 'collaborative,' it only allows sharing of *files* between different law enforcement agents, and does not support collaborative *work*. Unlike CAPER and most other OSINT tools, CoSINT supports virtually any number of users collaborating within and across teams on one or more investigations. Lastly, while investigations in general have been well-studied within CSCW and HCI, OSINT in particular — despite its popularity in practice — has garnered less attention [3, 12, 62]. Through this work, we show that OSINT can be a valuable framework for the CSCW and HCI community.

2.2 Systems for Supporting Collaborative and Crowdsourced Sensemaking

Investigations involve a complex sensemaking process [2, 9, 138], where investigators (e.g., journalists, historians) must collect, analyze, and make sense of disparate sources of information to arrive at a theory or conclusion presented in a final report. This sensemaking process also closely maps onto the OSINT cycle [12, 146], which involves similar steps (discover, archive, verify, and report).

Investigators face challenges in adequately using OSINT tools and techniques due to limited time, personnel, access to data (e.g., reach data, follower networks, content metadata), or the data science expertise required to analyze data [95]. Further, many OSINT tools and techniques become obsolete because of changes in online environments and social media platforms. These challenges, coupled with the complex and dynamic nature of misinformation investigations [126], result in many investigations being under- or unexplored.

To scale up and speed up their work, investigators turn to collaboration and crowdsourcing. Collaboration can support sensemaking by dividing discovery and verification tasks and providing diverse perspectives when analyzing data and generating reports [46]. Yet, collaboration also comes with coordination challenges: collaborators may be geographically distributed, have different skills and backgrounds, and have access to different data sources and tools [45, 48, 52, 65].

Prior research on supporting collaborative sensemaking has focused on co-located teams (2–10) and crowds (30+) [60, 67, 100, 137, 140] working synchronously, or distributed crowds (30+) working asynchronously [32, 35, 46, 81, 91]. We extend prior work by studying how to support a distributed crowd working *synchronously*.

Specifically, we study a semester-long field deployment with 46 university students who conduct investigations into misinformation online in a virtually co-located classroom.

Most closely related to our work are studies of the effectiveness of crowdsourced sensemaking and fact-checking. For example, Dailley and Starbird [32] found that large, distributed crowds engaged in collective sensemaking and rumor-rumoring during crises on social media. Arif et al. [9] showed that distributed crowds were able to successfully correct rumors online, and Saeed et al. [120] showed that crowdsourced fact-checking on Twitter frequently performed as well as professional fact-checkers. Apart from observational studies, researchers have also experimentally studied the effectiveness of crowdsourced fact-checking. Pennycook and Rand [109] as well as Allen et al. [5] found that crowdsourced trustworthiness ratings can distinguish between real and fake news sources, though Godel et al. [50] found that real-time crowdsourced veracity ratings performed worse than professional fact checkers. However, these prior crowdsourcing approaches studied online crowds working asynchronously or independently. Emergent collective behavior and asynchronous collaboration pose coordination challenges [9, 32], and aggregating independent ratings from novice crowds [e.g. 50, 109] does not allow members of the crowd to build consensus or learn from each other over time. CoSINT overcomes these limitations by leveraging trained crowds that synchronously collaborate with — and learn from — each other.

Scaling up collaborative sensemaking from small teams to large crowds can amplify coordination challenges [131]. To address these challenges, prior crowdsourcing systems have focused on dividing work into microtasks, such as collecting [106], extracting [23, 138], or schematizing [27, 69, 88] data. Other projects have crowdsourced all parts of the sensemaking loop to support more complex work, such as unraveling mysteries [81, 82] or drafting articles [6, 11, 14, 55]. Like this latter group of systems, CoSINT supports the entire sensemaking loop, with a focus on investigating misinformation. Still, Retelny et al. [116] found that rigid crowdsourcing workflows constrain complex and creative work. Instead, CoSINT leverages Retelny et al. and Alharthi et al.'s [2] suggestion of including more flexible rules and roles, sharing information obtained individually with others, and fostering social interaction within the group. CoSINT allows crowd workers to choose what to collect and archive, which verification tools and techniques to use, and how to structure the final report.

CoSINT differs from prior crowdsourcing systems and approaches in two additional ways. First, while prior work has enabled and enhanced collaborative sensemaking environments, our work begins with a competitive sensemaking environment where we introduce collaborative elements. Second, traditional crowdsourcing systems are often fixed in their functionality, relying on the developers of the system to introduce new features. CoSINT, though, handles complexity by leveraging appropriability [37, 51]. Roles and workflows are flexible and dynamic, and additional automated tools and crowdsourcing techniques can be integrated and configured through CoSINT's API.

2.3 Designing Collaboration Into Competitive Environments

Competition can benefit endeavors through providing an increased sense of urgency, immersion, and motivation [107, 148]. There are several types of competitive environments, ranging from hackathons [113], capture the flag competitions (CTFs) [21, 28], and innovation contests (ICs) [73, 132], to games [77, 101] and even self-competitions [97]. However, while competitions such as ICs and hackathons have been used to address misinformation, one notable exception is CTFs.

In CTFs, teams compete against each other by “capturing flags,” either from other competitors (attack / defense-style) or from the organizers (jeopardy-style) [21]. The team that captures the most flags or the highest cumulative value of flags — where different flags are worth different numbers of points — wins the CTF. Besides outdoor sports, CTFs are perhaps best-known in the field of cybersecurity [21, 96, 143], requiring programming expertise to solve cryptographic puzzles, make database and network queries, or uncover exploits in operating systems.

Prior work has shown that CTFs are best-suited for settings focused on collecting new information or uncovering new problems [66]. Researchers have also shown that CTFs introduce a sense of urgency and strongly motivate novices to participate [21, 96], while also helping novices learn through hands-on experience [28]. CoSINT leverages these features to enable a novice crowd to rapidly respond to misinformation at the message level, where a large quantity of content must be discovered and verified.

However, most CTFs are designed to be theoretical in nature [66], with little emphasis on solving real-world problems. This may partly be due to the origins of some CTFs in the cybersecurity field [28, 124], where unauthorized access of real-world computer systems is illegal in the U.S. [41]. Still, CTFs attract tens of thousands of participants yearly, engaging in millions of hours of collective effort. For example, DEF CON’s CTF attracted 3,229 teams from 2018 to 2021, each logging 276 hours of active game time [34]. Assuming three people per team, this amounts to over 2.5 million hours of effort spent on a theoretical competition that, while valuable per se for training and other other purposes, could alternatively have directly addressed a real-world problem.

Most closely related to our work, Trace Labs’ *Search Party OSINT CTF* is a rare instance of a real-world oriented CTF, with the goal of collecting information to help law enforcement find missing persons [30]. In 2022, Trace Labs’ CTF attracted 250 teams that collectively made nearly 4,000 submissions in four hours [135]. CoSINT builds on Trace Labs’ jeopardy-style model of assigning flags of greater strategic importance more points. In Trace Labs’ model, flags are independent of each other (i.e., there are no successive challenges to be completed), whereas in CoSINT, flags build on each other, starting with a discovery flag and ending with a reporting flag. Relatedly, Belghith et al. [12] studied OSINT organizations that exhibited elements of competition and collaboration (“social OSINT”) and found two other limitations of CTFs, and competitions more generally: 1) duplication of effort and 2) siloed information. These limitations are acceptable in theoretical environments, but may be less desirable in real-world investigations. Trace Labs’ CTF, despite

its real-world orientation, would also benefit from addressing these challenges.

Prior work has found that introducing collaboration within a competition can help overcome these limitations. Two types of collaboration can exist in a competition: intra-team and inter-team. While intra-team collaboration naturally benefits a single team, research has shown that inter-team collaboration can be beneficial even in competitive environments [132]. Analyzing 25 data science ICs on Kaggle, Tausczik and Wang found that sharing code between teams improved each individual teams’ performance [132]. In a design IC, Hutter et al. also found that teams who engaged in “communitition,” community-level collaboration among competing teams, made higher quality submissions and were more likely to win [61]. Different from prior CTFs, CoSINT includes beneficial elements of communitition found in ICs, such as sharing work products and building on others’ solutions. CoSINT also focuses on a novel domain — crowdsourced investigations — which poses unique analytical and ethical challenges [44].

3 DEVELOPING COSINT USING RESEARCH THROUGH DESIGN

Misinformation on social media is a wicked problem [117] because it is a symptom of another problem (e.g., political polarization or psychological biases), it can be interpreted and solved in many different ways (e.g., social, psychological, or technological), and solving it is identical to completely understanding it and there are no clear criteria for sufficient understanding [80, 99, 119].

To address wicked problems, Zimmerman and Forlizzi [152] propose an approach to conducting research called Research through Design (RtD). RtD is the process of iteratively designing and critiquing an artifact that acts as a proposed solution to a wicked problem. Solutions to wicked problems are not right or wrong, but “good” or “bad” depending on the initial framing [13]. Engaging in RtD enables researchers to investigate what a potential future might look like so as to reframe the wicked problem [153]. Hence, RtD has become a well-established design method in HCI [10, 13, 152, 153].

Artifacts produced through RtD can also help inform new theory and future work, provided the process is well-documented [17, 33, 71, 153], including “detailed documentation of the actions and rationale for actions taken during the design process” [152]. We thus describe the four phases of our RtD process for developing and evaluating the CoSINT platform: 1) frame the problem; 2) prototype workflows and interfaces; 3) clarify design goals; and 4) deploy, iterate on, and evaluate the design.

3.1 Phase 1: Frame the Problem

In response to Carroll et al.’s criticism that the *thing* often proceeds *theory* in HCI [22], Zimmerman and Forlizzi suggest that things in RtD should be informed by current theory and practice while spawning new theory and practice through the design and evaluation process [152]. Here, we describe our engagement with findings from prior work as well as practice.

Recently, RtD has been applied to address misinformation by envisioning new design artifacts and interaction modalities [e.g., 8, 85, 98, 150, 150]. For example, Zade et al. [150] employed an

RtD process to design contextual cues to inform credibility assessment on social media; while Løvlie et al. [85] designed a tool to help readers better understand evidence and uncertainty in science journalism. The authors of this paper also have complementary experience in: 1) designing and evaluating crowdsourcing systems to support sensemaking [e.g., 81, 137, 138], 2) studying competitive and collaborative online communities [e.g., 12, 86, 87, 149], and 3) participating in large-scale collaborative and competitive investigations, including hackathons and OSINT CTFs.

Through our prior experiences and engagement with the literature (Section 2), we found three influential themes. First, expert investigators who seek to investigate misinformation in real-time require additional resources, such as personnel [95]. Second, while competitive and collaborative crowdsourcing have been used to support investigators in combating misinformation [12], little work has explored how to combine both approaches in this context. Third, CTFs show promise in attracting and motivating large, novice crowds through non-monetary incentives [21, 26]. However, it is unclear how to adapt CTFs that are traditionally theoretical for a real-world application, as well as how to introduce elements of collaboration into a CTF. These themes led to the following research question:

RQ. How can we merge the complementary benefits of competition and collaboration to provide a rapid response to investigate misinformation?

3.2 Phase 2: Prototype Workflows and Interfaces

To inform the design of the CoSINT platform, we used existing social computing systems to piggyback prototype [53] interaction workflows and interfaces over one month with ten members of our lab acting as a small crowd. Our prototyping approach is common in classroom-oriented crowdsourcing research [e.g. 38, 110, 136, 151].

Piloting Investigation Structure. We needed to determine how to structure the crowd's investigation into misinformation. Based on prior work [74, 130, 138], we chose a format that involved the crowd identifying potential misinformation, followed by verifying or refuting claims made by that content. We focused on misinformation on social media platforms given experts' need to quickly verify or refute claims made on social media before a post "goes viral" [64, 127, 138]. Note that *potential misinformation* refers to content that has not been debunked (either verified or refuted), but appears to be falsifiable. To verify or refute claims, we used open source intelligence (OSINT) techniques, given that they are transparent and publicly accessible, and followed the OSINT cycle to structure our investigations [12, 124, 146].

In our pilot deployments, we provided the crowd with topics to investigate, such as *COVID-19*, *election security*, and *financial misinformation*. Then, we asked the crowd to search for these topics on social media platforms, such as Twitter and Facebook, and identify potential misinformation that was recently posted and appeared to reach a wide audience (i.e., large number of shares/retweets). We prototyped different information management systems (e.g., Slack, online forms, collaborative documents, and collaborative spreadsheets). We ultimately settled on a Google Forms front-end for the

crowd to quickly submit potentially misinformative social media posts in a structured manner.

Piloting CTF Structure. Next, we needed to understand how to apply CTFs to a real-world investigative setting, given that CTFs are typically used for theoretical investigations. CTFs are often structured as individuals or teams competing against each other to capture "flags" and score points, with the highest-scoring team winning the competition. Capturing a flag in CTFs can represent diverse actions, e.g., identifying a piece of information, solving a cryptographic puzzle, or completing some other task. These tasks can be independent or interdependent, i.e., solving one puzzle to use the results in the next task. Many CTFs incorporate human judges to evaluate creative or subjective tasks that cannot be automatically evaluated.

Teams. Given that our ultimate goal was to make the CTF more collaborative, we randomly assigned students into teams of two or three, with each team containing at least one political science and one computer science major. To help streamline each team's efforts, we also asked each team to nominate a team leader and choose a unique topic to investigate.

Flags. After consulting with the literature on OSINT and existing CTF structures, we chose four different flag types that closely map onto the OSINT cycle [12, 124, 146]: 1) discovery flags — the task of identifying content that is potential misinformation, 2) verification flags — verifying or refuting claims made within the content, 3) archival flags — permanently archiving the discovered content and any associated information, and 4) reporting flags — writing a report of the investigative process to enable transparency and reproducibility.

Points. To make it easier to compare performance across teams and to make the evaluation more objective, we assigned point values to flags, with each flag type being worth 20 points. We heightened the sense of urgency and competitiveness in the CTF by creating a leaderboard in Google Sheets. The leaderboard displayed the cumulative points earned by each team in a graph that updated every 15 minutes.

Judges. We introduced the concept of judging, where expert investigators awarded points and provided feedback on flags created by the crowd, for two reasons. First, the subjective nature of identifying potential misinformation and verifying or refuting it necessitated expert evaluation. Second, prior work has shown that judges can increase the quality of the crowd's work by providing frequent feedback [39, 92].

Judges assessed whether a flag was relevant to the topic chosen, if the contents were falsifiable, and if the crowd worker accurately debunked claims made within the content. In our pilot deployments, the authors acted as judges due to their prior expertise with conducting such investigations. To help judges evaluate crowd workers' submissions, we asked crowd workers to include a link to a Google Doc that described their investigative process in greater detail than was included in the submission form. To provide feedback, we used Google Doc's comment feature, while teams' cumulative points were tabulated and visualized in a leaderboard using Google Sheets.

Findings from Prototyping Process. From our prototyping process, we identified four themes that would need to be addressed in subsequent versions of our system. First, lab members said that

certain types of content were more difficult to discover, and that some verifications required more time and effort than others, so they should be worth more points. Second, we found that judging flags was time consuming. We decided to make judging faster by adding more structure to the flags.

Third, we found that some teams performed better than others due to differences in their composition (e.g., technical vs. topical expertise) and tools used (e.g., using reverse image search vs. manual searches). Fourth, we learned that most teams were often unaware of what other teams were working on. Taken together, these two findings align with prior work in other domains (data science [131] and innovation [61]) indicating that better performance may be achieved by allowing competing teams to work with (and against) each other.

3.3 Phase 3: Clarify Design Goals

After settling on the general format of our CTF investigation, we engaged in a cyclic process of brainstorming and designing possible features for a collaborative CTF (CoCTF) that would overcome the limits of traditional competitions by introducing beneficial elements of collaboration. To illustrate what these features would look like, we created low-fidelity and high-fidelity interface mockups (see Appendix A for examples), and solicited feedback from lab members.

By combining prior work and our experiences with our prototyping process, we identified two design goals: 1) support a rapid-response to investigating misinformation; and 2) give the crowd agency. However, it is challenging to design software that meets these goals in a complex setting involving a large number of people coordinating their actions [51].

One promising approach in CSCW for navigating this complexity is designing for *appropriation* [37]. That is, instead of trying to understand or anticipate all of the features of a system, we can design solutions that can be used in diverse and dynamically reconfigurable ways — thus creating more robust solutions for complex problems [51]. We instantiated our design goals by leveraging four of Dix's heuristics for software appropriation [37]. Dix suggests in his first heuristic that designers *expose the intentions* behind the system, that is, making design assumptions and decisions explicit, and “if they are wrong then they [can] be re-examined” [37]. Along these lines, we explicitly exposed the intentions behind the CoSINT platform through our two design goals:

Goal 1. Our primary goal for CoSINT was to enable a rapid-response to investigating misinformation on social media. We instantiate this goal by using two of Dix's other heuristics: *encourage sharing* and *provide visibility*.

Goal 2. Our second goal was to give the crowd agency. This is in contrast to typical crowdsourcing systems where crowds are given specific microtasks, with limited agency in how to complete them [18, 92]. This follows Dix's fourth heuristic to *support not control* users' actions, but to provide necessary functionality so users can achieve their goals without detailed instructions.

In the next section (Section 4), we describe how we designed CoSINT to meet these two goals.

3.4 Phase 4: Deploy, Iterate on, and Evaluate the Design

Having prototyped our designs and clarified our two design goals, we implemented them as a functional software prototype — the CoSINT system — and deployed it with our intended user base: a crowd of 46 students trained to investigate misinformation.

RtD creates new situations and practices for researchers to investigate, producing gaps in behavioral theory and technical opportunity [152]. Thus, our evaluation process was continuous and sought to better understand emergent sociotechnical interactions enabled by CoSINT and opportunities for further improvement.

Research and Class Setting. Due to the sensitive and contextual nature of investigations into misinformation, we sought crowd workers with whom we could build trust and foster accountability. In our design goals (Section 3.3), we also noted the importance of providing the crowd with more agency and allowing them to take on complex tasks through extensive training. We thus decided to deploy CoSINT in a semester-long course at our university and evaluated it using mixed methods. All authors helped to design the class and teach students OSINT investigative skills through hands-on training. This enabled a tight coupling between the skills students learned in class and how they applied them while using CoSINT. Our use of a “class as a crowd” is well-established, both within the domain of investigations [e.g., 103, 118] and in other domains [e.g., 38, 110, 136, 151].

For our evaluation, we required students with topical expertise in misinformation and technical expertise in testing and developing software. We thus recruited senior undergraduate and graduate students in the Computer Science and Political Science departments. The class met online due to the COVID-19 pandemic. Twice a week for approximately 90 minutes, we taught students techniques and tools in each of the four steps of the OSINT cycle. Modules for each step lasted approximately three to four weeks.

Classroom Deployment and Iteration Procedure. As we learned how students used the system, and analyzed their usage and feedback, we continued to iterate on the design of the software prototype for three months, until we arrived at a final system design (described in Section 4). Here, we briefly describe major system changes.

We demonstrated CoSINT to the class in week six and described our motivation for creating it. We also asked students to engage in an iterative, participatory design process with us. Our deployment procedure involved three steps that we repeated every two weeks (for a total of six deployments between week six and sixteen): 1) demonstrate CoSINT features and have students take part in a CTF for the entire class; 2) solicit student feedback and reflect on the CTF and system features by ourselves; 3) implement and refine new features.

For the pilot deployment, students were placed into teams through random assignment. For the final deployment and iteration during weeks six to sixteen, students self-organized into teams of four or five members to work on the class project and compete in the CTFs. To enable group cohesion and familiarity, teams remained fixed throughout the rest of the semester. To leverage the complementary technical and domain expertise, each team was

required to have least one political science and computer science major.

Eleven research assistants and two senior researcher served as judges across the five events, with three to six individuals per event. All research team members had prior experience with open source investigations and helped with the prototyping of CoSINT. Most also had experience with participating in capture the flag competitions. For events 3 and 4, we also recruited one student from each team (different each time) to serve as a judges to gain valuable experience evaluating other teams' flags. Student judges were not allowed to evaluate their own team's flags.

The major changes during the classroom deployment involved finding a balance of points for different flag types and for incentivizing competition versus collaboration. Based on student feedback, we modified the rubric categories and refined the user interface. Due to space restrictions, we only describe the final version of the CoSINT system in the next section.

Participant Recruitment, Data Collection, and Analysis

Methods. This study was approved by our university's IRB. The first author recruited students during an in-class guest lecture and stated that participation was voluntary. We provided consenting participants with \$20 gift cards. For the final CTF, we recognized the three top-scoring teams with prizes of \$55, \$45, and \$35. We collected both qualitative and quantitative data during and after each deployment. This included participant observation with detailed notes [125], notes from our weekly feedback and reflections, semi-structured interviews, and system log data.

Twenty out of 46 students (from six out of 11 teams) participated in our study, though we report system log data on anonymized aggregate results for all teams. The median age was 21 years (range = 19–23). Fifteen (of 20) participants majored in computer science or similar fields, and four majored in political science or similar. Six identified as women, and 14 as men. None of the political science students had participated in CTFs prior to the class, while six of the computer science students had.

Semi-structured interviews. The first author interviewed 16 of the 20 students who provided consent (the other four were unavailable) and took detailed notes [125], which were incorporated into the transcripts. The interview guide contained questions about how students worked with their team members and other teams, their perceptions of the platform and how it had changed over time, their feedback on various system components, and reflections on what they learned.

In total there were eight interviews with members from six different teams, with one to four team members in attendance for each. All participants were interviewed immediately before (KG6-9, SS-4, SS-5, OT-1) or after (SL-2, SL-3, KP14-17, DD-13) the final CTF. We interviewed DD-10 and DD-11 before and after the final CTF because they wanted to provide additional feedback. The first author recorded all interviews using Zoom and fully transcribed the recordings. Interviews ranged from 58 to 85 minutes (average = 65).

System log data. To gain insight into how teams performed using CoSINT, we collected system log data. We received IRB approval to analyze anonymized, aggregated (at the team level) system log data

from students who did not consent to participate in the interview portion of our study.

Data analysis. To analyze our data we conducted a deductive thematic analysis [19] of the transcripts, based on themes relevant to our research questions and the various system components. These themes largely aligned with the structure of the interview guide. After downloading and fully anonymizing the system log data, we analyzed log data for all but the first event using the pandas and numpy Python libraries. We omitted the very first event from our data analysis because we made significant changes to the interface and database schema, and students would not have had sufficient time to familiarize themselves with CoSINT.

4 THE COSINT COCTF PLATFORM

Here we describe the final design of the CoSINT CoCTF platform following our iterative classroom deployment process. CoSINT is designed to support experts and a team of crowd workers to rapidly respond to misinformation on social media. Teams compete against each other in an event by “capturing” flags to score points, and the team with the highest score at the end of the event wins the CTF. These points are shown in a leaderboard that dynamically updates after a team completes an action.

There are four possible actions that correspond to four flag types. Teams can 1) discover potential misinformation that makes falsifiable claims (discovery flag), 2) archive the content making those claims (archival flag), 3) verify or refute those claims and document their verification process (verification flag), and 4) write up a short report on their findings (reporting flag). Each flag type is worth a different set of points that varies according to a rubric that prioritizes higher quality flags.

Experts specify different narrative threads (topics) of misinformation to investigate. Teams document a rumor or potential misinformation as an evidence piece that consists of one of each flag type. An evidence piece can consist of one or more of each type (except for a discovery flag, where there is a one-to-one mapping). Competing teams can also collaborate with each other by contributing flags to other teams' evidence pieces. Experts act as judges to evaluate teams' flags against the rubric and award points.

4.1 System Description

We now describe how the CoSINT platform (Fig. 1) can be used to conduct an investigation into social media misinformation using a fictional scenario based on a real-life event³. Jane, the *expert*, is an investigative journalist who works for a large news agency. Last night, a fire broke out in a warehouse in Beirut, Lebanon. Minutes later, two large explosions rocked the city, and were felt 150 miles away. Located several thousand miles away, Jane cannot obtain drone footage herself; and with the recentness of the event, updated satellite imagery is unavailable. However, residents of the city quickly took to social media to share what they had witnessed. Rumors about the cause of the explosion also began to spread on social media. To investigate the event and debunk false and potentially harmful claims on social media, Jane must act fast. Jane requests help from two of her co-workers (*judges* in this scenario)

³<https://www.amnesty.org/en/latest/news/2021/08/lebanon-one-year-on-from-beirut-explosion-authorities-shamelessly-obstruct-justice/>

Team Name	Team Size	Participant [†]	Gender	Degree Major
OT	4	OT-1	Male	Political Science
SL	4	SL-2	Female	Computer Science
		SL-3	Male	Computer Science
SS	4	SS-4	Male	Computer Science
		SS-5	Male	Computer Science
KG	4	KG-6	Female	Political Science
		KG-7	Male	Computer Science
		KG-8	Male	Computer Science
		KG-9	Male	Computer Science
DD	4	DD-10	Male	Computer Science
		DD-11	Female	Political Science
		DD-12	Male	Computer Science
		DD-13	Male	Computer Science
KP	4	KP-14	Male	Computer Science
		KP-15	Male	Computer Science
		KP-16	Male	Computer Science
MH	4	–	–	–
TL	5	–	–	–
BF	4	–	–	–
KF	5	–	–	–
JE	4	–	–	–

Table 1: Table of all participants that used CoSINT over the semester, their team size, the [†] participants that we interviewed, their self-identified gender, and their degree major.

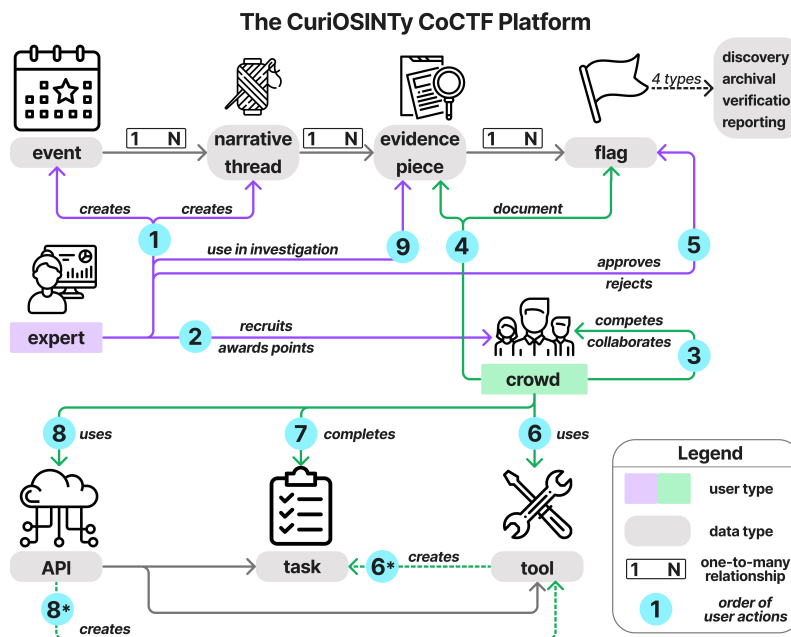


Figure 1: The structure of the CoSINT platform, described in Section 4. Screenshots of the platform can be found in Appendix A. The platform consists of events, narrative threads, evidence pieces, and flags, each hierarchically linked to the former. There are two types of users: experts and crowd workers. Experts perform actions (1), (2), (5), and (9). Crowd workers perform actions (3), (4), (6), (7), and (8). We omit further discussion of the API due to space constraints.

and a group of college students (*the crowd*), whom she has previously worked with. They use CoSINT to structure their work and collectively investigate the explosions in Beirut. Note: the steps below (e.g., '(1)') correspond to the numbered blue circles in Fig. 1.

(1) Create Event and Narrative Threads. In line with our second design goal, we incorporate Dix's fourth heuristic to *support not control* users' actions. CoSINT allows experts to provide some, but not complete, direction to the crowd. The expert specifies which misinformation narratives the crowd should track, and possibly which platforms to search. The crowd chooses how to search these platforms, which posts to investigate, and how to verify or refute a particular piece of misinformation. In this scenario, Jane first creates an *event* in CoSINT, followed by several distinct narrative threads for teams to focus their efforts on. This includes: potential causes of the explosion, imagery of the explosion, injuries and lives lost, recent news about the port, historical information about the port, among others.

(2) Recruit Crowd Workers and Judges. Through Jane's "OSINTvestigators" Discord group, she asks the college students to help her investigate. She also asks her colleagues, Alice and Bob, to serve as judges.

The students log in to the platform with their existing accounts and click on the events tab to access the current CoCTF event. The students form two teams (Gamma and Delta) and choose team leaders. The leaderboard displays all of the newly formed teams and their total points so far (currently zero).

(3) Compete and Collaborate. Teams Gamma and Delta simultaneously compete against each other for points, with the highest-scoring team winning the event. However, competitions suffer from informational silos and duplication of effort. To ameliorate these limitations and in line with our first design goal, we incorporated Dix's second heuristic: *encourage sharing*. Prior work studying competitions found that when communitition was encouraged (collaboration among competing teams), the individual [132] and collective performance of teams was higher [16, 73]. CoSINT incentivizes teams to collaborate with each other in three ways.

First, to prevent informational silos, all evidence pieces and flags are visible to all users, irrespective of team membership. Second, to enable members of Gamma and Delta to build on each others' work, any user can contribute a flag to another users' evidence piece — scoring a collaboration point bonus, along with the base value of points for that flag. Third, there may be instances when a particular task proves too cumbersome for any one team, or if a team does not have the requisite experience but another one does. To allow users to explicitly ask for help and support others, CoSINT provides a Mechanical Turk-like requester interface. This interface, described in Steps (6,7,8), lets users create microtasks that others can complete to score additional points.

On Discord, team leaders strategize with their team members on which narrative threads to address. To reduce context-switching and collaboration costs, Team Gamma (G1-4) assigns specific members to entire pieces of evidence. That is, one member will discover, verify, archive, and report on each evidence piece. Team Delta (D1-5) chooses to play off their members' strengths in discovery and verification. Delta assigns three members to discover and verify new evidence, while the other two members work on archiving the content and writing reports.

To avoid duplicate effort and further support inter-team collaboration and coordination, we incorporated Dix's third heuristic, *provide visibility*, to make clear how the platform works so that the users can devise their own uses. CoSINT not only makes all information accessible to all users, but it also displays the current status of various actions [121, 121], e.g., whether a piece of evidence is completed, or if a flag has been approved or rejected by a judge. Further, it shows users the maximum number of points they can score for each flag type and how many points a judge awarded them.

In our scenario, Team Gamma's leader, G1, views the evidence pieces that Team Delta is creating. She observes that Delta is focused on potential causes of the explosion. To avoid duplicate work, she directs her team to focus on historical information about the port. Halfway through the event, G3 learns of fertilizer storage facilities located near the explosion site, and that fertilizer is explosive. Instead of creating his own evidence piece, G3 contributes a verification flag to one of D4's evidence pieces, providing more evidence to refute a claim that the explosion was caused by a gas leak. Both G3 and D4 score points for the flags that they submitted, but also score additional points for collaborating with each other.

(4) Document Evidence Pieces and Flags. To meet our second design goal of providing the crowd more agency, CoSINT incorporates scaffolding and rubrics that enable novice crowds to match expert-level performance [29, 39]. CoSINT provides scaffolding by dividing each narrative into multiple evidence pieces focused on a particular claim. In turn, each evidence piece is divided into four different flag types (discovery, archival, verification, and reporting). However, the platform does not enforce 'hard' constraints on the exact process or order of completion, but encourages high quality submissions through the point system and through a self-assessment rubric.

In this scenario, D5 finds a video of the explosion on Twitter. To document this, D5 clicks on the New Evidence button. CoSINT prompts D5 to choose an associated narrative thread ("imagery of the explosion"), specify a name, and provide the URL of the original tweet. Simultaneously, on the same page, D5 documents the discovery flag for discovering that evidence piece. First D5 specifies the sub-type for his discovery flag: video. Next, to promote transparency in the investigative process, D5 is required to describe how they found this content. Finally, D5 self-evaluates the quality of their discovery flag. Once D5 clicks on the Add Evidence button, the new evidence piece is created.

Next, following the OSINT cycle [40], D5 adds verification, archival, and reporting flags; and G1 adds another verification flag. A typical evidence piece consists of one of each type, but also allows multiple of each type to accommodate more complicated evidence pieces with multiple claims. Verification, archival, and reporting flags have different evaluation schemes to incentivize factors that Jane determines are important. For example, the evaluation criteria for discovery flags includes: originality, influence, and recency. More original, influential, or recent discoveries are worth more points.

(5) Judge Flags and (2) Award Points. Although Jane trusts the college students she is working with, she leverages CoSINT's judging mechanism as a first-pass filter to focus her attention on the most relevant, urgent, and accurate evidence that teams identify.

Judging also gives the crowd feedback to improve the quality of flags they submit in the future. The self-assessment rubrics allow judges to evaluate flags faster, since a judge is shown the crowd worker's baseline, (hopefully) good-faith assessment of their flag.

In the scenario, Alice and Bob act as judges for the event. Judges can approve flags in any order, except for reporting flags, which can only be approved after the other three types of flags have been submitted and approved. In the first thirty minutes, teams documented several pieces of evidence and flags, which the judges begin to evaluate. Judges can view a user-submitted flag, the evidence piece it is part of, and the user's self-evaluation. A judge can then decide to approve or reject the flag, and modify the number of points that a team is awarded (compared to the original self-evaluation). For example, D6 submitted a verification flag where his self-evaluation totaled 600 points. However, Alice notices that some important details were missing — such as the time of day — awarding D6 500 points. Alice encourages D6 to submit a separate verification flag for the time of day mentioned in the original claim.

A team's points are calculated as the sum of its team members' points. If a judge rejects a flag or assigns a lower point value, a user can create a new flag with additional details and context. After D6 submits the time-of-day verification flag, he lets Alice know through the Discord group, asking her to evaluate it. Disagreements can also be clarified in a similar way.

(6,7,8) Create and Use Tools. CoSINT promotes appropriability and extensibility by providing an API that supports generic task, task-response, and reward formats, similar to Human Intelligence Tasks (HITS) on Amazon Mechanical Turk. The API allows users to develop custom tools that tie into CoSINT's event structure. We omit further discussion and evaluation of the API and tools due to space constraints.

(9) Incorporate Evidence into a Broader Investigation. CoSINT provides Jane with a birds-eye view of the CTF as teams and judges work. She can see all flags and evidence pieces that teams submit and filter them. She can also see the tasks that teams create using add-on tools.

This birds-eye view helps Jane direct the event and steer teams to focus on more important topics. For example, when Team Gamma discovered video footage of a nearby fertilizer storage facility, Jane realized its importance and asked them to identify the exact location where it was taken. She also asked both teams to look for other footage in nearby areas through Twitter's geotag search feature.

After two hours, teams have collected 200 unique pieces of evidence. Jane has already looked through half of them, and created threads on Twitter debunking some social media posts that had significant traction. She is now synthesizing this evidence to write a long-form article.

Implementation Details. We built CoSINT using the Python/Django web framework, a PostgreSQL database, and hosted it on Heroku.

5 FINDINGS

Having described our Research through Design process and the CoSINT system, we now discuss our findings. We focus on: (1) students' evolving attitudes towards OSINT CoCTFs, followed by (2) how students reported collaborating with each other during the

events, (3) their perceptions of the point system and an analysis of their actual performance based on our system log analysis, and (4) students' perceptions of judging during the events.

5.1 Evolving Attitudes Towards OSINT CoCTFs

Many of the computer science students said CoSINT's format was familiar to them because they had prior experience with CTFs. DD-12 said that the format of CoSINT — with teams competing against each other to capture flags and score the points — was similar to other cybersecurity CTFs he had participated in before. However, CS students pointed out four key differences they noticed. First, CoSINT had a real-world orientation (practical investigation vs. theoretical); in SL-3's words, "instead of just reading about [investigations], we were able to perform it ourselves." Second, DD-12 and KG-8 said that flags were not predetermined, so there was no limit on how many points a team could score. Third, there was a different area of focus (misinformation on social media vs. cybersecurity vulnerabilities). Fourth, the time duration was shorter (60 to 90 minutes vs. several hours or days).

In contrast with the CS students, none of the political science students was familiar with CTFs. DD-11 said that participating in gamified and fast-paced investigations was "very much a culture shock for us" because "quite literally, everything is different." Still, DD-11 felt that the format of the investigation was advantageous for political science students because of its novel, hands-on aspect: "[I]t isn't just writing a paper on this topic that we've researched for a few weeks." Instead, she said the CTF taught efficiency and teamwork with a focus on addressing a real-world problem. Despite the novelty, OT-1 said that the instructions we provided were clear and that participating in the CTF was "pretty easy once you get the hang of it."

5.1.1 Defining success. All students said that one form of success was scoring the most points and winning the CTF. One student, SL-2, said she was "very competitive in pretty much anything" and would feel successful if she could "find the way to most easily and effectively win within [the] bounds" of a competition's rules.

Students' definition of success evolved over the semester, to the point where many said that success was more than just winning the CTF. Other markers of success included finding actionable misinformation, achieving a "flow state" [31] with their teammates, having an enjoyable experience, and learning new skills to grow as an investigator.

Not all students were equally motivated by competing to win the CTF. Although DD-10, DD-12, and OT-1 acknowledged that success meant being on top of the leaderboard, they found the most enjoyment when they worked with their team members to find "a story or some sort of a coordinated campaign [which] much more successful than just collecting a bunch of unrelated flags" (DD-12).

5.2 Collaboration Styles During the CoCTFs

We asked students how they worked within their teams and found two types of workflows that they employed: (1) assembly line and (2) free-for-all.

5.2.1 Assembly line workflow. In the assembly line workflow, each member focused on certain phases of the investigation, with two or

more team members working on the same piece of evidence. Two teams (DD, KGB) developed an informal leadership structure with their assembly line. Team KP — which had largely used a free-for-all workflow (see below) — also set up an assembly line workflow for the final CTF.

DD-12 said he disliked how each team member worked independently, instead of deeply working together to complete every flag. Unlike DD-12, KP-14 enjoyed the efficiency of the assembly line:

We just were very fluid, moving very quickly, [...] and we kind of spread all that work out. And, you know, I would find evidence, [KP-16] would archive it ... I felt really good when the flags that I submitted would actually get approved.

Informal team leaders emerged over time for Teams DD and KGB. DD-12 pointed out how he and DD-10 became the *de facto* team leaders and that it “wasn’t done intentionally that way.” DD-10 and 12’s team member, DD-13, wished for a way to explicitly assign work to team members to more easily manage the assembly line workflow.

5.2.2 Free-for-all workflow. Different from the assembly line workflow, four teams (SS, SL, OT, KP) employed workflows that were largely “free-for-alls” where decisions were made in an ad hoc manner. These team members largely worked without coordination on a given evidence piece, submitting flags for each of the four phases. For example, SL-3 said, “it was mostly a free-for-all ... We’d be talking about our flags and what we found. But we wouldn’t really collaborate on or delegate specific tasks.”

Another student, SS-5, believed that the free-for-all workflow was better than an assembly line workflow. He said that his team initially followed an assembly line workflow, but quickly decided to switch to working largely independently. This was because it was difficult to communicate intention and context in an assembly line workflow: “The first person’s discovery flag doesn’t communicate well to the other person trying to do the verification. ‘Hey, why do you actually think this is misinformation?’ or ‘Why do you think this needs to be verified in the first place?’ So you can’t really just split it up purely into those stages” (SS-5).

Apart from Team SL, three of the four teams did not have a team leader. OT-1 said his team would take on tasks at the start of each CTF without a leader assigning them. For Team SL, SL-3 said that SL-2 became the *de facto* team leader because of her strong performance over the first few CTFs.

5.3 Point System Was Effective But Revealed New Tensions

We found that, in line with prior work [12, 132], CoSINT’s point system initially promoted a competitive environment. About two thirds of students said that they enjoyed the competitive environment, while one third said that they did not. As the semester progressed, we modified the point values in line with our goal to make CoSINT more collaborative, while also taking into account students’ feedback on the relative balance of points assigned to different flag categories. However, given the fast-paced, largely competitive, and gamified structure of CoSINT, we found two key tensions over the course of the semester. The first is a tension between competition

and collaboration, and the second is a tension between quantity and quality.

5.3.1 Performance. In Fig. 2, we see that for Event 1, teams submitted 227 flags across 148 evidence pieces (mean = 1.54 flags per evidence). By Event 5, teams submitted 597 flags across 228 evidence pieces (2.62 flags per evidence) — a 70.1% increase in flags per evidence. Despite a 163% increase in the number of flags, the flag approval rate was similar: 78.5% at Event 1 versus 80.4% for Event 5. For Events 2–4, we see that approval ratings were slightly higher at approximately 90%. This suggests that teams became more efficient at submitting more flags without a corresponding decrease in the approval rating.

5.3.2 Receptiveness to the Point System. We found that students’ receptiveness to the points system was affected by a combination of motivational factors — both intrinsic (sense of achievement, competence, and learning) and extrinsic (monetary incentives, grade incentives).

Most students, such as KG-9 and KP-15, were extrinsically motivated by the points system to participate and develop their skills. For them, the point-based leaderboard provided direct feedback indicating whether their strategies were effective. For example, KG-9 said, “I like being able to look at the scoreboard and be like, ‘Hey, we did pretty good today.’ Or sometimes we have bad days, too. And then you learn from the bad days, like, ‘Oh, maybe I should have found more misinformation.’”

Some students did not enjoy competition-based games in general. One graduate student, DD-13, said that it might be an “an age thing” where they were no longer motivated by competing for its own sake. Others were driven by more intrinsic motivations, such as the thrill of the hunt in conducting investigations. For example, KP-15 said,

I don’t think points necessarily correlate to how good of an investigator you are. I think it also has a lot to do with your strategy and which things you focus on. I really enjoyed things that felt engaging to me, like maybe an original verification, archival, or a really good discovery.

5.3.3 Changing Incentives Can Affect Desired Outcomes. Through the point system, we sought to value work that was of greater strategic importance (more recent, more reach, etc.), of higher quality, or required more effort to do. We found that we could encourage students to focus more (or less) on certain aspects of the investigation by changing the point values for different flags and flag categories. For example, KP-15 focused more on verifications once they noticed that verification flags were worth more points than discovery flags:

During the first few CTFs, I would mainly focus on discovery. And I don’t think I’d get a whole lot of points from discovery ... When I did one verification, it got me as many points as it took for five discoveries. Once I noted that, I was like, ‘What am I doing? I should focus on verification,’ because for the same amount of time I can get way more points.

Acting on this realization, KP-15 and his team placed second in Event 5, where 61.8% of their points came from verification flags. In previous events, the percent of Team KP’s points that came from

Type	Event 1	Event 2	Event 3	Event 4	Event 5
Total no. of flags	227	158	257	238	597
No. of approved flags	179	144	229	209	480
No. of rejected flags	48	14	28	29	117
Flag approval rate	78.9%	91.1%	89.1%	87.8%	80.4%
No. of verification flags	29	22	53	40	93
No. of verification flags identifying misinformation	N/A	7	37	27	83
Pct. of verification flags identifying misinformation	N/A	31.82%	69.81%	67.50%	89.25%
Total no. of evidence	148	97	112	114	228
Total no. of flags per evidence	1.54	1.63	2.3	2.09	2.62

Table 2: The number of evidence pieces and flags created per event. From Event 2, we introduced a way to track whether a verification flag refuted the original claim (i.e., identified misinformation).

Event / Team	Event 1		Event 2		Event 3			Event 4			Event 5				Avg. Rk
	Rk	Pts	Rk	Pts	Rk	Pct. Collab.	Pts	Rk	Pct. Collab.	Pts	Rk	Pct. Tasks	Pct. Collab.	Pts	
OT	5	1724	4	2753	4	0%	7637	1	0%	8716	4	0.20%	7.16%	12569	4.25
SL	2	3708	1	4826	1	4.44%	9003	3	0%	6401	1	4.73%	8.90%	17977	2.5
SS	4	1782	9	1097	8	4.41%	4531	10	0%	2536	10	2.73%	3.64%	8241	8.2
KG	6	1551	7	1681	10	0%	3991	8	0%	3216	8	1.50%	4.99%	10028	7.8
DD	1	3942	5	2635	11	0%	915	11	0%	2393	11	1.33%	1.33%	7552	7.8
KP	9	825	6	2597	6	3.47%	5760	6	10.24%	3908	2	0.16%	12.25%	15507	5.8
MH	8	1051	8	1421	5	0%	5979	5	1.83%	5452	3	1.01%	0%	14795	5.8
TL	11	366	3	3412	3	0%	8166	4	0%	5485	9	1.87%	2.14%	9337	6
BF	10	740	2	4215	7	9.14%	5468	9	0%	2685	7	6.92%	2.87%	10472	7
KF	3	2889	11	282	2	9.46%	8445	2	0%	8395	5	6.92%	1.62%	12282	4.6
JE	7	1492	10	470	9	0%	4250	7	2.74%	3651	6	0.85%	0%	11753	7.8

Table 3: The rank and percent of points that each team scored for each of the five events. Percent collaboration refers to the percent of total points that a team scored through collaboration, while percent tasks refer to the percent of total points that a team scored by completing tasks.

verification ranged from 0% in Event 1 to 54.9% in Event 4. For Event 5, Teams KP and OT (who placed fourth) were the only two teams who received 60% or more of their points from verification flags. On the other hand, for all other teams in Event 5, the average percent of points received from verification flags was 33% (min. = 0% and max. = 47.3%).

Interestingly, the team that placed first in Event 5, Team SL, submitted only one verification flag (which was rejected). Instead, it appears that Team SL opted to obtain the majority of their points from discovery and archival flags (40.8% and 39.5%, respectively). Team MH, who placed third, chose a more evenly distributed strategy, obtaining 31.5% of points from verification, 20% from discovery, 16% from archival, and 8.8% from reporting flags.

Initially, DD-10 noted that “no one does reporting flags either because that takes way more work to really put together a report.” For Event 1, two reporting flags were submitted (0.88% of all flags); and none were submitted for Event 2. After we increased the point values for submitting reporting flags for the third CTF, we found that students positively responded to this change. For Events 3, 4, and 5, the ratio of reporting flags to total flags increased to 5.45%,

3.78%, and 11%, respectively. As a percentage of total points scored, this was 3.9%, 2%, and 5% respectively.

Team KP and SL’s strategies seemed to be prioritizing actions that would maximize the number of points scored. Along these lines, KP-15 said he largely focused on completing his own team’s flags. However, once we added the collaboration incentive, “any flag that would pop up once I refreshed, I would just go for it, because we’re going to get those extra points by doing another team’s flags.” In this way, Team KP consistently increased the number of points they obtained through collaboration, from 3.5% and 10.3% in Events 3 and 4, respectively, to 12.3% in Event 5 (see Table 3).

In Table 3, for Event 5, we see a slight correlation with respect to how well teams ranked and whether they leveraged the collaboration and task features. For example, Team SL placed first, receiving 13.6% of their points through collaboration and tasks, and Team KP placed second, receiving 12.4% of points in a similar manner. However, Team MH did not obtain any points through collaboration and only 1% of points through tasks, but still placed third.

Perceptions of fairness. Multiple students, including OT-1, DD-10, DD-11, and KG-9, said that the balance of points improved over the course of the semester as we incorporated their feedback into the

rubric. For instance, DD-10 said, “Before, the more laborious tasks weren’t rewarded nearly as much as they should have been. I think now they are [rewarded] more.” OT-1 added, “I think originally archiving was 100 points ... To me that was way too much. And I think you guys lowered it to 50 or something now. So that makes sense, because archiving is easy, right?”

Still, we found opposing perspectives around how many points should be assigned to certain types of flags. KG-8 said that verification and reporting flags should be worth more points because they were more crucial to the investigation. In contrast, while OT-1 said that the balance of points for discovery, archival and reporting flags were fair, he believed that verification flags were worth too many points. This may be because OT-1, who has worked in the US Intelligence Community, said it’s “really hard to say with 100% probability [that something has been debunked].”

5.3.4 Balancing Competition and Collaboration. As the semester progressed, we modified the point values in line with our goal to make CoSINT more collaborative, while also taking into account students’ feedback on the relative balance of points assigned to different flag categories. Still, we found a tension between competition and collaboration. Some students felt that collaboration was not incentivized enough for it to be worth the effort, or that it was unclear how they could collaborate with other teams. We conclude with ways to better incentivize collaboration.

Competitions promote efficiency and intra-team collaboration. We found that the competitive environment encouraged students to work more efficiently, both on their own and with others. While working individually, KP-14 said that CoSINT “definitely enforced my thought process of, ‘How do I compete in a CTF? Where do I get my points from?’” SS-4 and SS-5 felt similarly, saying, “It helped us to understand the process overall, but it encouraged us to be more efficient in how we look at the process.”

Inter-team collaboration can be useful in competitions but is difficult to structure. Apart from encouraging competition and intra-team collaboration (within teams), we observed that CoSINT also promoted inter-team collaboration (between teams). Some students found the ability to collaborate with other teams useful, but others were unsure how to do so effectively.

We found that students appreciated the ability to gain points through collaboration for two reasons. First, this incentivized people to work together, and second, working on other teams’ flags gave them access to a wider variety of topics to investigate and methods to use.

For example, when KG-8 and KG-9 learned that Team KF was performing well because they were using a Twitter scraping tool, they decided to look into it and have their team use it as well, because otherwise they thought that “we’re totally going to get destroyed.” KG-9 liked this balance between competition and collaboration in CoSINT where it is “half collaboration and half competition.”

OT-1 also said that there were multiple instances where he was contributing to another team’s evidence piece to gain points. By Event 5, Team OT obtained 7.36% of their total points from collaboration and tasks. OT-1 also found leads from other teams that were beneficial for his own work. In one example, OT-1 said “I was looking at another team’s discovery post, because I was going

to archive it. And that account had over 100,000 followers, and I was like, ‘Well, I haven’t heard of this account before.’ ... So it was helpful to find other accounts through the CTF.”

Many students saw the ability to build on other teams’ flags as a turning point in the semester. KP-14 recalled, “As people got better at the CTF, they became more competitive and more collaborative. But as far as adding the feature of actually being able to verify other people’s stuff, that definitely had a significant boost on collaboration.” KP-14’s team saw the collaboration feature as an “opportunity to be a shark” and earned many of their points in the final CTF this way. Specifically, KP-14’s team searched and filtered evidence tab for specific teams’ evidence pieces. Then, they inspected the status to determine if an archival, verification, or reporting flag was present. If there was a missing flag, someone from KP-14’s team would attempt to create it themselves. For Event 4, Team KP was the most collaborative team, obtaining 373% more points through collaboration than the second-most collaborative team. For Event 5, Team KP placed second and was the second-most collaborative team at 12.41% of total points, versus 13.63% for Team SL who placed first (see Table 3).

Some teams did not collaborate, and students described several reasons why not. First, SL-2 and DD-11 both said that collaborating with teams was not incentivized enough in the most recent version of CoSINT because it was not “worth as much points as the time that went into doing them properly, [versus] making flags yourself” (SL-2). Even though SL-2 said collaboration was not sufficiently incentivized, Team SL still obtained 13.63% of their points through collaboration and tasks.

Second, some teams found collaboration was confusing, time-consuming, or required cognitively demanding tasks like context-switching and sensemaking of the other team’s work products. SL-2 was “not really sure how to collaborate through the CoSINT platform.” DD-11 decided not to collaborate because “there wasn’t really a lot of time to understand what the other teams were working on and what their objectives are, both from the flag and evidence perspective, and then from the tool perspective.” SS-5 recalled, “I haven’t helped another team’s flag yet. Even with the points I’m just not inclined to, because at the end of the day, I have to read through theirs, understand what it is. That’s almost like stopping in my tracks what I’m doing already, and then trying to understand what they’re doing.”

Third, some teams worried about how collaboration might negatively impact other teams. SS-4 and KP-15 voiced concerns that unexpected collaboration could be distracting:

There’s a fine line between competition and collaboration with some things, because if I verify some other group’s piece of evidence, we’re technically collaborating, but perhaps their strategy is to have them focus on their own pieces of evidence. So maybe I’m disrupting [their] strategy as well. (KP-15)

More broadly, some teams felt that the novelty of collaboration required changing norms or reframing expectations, especially for those with prior CTF experience. Although Teams OT, KG, and KP took advantage of collaboration features during the CTFs, even these teams recognized it as unusual. OT-1 said that collaboration

was not common in typical OSINT investigations he had participated in. SS-4 thought the idea of collaborating across competing teams was a “really cool idea and it definitely works [but] if all the teams are on board with doing that, it’ll go a lot smoother.” OT-1 suggested a change in mindset for all participants at the CTF might help improve collaboration. Participants should view the CTF as teams collectively working towards a common goal “instead of separate teams working on separate things, trying to win.”

5.3.5 Balancing Breadth and Depth. We described in Section 3.3 how one of the goals for CoSINT was to rapidly identify and debunk misinformation. This required casting a wide net, both in terms of covering a wide variety of topics but also collecting a large quantity of content. However, we found a tension between our design goals that emphasized breadth versus students’ desire to conduct in-depth investigations.

Defining a good investigation: breadth vs. depth. In Section 5.3.4, we described how the competitive environment promoted efficiency and breadth. However, OT-1, who had prior experience with OSINT investigations, believed that a good player needed to balance discovering a large quantity of content while making sure that it is also of high quality through careful research. He explained that “a lot of times it’s easy to discover poor quality tweets made by bots, you know, it’s obvious, but then the real exploitable information is a little bit harder to find.”

Five other students also said that they preferred an environment that incentivized conducting in-depth investigations. For example, KG-8 said her team ended up finding “a lot of small pieces of misinformation [because] a lot of the bigger fish had sort of been fried already.” DD-10 also pointed out how his team had not spent much time looking into any single piece of evidence, but rather “trying to just cast a huge net.” He went on to describe what he perceived as the tension between breadth and depth within a competition: “Stuff like that, that takes a lot of time, and quantitatively it’s not very much actual result at all, is actually the most [intrinsically] rewarding. You really have to be clever about this one image instead of finding all of them.”

Rewarding and assessing depth. Through our classroom deployment, we found that students were receptive to changes in the point structure. In turn, rewarding in-depth work with more points may satisfy some students’ desires to investigate in depth. For example, KP-15 said that they valued original verifications where they “extrapolated on some knowledge from a few different sources ... as opposed to just using a fact check article,” but also added that there was a big “point boost” for original verifications, which he described as a “win-win” scenario.

From Event 2, we began tracking if a verification flag identified an instance of *misinformation* — that is, whether it refuted the original claim made in the discovery flag. For example, one member of Team JE found a video posted online that claimed to show an instance of voter fraud. However, this team member was able to debunk the video by finding an alternative source that had investigated the same video. In this case, that verification flag identified an instance of misinformation. From Event 2 through Event 5, the percentage of approved verification flags with instances of misinformation rose from 31.8% to 89.25% (see Fig. 2), indicating

that teams increasingly submitted and investigated content fitting our definition of misinformation. (As a caveat, Event 5 focused on topics such as 9/11 and chemtrail conspiracy theories, which are more likely to contain misinformation.)

5.4 Judging Improved Quality But May Be Difficult to Scale Up

5.4.1 Self-assessment rubric and judging improved quality of flags. We found two aspects of the judging process that students said helped improve the quality of flags that they submitted. First, students perceived the self-assessment rubric as valuable. For example, KG-7 noted that the self-assessment rubric helped them better understand the requirements for a high-quality flag to “make sure I can get the most points and I can justify the points.”

Second, students said that the judges’ feedback encouraged them to submit higher quality, more detailed flags that were not only more likely to be approved, but would be worth more points. Many of SS-5’s discovery flags were rejected early on because he did not sufficiently describe why what he had found was potential misinformation. However after he resubmitted the same flags with more details, they were approved.

5.4.2 Judging misinformation may be subjective. Despite our use of a rubric to make judging more fair and objective, three students — DD-13, KP-14, and OT-1 — pointed out that judging whether something is misinformation may be a subjective task. Further, KP-14 worried that the evidentiary standard required for verifying or refuting a claim differed between judges. He said that he would sometimes reject flags because he did not think that a student had submitted enough evidence, but was not sure if another judge would have rejected the flag for the same reason, perhaps because they were “a little more timid to reject it.”

To overcome these challenges, KP-14 and KP-15 suggested rotating judges between teams. Alternatively, DD-13 suggested a tiered judging system where one judge would go over another judges’ evaluation, acknowledging the potential drawbacks of increased reviewing workload and confirmation bias among trusting judges.

5.4.3 Rubric enables judging to scale up. The research team, including three to six research assistants, acted as judges for all of the CTFs. During the first two CTFs, we maintained our rate of evaluation in line with students’ rate of submission. Students soon became quicker and more adept at submitting flags, leading to two occasions where we could not evaluate all flags before class ended. For Events 3 and 4, to decrease judging bottlenecks and provide students with practice being a judge, we asked students from each team to sign up as a judge. With the students’ help, we found that judges were better able to keep up with the rate of submission. In Event 3, there were 25 judges who evaluated 292 flags, and in Event 4, there were 24 judges who evaluated 238 flags. In both events, judges took on average 10.6 minutes to judge a flag after it was submitted. For Event 5, we wanted students to fully participate in flag submission, so we recruited additional research assistants as judges. Here, 11 judges took 20.6 minutes on average to judge 597 flags. Thus, the judges were able to review four times as many flags, but the judging time per flag doubled.

Two potential reasons for this increased efficiency could be that students became more adept at submitting higher quality flags (that took less time to evaluate), but also that judges themselves became more adept at evaluating flags. We found that judges also perceived that the self-assessment rubric enabled them to evaluate flags more quickly. In KG-9's words, the rubric "streamlines the process — 'Okay, well, I need to follow this link, I need to check all these things' — it makes it easier to approve it." SS-5 also felt that the rubric encouraged students to more accurately rate themselves, such that as a judge, he rarely adjusted the rating.

6 DISCUSSION

In this work, we engaged in a four-month-long Research through Design (RtD) process to develop the CoSINT platform. We find that a RtD process helped us to improve and validate the design of the platform, moving closer towards our ideal preferred state described in our two design goals.

First, we find that CoSINT enabled a rapid response to misinformation on social media by merging the complementary benefits of competition and collaboration. CoSINT's point system not only motivated teams to compete *against* each other, but to also collaborate *with* each other. Second, our findings show that CoSINT structured students' work, allowing them to perform complex investigative tasks ranging from discovery and archiving to verification and reporting. CoSINT was flexible enough that students could investigate a range of topics, from COVID-19 and election misinformation to human rights violations and stock market rumors.

Recall that we instantiated our two design goals using four of Dix's heuristics for software appropriation [37]. However, Dix provides another important heuristic that we discuss next: *learn from appropriation*. By observing how a system has been used and appropriated, we can redesign the system to better support users. Our mixed-methods evaluation allowed us to assess CoSINT against our two design goals, which we revisit below.

6.1 Goal 1: Enable a Rapid Response to Misinformation on Social Media

6.1.1 CoSINT reduced inefficiencies compared to current CTF competitions. We showed that a crowd of 46 students could be motivated to quickly identify and debunk hundreds of pieces of potential misinformation in sessions as short as 60 minutes by creating a competitive environment with a points-based incentive structure. Building on prior work showing the benefits of competition in crowdsourcing [12, 131], CoSINT demonstrated that traditionally theoretical CTFs can be adapted for real-world misinformation investigations. Second, we mitigated some key limitations of competitions, such as information silos, by allowing competing teams to view and build upon each others' evidence and flags. In fact, some teams scored up to 12.25% of their points by contributing to other teams' evidence pieces (average = 4.07%), whereas such collaboration would not be possible in traditional CTFs.

Despite the noticeable increase in collaboration over the last three events for our CoCTF platform, our findings suggest that collaboration can be further encouraged. For example, some students indicated that they were hesitant to contribute to other teams' flags

without explicit calls for help or social norms encouraging collaboration within the CTF. As Lessig posits in his New Chicago School theory [79], there are four ways to regulate human behavior: laws, norms, markets, and architecture. While CoSINT leverages markets (extra points for collaboration) and architecture (information sharing, contributing flags to other teams' evidence, and tasks), future CoCTFs should explore how to frame policies and develop social norms to encourage collaboration. For example, sets of two competing teams could be required to be physically or virtually co-located to minimize redundancy and maximize information sharing [137]. In terms of social norms, the expert could emphasize the shared goal that teams are working towards, and encourage members of different teams to build rapport with each other [83, 104]. The architecture of CoCTF systems could also facilitate social translucence [43] where teams can press a help wanted button to indicate that they are open to collaboration.

6.1.2 CoSINT reduced inefficiencies compared to traditional crowdsourcing approaches in three ways. In many traditional crowdsourcing systems with monetary compensation, designers implement "attention checks" to make sure crowd workers are making an honest effort to complete the work. They also aggregate multiple, repetitive crowd inputs for the same microtasks to mitigate the effects of low-quality work or biases [93]. Instead, CoSINT enabled high-quality work through a combination of a trusted and trained crowd [93], a point-based incentive system [141], self-assessment rubrics [39], and real-time feedback from judges [72]. Because we knew the students and developed a working relationship with them, we could trust them to submit higher-quality work compared to an anonymous crowd. We could also delineate and communicate low- and high-quality work to students through the point system and rubric. Finally, students felt that the self-assessment rubrics and judging mechanism improved their work in the short- and long-term.

Still, more structure within teams could lead to greater efficiency gains. For instance, some teams organically devised assembly line workflows and team leaders emerged over time; these teams frequently placed high on the leaderboard. In contrast, teams that employed free-for-all workflows with minimal collaboration among team members and no explicit team leader did not perform as well.

Our findings suggest that both types of teams may benefit from more explicit structure and roles, such as delineating the responsibilities for each team leader and assigning roles to each team member [57, 87, 116, 137]. While Retelny et al. [116] suggest that rigid workflows restrict adaptability, we find that too much freedom can hamper performance. Future work should explore providing flexible structures that teams can choose to use and modify based on their working styles. For example, the leader could mitigate unwanted redundancy by assigning team members to work on a specific topic or social media platform. To prevent judges from being overwhelmed by work, the leader could also conduct a preliminary evaluation of their flags before forwarding it to the judge. To further increase efficiency, individuals could be assigned or encouraged to focus on tasks that they preferred or excelled at, such as content discovery versus verification.

To increase the impact of CoCTFs, events can increase the overall number of participants and also involve a greater number of professional investigators to lead teams. However, as we learned in

our four-month-long deployment, judges occasionally struggled to keep up with the rate of submissions as teams became more adept at creating flags. To address this bottleneck, designers could explore developing automated judging systems trained on past judges' evaluations [68] and provide participants with automated tailored feedback from large language models (LLMs) [20].

6.2 Goal 2: Give the Crowd (More) Agency

In traditional crowdsourcing systems, complex tasks are divided into microtasks that crowd workers complete independently, with little to no interaction with each other or agency in how to complete these tasks [14, 70]. However, CoSINT builds on a growing body of literature that shows that crowds can perform more complex tasks, provided that they are sufficiently motivated and given adequate scaffolding, training, and agency [39, 57, 116].

6.2.1 Providing more agency can lead to a virtuous cycle. We also found that CoSINT helped students learn to more critically examine information online and develop a mental model for conducting investigations. This proved to be a virtuous cycle: between the first and fifth events, students submitted 65% more evidence pieces and 163% more flags, while maintaining flag approval ratings. In addition, students said that they enjoyed using CoSINT, possibly motivating them to continue participating in the events.

Students desired even greater agency to investigate topics in greater depth. While we designed CoSINT to provide a rapid response to misinformation (60–90 minutes), future work should study how to provide the crowd with greater agency and design longer-duration CoCTF events. Participants in our study also noted that greater quantity did not always imply greater quality, and it may be beneficial for CoCTF organizers to empirically analyze the trade-offs between quantity and quality. Teams could be limited to a certain number of submissions per hour, or high-quality flags could be emphasized — through point incentives and community norms [78] — over low-quality flags.

6.2.2 Dynamically modify competition and collaboration affordances. Though CoSINT motivated most students to participate, some were less motivated by competition, preferring collaboration instead. To better engage crowd workers, future CoCTF systems should consider alternative team structures and more flexible incentive and feedback mechanisms.

Prior to starting the event, the system could survey the crowd to signal to the expert what motivates them [111], and allow the expert to modify the system accordingly. One option would be to divide teams into two groups that work in the same environment but have different work arrangements. One group — those motivated by a sense of urgency and competition — could conduct rapid data collection and analysis with a group-wide competition or even self-competition [97]. A second group — motivated by conducting in-depth analyses and collaboration — could collaboratively investigate the first groups' work in greater detail and over a longer period of time (days versus hours).

A second option would be to emphasize different types of feedback [133]. For crowd workers who are more motivated by qualitative assessments of their work, the system could prompt judges to provide detailed written feedback, and emphasize this in the crowd

worker's interface over the point values that the judge awarded them.

6.3 Extending CoCTFs to Other Domains

Dix also recommends that designers *allow for [re]interpretation* of the system. In other words, this intentional “absence of meaning” allows users to appropriate the system for other purposes [51]. As mentioned in Section 2, CTFs are best suited for settings focused on collecting new information or uncovering new problems [66]. Indeed, CoSINT could be easily adapted for domains other than online investigations of misinformation, such as to coordinate physical search-and-rescue efforts for missing persons or animals [135, 145] or to assess damage after natural and man-made disasters [15]. In these situations, flag types could remain the same (discovery, archival, verification, and reporting), but the evaluation criteria and point values for each would differ. For instance, the evaluation criteria for a discovery flag are currently originality, influence, and recency. These could be modified to focus on factors more relevant during crises, such as reliability and recency of information and level of danger posed. Given the potential consequences of errors when lives may be in immediate danger, organizers of such CoCTFs may also need to de-emphasize certain “fun” gamified and competitive elements [129], and instead encourage greater communication and collaboration.

To incorporate individuals with relevant skills and from diverse backgrounds, CoCTF organizers should consider working closely with existing communities of practice [144] to understand their intrinsic and extrinsic motivations and enable novices to join these communities through legitimate peripheral participation [75]. For example, novices could be required to “shadow” judges or join an experienced team and make micro-contributions [84].

6.4 Limitations and Broader Impacts

Giving crowd workers more agency is often viewed in a positive light [63]. In addition, CoSINT itself empowers regular citizens to work together to rapidly uncover misinformation, holding governments and corporations accountable for their words and actions. We must also grapple with the potential negative impacts that sociotechnical systems like CoSINT can have on individuals and societies [58]. For example, authoritarian governments could use CoSINT to crowdsource investigations into dissidents, or rogue crowds could use it to investigate members of marginalized communities. However, by incorporating expert supervision and evaluation, as well as training on professional and ethical investigative standards — such as OSINT's “no touch” or passive reconnaissance ethos — CoSINT reduces the likelihood of these potential harms.

In addition, addressing misinformation by delineating factual information from false and misleading information is not a panacea. Corrections may lead to a “backfire” effect and increase partisanship [115]. Fact-checking may also not be relevant during crises and mass-convergence events — such as natural disasters, protests, or political events — where it may not be possible to immediately determine the veracity of information. In these situations, CoSINT can still be leveraged to improve investigators' contextual understanding of these events. On the whole, we believe that by enabling democratic participation in understanding our online information

ecosystems, CoSINT can be used to do significantly more good than harm.

7 CONCLUSION

We engaged in a four month-long Research through Design process to develop and evaluate CoSINT, a platform for collaborative capture the flag competitions (CoCTFs) that enabled a trained crowd to investigate misinformation on social media. CoSINT showed that traditionally theoretical CTFs can be adapted for real-world misinformation investigations; and that novice crowd workers can be provided with greater agency in OSINT work when coupled with training, scaffolding, and expert guidance. Further, by incorporating beneficial elements of collaboration into a CTF, CoSINT ameliorates two limitations of purely competitive CTFs: information silos and duplication of effort. In turn, this novel CoCTF concept allowed a trained crowd of 46 students to identify and debunk hundreds of pieces of misinformation in less than ninety minutes, while collaborating up to 12% of the time. By merging competition and collaboration, CoCTFs can be a powerful site of collective action that is both effective and enjoyable.

ACKNOWLEDGMENTS

This paper about crowdsourcing was itself an effort in crowdsourcing. We would like to thank members of the Crowd Intelligence Lab for their contributions to this work — especially Emily A., Ryan B., Charles C., Katie F., Alex H., Yunis H., Mariela J., Rissa M., Brandon N., Sophia P., and Raymar R. This work would also not have been possible without the involvement of students in the OSINT Lab course. We appreciate the insightful and detailed comments provided by Kate Starbird, Tanushree Mitra, Chris North, Eugenia Rho, and the anonymous reviewers. This work was supported by NSF award IIS-1651969 and the Virginia Commonwealth Cyber Initiative. Sukrit Venkatagiri was additionally supported by the University of Washington Center for an Informed Public, Craig Newmark Philanthropies, and the John S. and James L. Knight Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the above supporting organizations or the National Science Foundation.

REFERENCES

- [1] Alwan Abdullah, Shams A. Laghari, Ashish Jaisan, and Shankar Karuppayah. 2021. OSINT Explorer: A Tool Recommender Framework for OSINT Sources. In *Advances in Cyber Security (Communications in Computer and Information Science)*, Nibras Abdullah, Selvakumar Manickam, and Mohammed Anbar (Eds.). Springer, Singapore, 389–400. https://doi.org/10.1007/978-981-16-8059-5_24
- [2] Sultan A. Alharthi, Nicolas James LaLone, Hitesh Nidhi Sharma, Igor Dolgov, and Z. O. Touns. 2021. An Activity Theory Analysis of Search and; Rescue Collective Sensemaking and Planning Practices. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 146, 20 pages. <https://doi.org/10.1145/3411764.3445272>
- [3] Carlo Aliprandi, Juan Arraiza Irujo, Montse Cuadros, Sebastian Maier, Felipe Melero, and Matteo Raffaelli. 2014. CAPER: Collaborative Information, Acquisition, Processing, Exploitation and Reporting for the Prevention of Organised Crime. *Communications in Computer and Information Science* (2014), 6.
- [4] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2020. Scaling up fact-checking using the wisdom of crowds. *Preprint at https://doi.org/10.31234/osf.io/9qdz2* (2020).
- [5] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2021. Scaling up fact-checking using the wisdom of crowds. *Science advances* 7, 36 (2021), eabf4393.
- [6] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *CHI Conference on Human Factors in Computing Systems*. 1–19.
- [7] Deb Amos. 2022. Open source intelligence methods are being used to investigate war crimes in Ukraine. *NPR* (June 2022). <https://www.npr.org/2022/06/12/1104460678/open-source-intelligence-methods-are-being-used-to-investigate-war-crimes-in-ukr>
- [8] Ahmer Arif. 2018. Designing to Support Reflection on Values & Practices to Address Online Disinformation. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) (CSCW '18). Association for Computing Machinery, New York, NY, USA, 61–64. <https://doi.org/10.1145/3272973.3272974>
- [9] Ahmer Arif, John J. Robinson, Stephanie A. Stanek, Elodie S. Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A Closer Look at the Self-Correcting Crowd: Examining Corrections in Online Rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 155–168. <https://doi.org/10.1145/2998181.2998294>
- [10] Jeffrey Bardzell, Shaowen Bardzell, Peter Dalsgaard, Shad Gross, and Kim Halskov. 2016. Documenting the Research Through Design Process. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (Brisbane, QLD, Australia) (DIS '16). Association for Computing Machinery, New York, NY, USA, 96–107. <https://doi.org/10.1145/2901790.2901859>
- [11] Charlie Beckett. 2017. Wikitribune: can crowd-sourced journalism solve the crisis of trust in news? *POLIS: journalism and society at the LSE* (2017).
- [12] Yasmine Belghith, Sukrit Venkatagiri, and Kurt Luther. 2022. Compete, Collaborate, Investigate: Exploring the Social Structures of Open Source Intelligence Investigations. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (CHI '22). Association for Computing Machinery, New York, NY, USA, 1–18. <https://doi.org/10.1145/3491102.3517526>
- [13] Arne Berger, Sören Totzauer, Kevin Lefevre, Michael Storz, Albrecht Kurze, and Andreas Bischof. 2017. Wicked, Open, Collaborative: Why Research through Design Matters for HCI Research. *i-com* 16, 2 (2017), 131–142.
- [14] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Solynt: a word processor with a crowd inside. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*. ACM, 313–322.
- [15] Christian Bittner, Boris Michel, and Cate Turk. 2016. Turning the spotlight on the crowd: Examining the participatory ethics and practices of crisis mapping. *ACME: An International Journal for Critical Geographies* 15, 1 (2016), 207–229.
- [16] Kevin J. Boudreau and Karim R. Lakhani. 2015. "Open" disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology. *Research Policy* 44, 1 (Feb. 2015), 4–19. <https://doi.org/10.1016/j.respol.2014.08.001>
- [17] John Bowers. 2012. The Logic of Annotated Portfolios: Communicating the Value of 'Research through Design'. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) (DIS '12). Association for Computing Machinery, New York, NY, USA, 68–77. <https://doi.org/10.1145/2317956.2317968>
- [18] Jonathan Bragg, Mausam, and Daniel S. Weld. 2018. Sprout: Crowd-Powered Task Design for Crowdsourcing. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) (UIST '18). ACM, New York, NY, USA, 165–176. <https://doi.org/10.1145/3242587.3242598>
- [19] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [20] Chen Cao. 2023. Leveraging Large Language Model and Story-Based Gamification in Intelligent Tutoring System to Scaffold Introductory Programming Courses: A Design-Based Research Study. *arXiv preprint arXiv:2302.12834* (2023).
- [21] Benjamin Carlisle, Michael Reininger, Dylan Fox, Daniel Votipka, and Michelle L Mazurek. 2020. On the other side of the table: Hosting capture the flag (ctf) competitions. In *Proceedings of the 6th Workshop on Security Information Workers, ser. WSIW*, Vol. 20.
- [22] John M Carroll, Wendy A Kellogg, and Mary Beth Rosson. 1991. The task-artifact cycle. *Designing interaction: Psychology at the human-computer interface* (1991), 74–102.
- [23] Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies Between Research Papers. 2 (2018), 31:1–31:21. Issue CSCW. <https://doi.org/10.1145/3274300>
- [24] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. 2017. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2334–2346.
- [25] Lennon YC Chang and Andy KH Leung. 2015. An introduction to cyber crowdsourcing (human flesh search) in the Greater China region. In *Cybercrime Risks and Responses*. Springer, 240–252.
- [26] Peter Chapman and David Brumley. 2013. picoCTF: Teaching 10,000 high school students to hack.

- [27] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA) (*CHI '13*). ACM, 1999–2008. <https://doi.org/10.1145/2470654.2466265>
- [28] Kevin Chung and Julian Cohen. 2014. Learning obstacles in the capture the flag model. In *2014 USENIX Summit on Gaming, Games, and Gamification in Security Education* (3GSE 14).
- [29] Amy Cook, Steven Dow, and Jessica Hammer. 2020. Designing Interactive Scaffolds to Encourage Reflection on Peer Feedback. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) (*DIS '20*). Association for Computing Machinery, New York, NY, USA, 1143–1153. <https://doi.org/10.1145/3357236.3395480>
- [30] Joseph Cox. 2018. The Hackers Hunting Down Missing People: Nonprofit TraceLabs ran DEF CON's first crowdsourced event for tracking missing people through public information. *Vice* (2018). https://www.vice.com/en_us/article/qvmm3x/hackers-hunting-missing-people-osint-defcon-tracelabs
- [31] Mihaly Csikszentmihalyi and Mihaly Csikszentmihalyi. 1990. *Flow: The psychology of optimal experience*. Vol. 1990. Harper & Row New York.
- [32] Dharma Dailey and Kate Starbird. 2014. Journalists as Crowdsourcers: Responding to Crisis by Reporting with a Crowd. *Computer Supported Cooperative Work (CSCW)* 23, 4 (01 Dec 2014), 445–481. <https://doi.org/10.1007/s10606-014-9208-z>
- [33] Peter Dalsgaard and Kim Halskov. 2012. Reflective design documentation. In *Proceedings of the Designing Interactive Systems Conference*. 428–437.
- [34] Oliver Dean. 2021. Cybersecurity competition challenges next generation of security experts. <https://news.asu.edu/20210930-solutions-cybersecurity-competition-challenges-next-generation-security-experts>
- [35] Katie Derthick, Patrick Tsao, Travis Kriplean, Alan Borning, Mark Zachry, and David W McDonald. 2011. Collaborative sensemaking during admin permission granting in Wikipedia. In *International Conference on Online Communities and Social Computing*. Springer, 100–109.
- [36] Nicholas Diakopoulos, Daniel Trielli, and Grace Lee. 2021. Towards Understanding and Supporting Journalistic Practices Using Semi-Automated News Discovery Tools. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct. 2021), 406:1–406:30. <https://doi.org/10.1145/3479550>
- [37] Alan Dix. 2007. Designing for appropriation. In *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but not as we know it - Volume 2 (BCS-HCI '07)*. BCS Learning & Development Ltd., Swindon, GBR, 27–30.
- [38] Steven Dow, Elizabeth Gerber, and Audris Wong. 2013. A pilot study of using crowds in the classroom. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 227–236.
- [39] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the Crowd Yields Better Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) (*CSCW '12*). Association for Computing Machinery, New York, NY, USA, 1013–1022. <https://doi.org/10.1145/2145204.2145355>
- [40] Sam Dubberley, Alexa Koenig, and Daragh Murray. 2020. *Digital witness: Using open source information for human rights investigation, documentation, and accountability*. Oxford University Press, USA.
- [41] Robert L Dunne. 1994. Deterring Unauthorized Access to Computers: Controlling Behavior in Cyberspace Through a Contract Law Paradigm. *Jurimetrics J.* 35 (1994), 1.
- [42] Mick Endors and Bill Peace. 2015. A Call to Arms: Open Source Intelligence and Evidence Based Policymaking. <https://www.bellingcat.com/resources/articles/2015/01/20/a-call-to-arms-open-source-intelligence-and-evidence-based-policy-making/>
- [43] Thomas Erickson and Wendy A Kellogg. 2000. Social translucence: an approach to designing systems that support social processes. *ACM transactions on computer-human interaction (TOCHI)* 7, 1 (2000), 59–83.
- [44] Gregory J. Feist. 1991. Synthetic and analytic thought: Similarities and differences among art and science students. *Creativity Research Journal* 4, 2 (Jan. 1991), 145–155. <https://doi.org/10.1080/10400419109534382>
- [45] Gerhard Fischer, Elisa Giaccardi, Hal Eden, Masanori Sugimoto, and Yunwen Ye. 2005. Beyond binary choices: Integrating individual and social creativity. 63, 4 (2005), 482–512. <https://doi.org/10.1016/j.ijhcs.2005.04.014>
- [46] Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed Sensemaking: Improving Sensemaking by Leveraging the Efforts of Previous Users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA) (*CHI '12*). ACM, 247–256. <https://doi.org/10.1145/2207676.2207711>
- [47] Richard Fletcher, Alessio Cornia, Lucas Graves, and Rasmus Kleis Nielsen. 2018. Measuring the reach of “fake news” and online disinformation in Europe. *Australasian Policing* 10, 2 (2018).
- [48] Susan Gasson. 2005. The Dynamics of Sensemaking, Knowledge, and Expertise in Collaborative, Boundary-Spanning Design. 10, 4 (2005). <https://doi.org/10.1111/j.1083-6101.2005.tb00277.x>
- [49] Michael Glassman and Min Ju Kang. 2012. Intelligence in the internet age: The emergence and evolution of Open Source Intelligence (OSINT). *Computers in Human Behavior* 28, 2 (2012), 673–682.
- [50] William Godel, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A Tucker. 2021. Moderating with the mob: Evaluating the efficacy of real-time crowdsourced fact-checking. *Journal of Online Trust and Safety* 1, 1 (2021).
- [51] Joseph A. Gonzales, Casey Fiesler, and Amy Bruckman. 2015. Towards an Appropriate CSCW Tool Ecology: Lessons from the Greatest International Scavenger Hunt the World Has Ever Seen. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) (*CSCW '15*). Association for Computing Machinery, New York, NY, USA, 946–957. <https://doi.org/10.1145/2675133.2675240>
- [52] Nitesh Goyal and Susan R. Fussell. 2016. Effects of Sensemaking Translucence on Distributed Collaborative Analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (New York, NY, USA) (*CSCW '16*). ACM, 288–302. <https://doi.org/10.1145/2818048.2820071>
- [53] Catherine Greve and Eric Gilbert. 2015. Piggyback Prototyping: Using Existing, Large-Scale Social Computing Systems to Prototype New Ones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) (*CHI '15*). Association for Computing Machinery, New York, NY, USA, 4047–4056. <https://doi.org/10.1145/2702123.2702395>
- [54] Hacktoria. 2021. Hacktoria – Story Based OSINT Capture The Flag Challenges. <https://hacktoria.com/>
- [55] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA) (*CHI '16*). ACM, 2258–2270. <http://doi.acm.org/10.1145/2858036.2858364>
- [56] Shalin Hai-Jew. 2017. Real-Time Sentiment Analysis of Microblog Messages with the Maltego “Tweet Analyzer” Machine. In *Social Media Listening and Monitoring for Business Applications*. IGI Global, 316–337.
- [57] Alexa M Harris, Diego Gómez-Zarà, Leslie A DeChurch, and Noshir S Contractor. 2019. Joining together online: the trajectory of CSCW scholarship on group formation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.
- [58] Brent Hecht, L Wilcox, JP Bigham, J Schöning, E Hoque, J Ernst, Y Bisk, L De Russis, L Yarosh, B Anjum, and others. 2018. It's time to do something: Mitigating the negative impacts of computing through a change to the peer review process. *ACM Future of Computing Blog*. Mar. 29, 2018. <https://acmfca.org/2018/03/29/negativeimpacts/>
- [59] Patricia Hswe, Joanne Kaczmarek, Leah Houser, and Janet Eke. 2009. The Web Archives Workbench (WAW) Tool Suite: Taking an Archival Approach to the Preservation of Web Content. *Library Trends* 57, 3 (2009), 442–460. <https://doi.org/10.1353/lib.0.0046> Publisher: Johns Hopkins University Press.
- [60] William Hue. 2019. Supporting Collaborative Use of Self-Tracking Data in the Context of Healthcare and Chronic Conditions. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion* (San Diego, CA, USA) (*DIS '19 Companion*). Association for Computing Machinery, New York, NY, USA, 93–96. <https://doi.org/10.1145/3301019.3324877>
- [61] Katja Hutter, Julia Hautz, Johann Füller, Julia Mueller, and Kurt Matzler. 2011. Communitition: The Tension between Competition and Collaboration in Community-Based Design Contests. *Creativity and Innovation Management* 20, 1 (2011), 3–21. <https://doi.org/10.1111/j.1467-8691.2011.00589.x> <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8691.2011.00589.x>
- [62] Denis Iorga, Octavian Grigorescu, Mihai Predoiu, Cristian Sandescu, Mihai Dascalu, and Razvan Rughinis. 2021. Early Usability Evaluation to Enhance User Interfaces-A Use Case on the Yggdrasil Cybersecurity Mockup. In *RoCHI*. 103–110.
- [63] Lilly C. Irani and M. Six Silberman. 2013. *Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk*. Association for Computing Machinery, New York, NY, USA, 611–620. <https://doi.org/10.1145/2470654.2470742>
- [64] Prerna Juneja and Tanushree Mitra. 2022. Human and Technological Infrastructures of Fact-Checking. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 418 (nov 2022), 36 pages. <https://doi.org/10.1145/3555143>
- [65] Ruogu Kang, Aimee Kane, and Sara Kiesler. 2014. Teammate Inaccuracy Blindness: When Information Sharing Tools Hinder Collaborative Analysis. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (New York, NY, USA) (*CSCW '14*). ACM, 797–806. <https://doi.org/10.1145/2531602.2531681> event-place: Baltimore, Maryland, USA.
- [66] Stylianos Karagiannis, Elpidoforos Maragkos, and Emmanouil Magkos. 2020. *An Analysis and Evaluation of Open Source Capture the Flag Platforms as Cybersecurity e-Learning Tools*. https://doi.org/10.1007/978-3-030-59291-2_5 Pages: 77.
- [67] N. Kerle and R. R. Hoffman. 2013. Collaborative damage mapping for emergency response: the role of Cognitive Systems Engineering. 13, 1 (2013), 97–113. <https://doi.org/10.5194/nhess-13-97-2013>
- [68] Sung-Kyung Kim, Eun-Tae Jang, and Ki-Woong Park. 2020. Toward a fine-grained evaluation of the Pwnable CTF. In *Information Security Applications: 21st International Conference, WISA 2020, Jeju Island, South Korea, August 26–28, 2020, Revised Selected Papers*. Springer, 179–190.

- [69] Aniket Kittur, Andrew M. Peters, Abdigani Diriye, and Michael Bove. 2014. Standing on the Schemas of Giants: Socially Augmented Information Foraging. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (New York, NY, USA) (CSCW '14). ACM, 999–1010. <https://doi.org/10.1145/2531602.2531644>
- [70] Aniket Kittur, Boris Smus, Sushel Khamkar, and Robert E. Kraut. 2011. CrowdForge: crowdsourcing complex work. In *Proceedings of the 24th annual ACM symposium on User interface software and technology* (New York, NY, USA) (UIST '11). ACM, 43–52. <https://doi.org/10.1145/2047196.2047202>
- [71] Ilpo Koskinen, Finland Thomas Binder, and Johan Redström. 2008. Lab, field, gallery, and beyond. *Artifact: Journal of Design Practice* 2, 1 (2008), 46–57.
- [72] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2011. Turkomatic: Automatic, Recursive Task and Workflow Design for Mechanical Turk. In *Proceedings of the 11th AAAI Conference on Human Computation (AAAIWS'11-11)*. AAAI Press, 91–96. <http://dl.acm.org/citation.cfm?id=2908698.2908716>
- [73] Karim R Lakhani and Robert G Wolf. 2003. Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Open Source Software Projects (September 2003)* (2003).
- [74] Nicolas J LaLone, Jess Kropczynski, and Andrea H Tapia. 2018. The Symbiotic Relationship of Crisis Response Professionals and Enthusiasts as Demonstrated by Reddit's User-Interface Over Time. In *ISCRAM*.
- [75] Jean Lave and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge university press.
- [76] ArLuther Lee. 2019. Police deny link to mysterious 'Umbrella Man,' who broke windows during riot. *Atlanta Journal Constitution* (2019). <https://www.ajc.com/news/police-deny-link-mysterious-umbrella-man-who-broke-windows-during-riot/3j8eSrRBHjvnx0PRwLYIP/>
- [77] Sunwha Lee, Sungho Lee, Yoojin Lee, Sanghoo Park, and Jinwoo Kim. 2014. Effect of competition and collaboration in social network game on intimacy among players. In *Proceedings of HCI Korea (HCiK '15)*. Hanbit Media, Inc., Seoul, KOR, 425–433.
- [78] Lawrence Lessig. 1998. The new Chicago school. *The Journal of Legal Studies* 27, S2 (1998), 661–691.
- [79] Lawrence Lessig. 1999. *Code and Other Laws of Cyberspace*. Basic Books, Inc., USA.
- [80] Ava Lew. 2021. The Unanticipated Use of Platforms in Disseminating Misinformation. (2021), 6.
- [81] Tianyi Li, Kurt Luther, and Chris North. 2018. CrowdIA: Solving Mysteries with Crowdsourced Sensemaking. 2 (2018), 105:1–105:29. Issue CSCW. <https://doi.org/10.1145/3274374>
- [82] Tianyi Li, Chandler J Manns, Chris North, and Kurt Luther. 2019. Dropping the baton? Understanding errors and bottlenecks in a crowdsourced sensemaking pipeline. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.
- [83] Kecheng Liu and Alan Dix. 1997. Norm governed agents in CSCW. In *The First International Workshop on Computational Semiotics*.
- [84] Ryan Louie, Kapil Garg, Jennie Werner, Allison Sun, Darren Gergle, and Haoqi Zhang. 2021. Opportunistic Collective Experiences: Identifying Shared Situations and Structuring Shared Activities at Distance. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 269 (jan 2021), 32 pages. <https://doi.org/10.1145/3434178>
- [85] Anders Sundnes Løvlie, Astrid Waagstein, and Peter Hyldgård. 2023. "How Trustworthy is this research?" Designing a Tool to Help Readers Understand Evidence and Uncertainty in Science Journalism. *Digital Journalism, forthcoming* (2023).
- [86] Kurt Luther, Scott Counts, Kristin B. Stecher, Aaron Hoff, and Paul Johns. 2009. Pathfinder: An Online Collaboration Environment for Citizen Scientists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, USA) (CHI '09). Association for Computing Machinery, New York, NY, USA, 239–248. <https://doi.org/10.1145/1518701.1518741>
- [87] Kurt Luther, Casey Fiesler, and Amy Bruckman. 2013. Redistributing Leadership in Online Creative Collaboration. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. Association for Computing Machinery, New York, NY, USA, 1007–1022. <https://doi.org/10.1145/2441776.2441891>
- [88] Kurt Luther, Nathan Hahn, Steven P. Dow, and Aniket Kittur. 2015. Crowdlines: Supporting Synthesis of Diverse Information Sources through Crowdsourced Outlines. In *Third AAAI Conference on Human Computation and Crowdsourcing*. <https://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11603>
- [89] Leticia S. Machado, Ricardo Rodrigo M. Melo, Cleidson R. B. de Souza, and Rafael Prikladnicki. 2021. Collaborative Behavior and Winning Challenges in Competitive Software Crowdsourcing. *Proceedings of the ACM on Human-Computer Interaction* 5, GROUP (July 2021), 220:1–220:25. <https://doi.org/10.1145/3463932>
- [90] Alexis C. Madrigal. 2013. Hey Reddit, Enough Boston Bombing Vigilantism. *The Atlantic* (April 2013). <http://www.theatlantic.com/technology/archive/2013/04/hey-reddit-enough-boston-bombing-vigilantism/275062/>
- [91] Narges Mahyar, Diana V. Nguyen, Maggie Chan, Jiayi Zheng, and Steven P. Dow. 2019. The Civic Data Deluge: Understanding the Challenges of Analyzing Large-Scale Community Input. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (San Diego, CA, USA) (DIS '19). Association for Computing Machinery, New York, NY, USA, 1171–1181. <https://doi.org/10.1145/3322276.3322354>
- [92] V. K. Chaithanya Manam and Alexander J. Quinn. 2018. WingIt: Efficient Refinement of Unclear Task Instructions. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*. <https://aaai.org/ocs/index.php/HCOMP/HCOMP18/paper/view/17931>
- [93] Winter Mason and Siddharth Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* 44, 1 (2012), 1–23.
- [94] Hana Matatov, Adina Bechhofer, Lora Aroyo, Ofra Amir, and Mor Naaman. 2018. DeJaVu: A System for Journalists to Collaboratively Address Visual Misinformation.
- [95] Melinda McClure Haughey, Meena Devii Muralikumar, Cameron A. Wood, and Kate Starbird. 2020. On the Misinformation Beat: Understanding the Work of Investigative Journalists Reporting on Problematic Information Online. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 133 (Oct. 2020), 22 pages. <https://doi.org/10.1145/3415204>
- [96] Lucas McDaniel, Erik Talvi, and Brian Hay. 2016. Capture the Flag as Cyber Security Introduction. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*. 5479–5486. <https://doi.org/10.1109/HICSS.2016.677>
- [97] Alexander Michael and Christof Lutteroth. 2020. Race Yourselves: A Longitudinal Exploration of Self-Competition Between Past, Present, and Future Performances in a VR Exergame. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–17. <https://doi.org/10.1145/3313831.3376256>
- [98] John Mills, Mark Lochrie, Tom Metcalfe, and Peter Bennett. 2018. NewsThings: Exploring Interdisciplinary IoT News Media Opportunities via User-Centred Design. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction* (Stockholm, Sweden) (TEL '18). Association for Computing Machinery, New York, NY, USA, 49–56. <https://doi.org/10.1145/3173225.3173267>
- [99] Molly Montgomery. 2020. Disinformation as a wicked problem: Why we need co-regulatory frameworks. <https://www.brookings.edu/research/disinformation-as-a-wicked-problem-why-we-need-co-regulatory-frameworks/>
- [100] Meredith Ringel Morris, Jarrod Lombardo, and Daniel Wigdor. 2010. WeSearch: Supporting Collaborative Search and Sensemaking on a Tabletop Display. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (Savannah, Georgia, USA) (CSCW '10)*. Association for Computing Machinery, New York, NY, USA, 401–410. <https://doi.org/10.1145/1718918.1718987>
- [101] Benedikt Morschheuer, Alexander Maedche, and Dominic Walter. 2017. Designing Cooperative Gamification: Conceptualization and Prototypical Implementation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2410–2421. <https://doi.org/10.1145/2998181.2998272>
- [102] Johnny Nhan, Laura Huey, and Ryan Broll. 2017. Diligantism: An analysis of crowdsourcing and the Boston marathon bombings. *The British journal of criminology* 57, 2 (2017), 341–361.
- [103] University of California Berkeley Human Rights Center. 2020. The Berkeley Protocol on Open Source Investigations | Berkeley Social Science. <https://matrix.berkeley.edu/research/berkeley-protocol-open-source-investigations>
- [104] Michael A Oren and Stephen B Gilbert. 2011. Framework for measuring social affinity for CSCW software. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 1387–1392.
- [105] Wanda J. Orlikowski. 1992. Learning from Notes: organizational issues in groupware implementation. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work (CSCW '92)*. Association for Computing Machinery, New York, NY, USA, 362–369. <https://doi.org/10.1145/143457.143549>
- [106] Alexandra Papoutsaki, Hua Guo, Danae Metaxa-Kakavouli, Connor Gramazio, Jeff Rasley, Wenting Xie, Guan Wang, and Jeff Huang. 2015. Crowdsourcing from Scratch: A Pragmatic Experiment in Data Collection by Novice Requesters. In *Third AAAI Conference on Human Computation and Crowdsourcing*. <https://www.aaai.org/ocs/index.php/HCOMP/HCOMP15/paper/view/11582>
- [107] Juhong Park, Alice Oh, and Suin Kim. 2017. Analysis of the Effect of Competition on Player Immersion and Engagement in a Mobile Game. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 6.
- [108] Election Integrity Partnership. 2021. The Long Fuse: Misinformation and the 2020 Election.
- [109] Gordon Pennycook and David G. Rand. 2019. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences* 116, 7 (Feb. 2019), 2521–2526. <https://doi.org/10.1073/pnas.1806781116> Publisher: National Academy of Sciences Section: Social Sciences.
- [110] Christoph Perger, Ellsworth LeDrew, Linda See, and Steffen Fritz. 2014. Geography Geo-Wiki in the classroom: Using crowdsourcing to enhance geographical teaching. *Future Internet* 6, 4 (2014), 597–611.

- [111] Paul R Pintrich. 2004. A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational psychology review* 16, 4 (2004), 385–407.
- [112] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.
- [113] Emily Porter, Chris Bopp, Elizabeth Gerber, and Amy Volda. 2017. Reappropriating Hackathons: The Production Work of the CHI4Good Day of Service. (2017), 5.
- [114] Nicolas Pröllochs. 2022. Community-based fact-checking on Twitter's Birdwatch platform. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 16. 794–805.
- [115] Diego A Reinero, Elizabeth A Harris, Steve Rathje, Annie Duke, and Jay J Van Bavel. 2023. Partisans are more likely to entrench their beliefs in misinformation when political outgroup members fact-check claims. <https://doi.org/10.31234/osf.io/z4df3>
- [116] Daniela Retelny, Michael S. Bernstein, and Melissa A. Valentine. 2017. No Workflow Can Ever Be Enough: How Crowdsourcing Workflows Constrain Complex Work. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 89 (Dec. 2017), 23 pages. <https://doi.org/10.1145/3134724>
- [117] Horst WJ Rittel and Melvin M Webber. 1973. Dilemmas in a general theory of planning. *Policy sciences* 4, 2 (1973), 155–169.
- [118] Dylan Rogers, Ray Adams, Row Farr, Matt Mahmoudi, and Mitra Dastbaz. 2020. How Amnesty's Digital Verification Corps documented the November 2019 protests in Iran. <https://citizenvidence.org/2020/12/07/cambridge-dvc-iran/Section: Digital Verification>.
- [119] Lara S. G. Piccolo, Somya Joshi, Evangelos Karapanos, and Tracie Farrell. 2019. Challenging Misinformation: Exploring Limits and Approaches | SpringerLink. https://link.springer.com/chapter/10.1007/978-3-030-29390-1_68
- [120] Mohammed Saeed, Nicolas Traub, Maelle Nicolas, Gianluca Demartini, and Paolo Papotti. 2022. Crowdsourced Fact-Checking at Twitter: How Does the Crowd Compare With Experts?. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1736–1746.
- [121] Kjeld Schmidt. 1994. Cooperative work and its articulation: requirements for computer support. *Le travail humain* (1994), 345–366.
- [122] Kjeld Schmidt and Ina Wagner. 2004. Ordering systems: Coordinative practices and artifacts in architectural design and planning. *Computer Supported Cooperative Work (CSCW)* 13 (2004), 349–408.
- [123] Chengcheng Shao, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. Hoaxy: A Platform for Tracking Online Misinformation. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 745–750. <https://doi.org/10.1145/2872518.2890098>
- [124] Aditya K Sood and Richard Endbody. 2014. Chapter 2 - Intelligence Gathering. In *Targeted Cyber Attacks*. Aditya K Sood and Richard Endbody (Eds.). Syngress, Boston, 11–21. <https://doi.org/10.1016/B978-0-12-800604-7.00002-4>
- [125] James P Spradley. 2016. *Participant observation*. Waveland Press.
- [126] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.
- [127] Kate Starbird, Dharma Dailey, Owla Mohamed, Gina Lee, and Emma S. Spiro. 2018. *Engage Early, Correct More: How Journalists Participate in False Rumors Online during Crisis Events*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173679>
- [128] Kevin Stowe, Martha Palmer, Jennings Anderson, Marina Kogan, Leysia Palen, Kenneth M. Anderson, Rebecca Morss, Julie Demuth, and Heather Lazrus. 2018. Developing and Evaluating Annotation Procedures for Twitter Data during Hazard Events. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 133–143. <https://www.aclweb.org/anthology/W18-4915>
- [129] Tarja Susi, Mikael Johannesson, and Per Backlund. 2007. Serious games: An overview. (2007).
- [130] Andrea H Tapia and Nicolas J LaLone. 2019. Crowdsourcing investigations: Crowd participation in identifying the bomb and bomber from the Boston marathon bombing. In *Crowdsourcing: Concepts, Methodologies, Tools, and Applications*. IGI Global, 1433–1450.
- [131] Yla Tausczik and Mark Boons. 2018. Distributed Knowledge in Crowds: Crowd Performance on Hidden Profile Tasks. In *Twelfth International AAAI Conference on Web and Social Media*. <https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17817>
- [132] Yla Tausczik and Ping Wang. 2017. To Share, or Not to Share? Community-Level Collaboration in Open Innovation Contests. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 100:1–100:23. <https://doi.org/10.1145/3134735>
- [133] Ara Tekian, Christopher J Watling, Trudie E Roberts, Yvonne Steinert, and John Norcini. 2017. Qualitative and quantitative feedback in the context of competency-based education. *Medical teacher* 39, 12 (2017), 1245–1249.
- [134] The Global Disinformation Lab. 2022. The Global Disinformation Lab at the University of Texas at Austin. <https://gdi.org/>
- [135] Trace Labs Twitter Account. 2022. TraceLabs Twitter Account. <https://twitter.com/tracelabs/status/1558831625986777088>
- [136] Gustavo Umbelino, Matin Yarmand, Samuel Blake, Vivian Ta, Amy Luo, and Steven P. Dow. 2021. ProtoTeams: Supporting Team Dating in Co-Located Settings. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 273 (jan 2021), 27 pages. <https://doi.org/10.1145/3434182>
- [137] Sukrit Venkatagiri, Aakash Gautam, and Kurt Luther. 2021. CrowdSolve: Managing Tensions in an Expert-Led Crowdsourced Investigation. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 118 (April 2021). <https://doi.org/10.1145/3449192>
- [138] Sukrit Venkatagiri, Jacob Thebault-Spieker, Rachel Kohler, John Purviance, Rifat Sabbir Mansur, and Kurt Luther. 2019. GroundTruth: Augmenting Expert Image Geolocation with Crowdsourcing and Shared Representations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 107 (Nov. 2019), 30 pages. <https://doi.org/10.1145/3359209>
- [139] Sukrit Venkatagiri, Tianjiao Yu, Vikram Mohanty, and Kurt Luther. 2021. Sedition Hunters: Countering Extremism through Collective Action. In *CSCW 2021 Workshop on Addressing Challenges and Opportunities in Online Extremism Research: An Interdisciplinary Perspective*.
- [140] Katherine Vogt, Lauren Bradel, Christopher Andrews, Chris North, Alex Endert, and Duke Hutchings. 2011. Co-located Collaborative Sensemaking on a Large High-Resolution Display with Multiple Input Devices. In *Human-Computer Interaction – INTERACT 2011*. Springer Berlin Heidelberg, Berlin, Heidelberg, 589–604.
- [141] Luis von Ahn and Laura Dabbish. 2008. Designing Games with a Purpose. *Commun. ACM* 51, 8 (Aug. 2008), 58–67. <https://doi.org/10.1145/1378704.1378719>
- [142] Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policymaking.
- [143] Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. 2019. WhatHack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3290605.3300338>
- [144] Etienne Wenger et al. 1998. Communities of practice: Learning as a social system. *Systems thinker* 9, 5 (1998), 2–3.
- [145] Joanne I White, Leysia Palen, and Kenneth M Anderson. 2014. Digital mobilization in disaster response: the work & self-organization of on-line pet advocates in response to hurricane sandy. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. ACM, 866–876.
- [146] Heather J. Williams and Ilana Blum. 2018. *Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise*. Technical Report. RAND Corporation. https://www.rand.org/pubs/research_reports/RR1964.html
- [147] Stefan Wojcik, Sophie Hilgard, Nick Judd, Delia Mocanu, Stephen Ragain, MB Hunzaker, Keith Coleman, and Jay Baxter. 2022. Birdwatch: Crowd Wisdom and Bridging Algorithms can Inform Understanding and Reduce the Spread of Misinformation. *arXiv preprint arXiv:2210.15723* (2022).
- [148] Nick Yee, Nicolas Ducheneaut, and Les Nelson. 2012. Online gaming motivations scale: development and validation. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 4.
- [149] Tianjiao Yu, Sukrit Venkatagiri, Ismini Lourentzou, and Kurt Luther. 2023. Sedition Hunters: A Quantitative Study of the Crowdsourced Investigation into the 2021 U.S. Capitol Attack. In *Proceedings of the ACM Web Conference 2023 (Austin, TX, USA) (WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3849–3858. <https://doi.org/10.1145/3543507.3583514>
- [150] Himanshu Zade, Megan Woodruff, Erika Johnson, Mariah Stanley, Zhennan Zhou, Minh Tu Huynh, Alissa Elizabeth Acheson, Gary Hsieh, and Kate Starbird. 2023. Tweet Trajectory and AMPS-based Contextual Cues can Help Users Identify Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–26.
- [151] Haoqi Zhang, Matthew W. Easterday, Elizabeth M. Gerber, Daniel Rees Lewis, and Leesha Maliakal. 2017. Agile Research Studios: Orchestrating Communities of Practice to Advance Research Training. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (Portland, Oregon, USA) (CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 220–232. <https://doi.org/10.1145/2998181.2998199>
- [152] John Zimmerman and Jodi Forlizzi. 2014. Research through design in HCI. In *Ways of Knowing in HCI*. Springer, 167–189.
- [153] John Zimmerman, Erik Stolterman, and Jodi Forlizzi. 2010. An Analysis and Critique of Research through Design: Towards a Formalization of a Research Approach. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems (Aarhus, Denmark) (DIS '10)*. Association for Computing Machinery, New York, NY, USA, 310–319. <https://doi.org/10.1145/1858171.1858228>

A APPENDIX

New Evidence

Name:

URL:

Topic:

Related Evidence:

Capture Flags:

☒ -Discovery Flag

Details:

Evaluate Your Flag

Research Method:

Recency:

Difficulty:

New Flag

Figure 2: An example of an early low-fidelity wireframe that we created in Balsamiq. This “New Evidence” page shows how a user would create a new evidence piece by specifying the name of the evidence, the URL for a social media post and what topic (now narrative thread) it is related to. This early version also allowed users to reference other evidence pieces to construct a broader narrative or story. At the bottom of the page, the user “captures” the discovery flag that awards them points for documenting this new evidence piece.

OSINT CTF

About | Events | Tools

Logged in as [user1](#) | [Log out](#)

Event1

Threads | **Flags** | Tasks | Submissions

Leaderboard | MyTeam (5000pts)

All Flags | New Flag

Sort by:

Filter by...

☒ Thread

☒ Flag Status


☒ Your Team

☐

☐

Ballots found dumpster

approved




Ballots reportedly found in dumpster...

Submitted 5min ago by [user3](#)

2000 mules voter fraud

rejected




2000 mules video suggests rampant voter fraud

Submitted 3min ago by [user3](#)

Mail-in voting confusion

approved




Arizona voters report confusion about mail-in voting

Submitted 7min ago by [user1](#)

Election official biased

pending




Reports of an election official invalidating ballots

Submitted 1min ago by [user6](#)

Sharpiegate (ballots)

approved




Sharpies reportedly used to disenfranchise voters

Submitted 15min ago by [user4](#)

Ballots in dumpster

pending



Ballots reportedly found in dumpster

Submitted 5min ago by [user2](#)

Figure 3: An example of a high-fidelity wireframe that we created in Google Slides. This “Flag List View” page shows users a card-like interface of all flags submitted during an event (by all users). On the left are filters allowing the user to narrow the set of flags based on the associated thread, its status, the team that created it, etc. On top are two menu bars that link to other pages. The topmost menu bar links to an about page, a page displaying all prior and future events, and all tools created by other users. The second menu bar links to pages that display narrative threads, flags, tasks, submissions (now renamed to evidence), the leaderboard, and the users’ “My Team” page.

2571

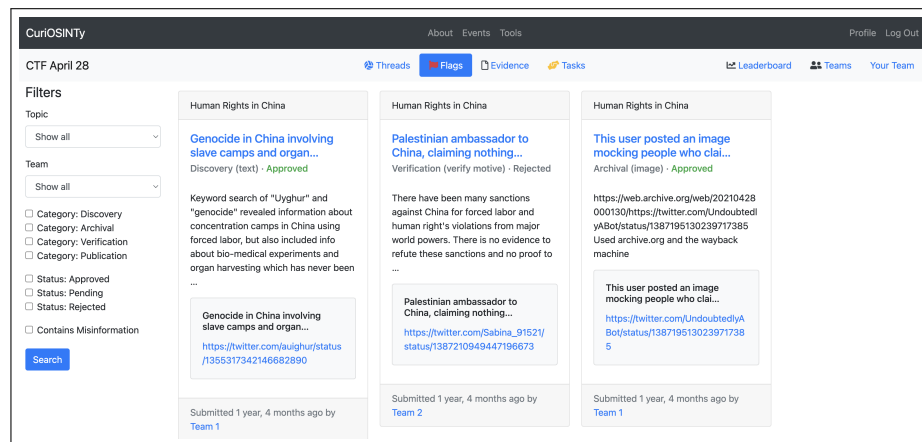


Figure 4: A screenshot of the actual CoSINT platform for the “Flag List View.” This page shows users a card-like interface of all flags submitted during an event (by all users).

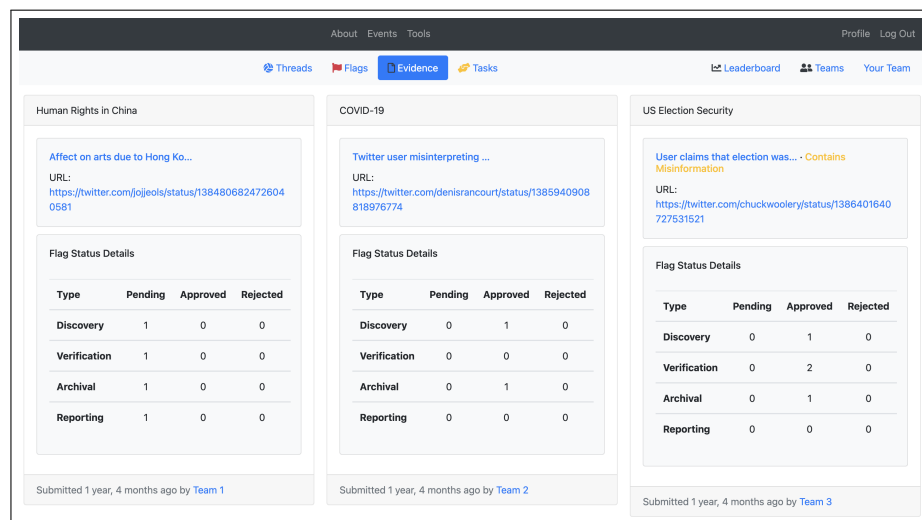


Figure 5: A screenshot of the actual CoSINT platform for the “Evidence List View.”