OSINT Research Studios: A Flexible Crowdsourcing Framework to Scale Up Open Source Intelligence Investigations

ANIRBAN MUKHOPADHYAY, Department of Computer Science, Virginia Tech, USA
SUKRIT VENKATAGIRI*, Department of Computer Science, Swarthmore College, USA
KURT LUTHER, Department of Computer Science, Virginia Tech, USA

Open Source Intelligence (OSINT) investigations, which rely entirely on publicly available data such as social media, play an increasingly important role in solving crimes and holding governments accountable. The growing volume of data and complex nature of tasks, however, means there is a pressing need to scale and speed up OSINT investigations. Expert-led crowdsourcing approaches show promise, but tend to either focus on narrow tasks or domains, or require resource-intense, long-term relationships between expert investigators and crowds. We address this gap by providing a flexible framework that enables investigators across domains to enlist crowdsourced support for discovery and verification of OSINT. We use a design-based research (DBR) approach to develop OSINT Research Studios (ORS), a sociotechnical system in which novice crowds are trained to support professional investigators with complex OSINT investigations. Through our qualitative evaluation, we found that ORS facilitates ethical and effective OSINT investigations across multiple domains. We also discuss broader implications of expert–crowd collaboration and opportunities for future work.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing systems and tools**; **Empirical studies in HCI**.

Additional Key Words and Phrases: OSINT, open source intelligence, design-based research, complex sense-making, collaboration, crowdsourcing

1 INTRODUCTION

Open Source Intelligence (OSINT) involves the use of publicly available information to generate intelligence that addresses a particular need [133]. OSINT analysis is increasingly used by journalists [75], human rights activists [56, 80], and law enforcement [59, 100], among other professions. For example, OSINT is used to verify breaking news and combat disinformation, monitor international weapons development, locate suspected terrorists and victims of human trafficking, and document war crimes [36, 42, 80, 134]. These investigations are widely recognized for their ability to use data sources like social media, satellite imagery, flight tracking information, and metadata from smartphones and IoT devices to conduct investigations [93, 134].

---

*This work was completed while the author was at Virginia Tech and the University of Washington.

Authors' addresses: Anirban Mukhopadhyay, Department of Computer Science, Virginia Tech, Blacksburg, VA, USA, anirban@vt.edu; Sukrit Venkatagiri, sukrit@swarthmore.edu, Department of Computer Science, Swarthmore College, Swarthmore, PA, USA; Kurt Luther, Department of Computer Science, Virginia Tech, Arlington, VA, USA, kluther@vt.edu.

There is a pressing need to scale and speed up OSINT investigations, especially those focused on time-sensitive topics such as documenting war crimes or addressing disinformation. Apart from time pressure, investigators find it difficult to manage the growing volume of data that they must process [5, 65]. The ephemerality of online information [56] and prevalence of misleading or misrepresented content [56, 78] pose additional challenges. Many investigators may also lack the data science and software development skills required to fully utilize OSINT tools and techniques [91], presenting another barrier to the adoption of OSINT.

There are two common approaches to scale and speed up such investigations: automation [62, 73, 115] and crowdsourcing [3, 53, 90, 104]. OSINT investigations require creatively leveraging multiple tools and techniques [36]. Thus, software tools cannot fully automate and scale up the complex sensemaking involved in investigative work [88]. Computational methods have produced high volumes of unverifiable information, which has low utility for experts [95]. OSINT analysts also aim to minimize dependency on custom-built tools, as they can easily become obsolete [97].

Crowdsourcing provides a second, more flexible way to augment investigators' complex sense-making efforts. However, unfettered access to information, as with OSINT investigations, has resulted in "bottom-up" crowdsourced investigations that exhibit biased results [29], doxxing [98], and even sabotaging of ongoing investigations [121]. Expert supervision has sometimes resulted in more successful investigations in terms of both process (ethical, safe, and privacy-protecting) and outcomes (results). For example, CrowdSolve [124] described an effort in which law enforcement officials and experts supervised a crowd of 250 true crime enthusiasts in investigating two cold cases in a co-located, weekend-long event where information was tightly controlled rather than publicly available or open source. The Human Rights Center (HRC) at the University of California, Berkeley trains law students to partner with professional human rights investigators on OSINT projects lasting months or longer [2]. GroundTruth [126] brought together the complementary strengths of experts and an online, novice crowd for performing image geolocation, an important but very specific type of OSINT task. While these diverse projects show promise, they tend to either focus on narrow tasks or domains, or require resource-intense, long-term relationships between expert investigators and crowds. There is a need for a more flexible approach that enables investigators across a diverse set of domains to enlist crowdsourced support for a wide variety of OSINT investigation tasks.

In this paper, we use a design-based research (DBR) approach [47] to develop OSINT Research Studios (ORS), a sociotechnical framework in which novice crowds are trained and provided with scaffolding to support professional investigators with complex OSINT investigations. We developed ORS in a semester-long class with 30 trained students serving as the crowd. Using an industry-standard OSINT model [133], we developed five types of macrotasks, i.e., discovery and verification techniques for OSINT. We recruited five OSINT experts who worked as journalists, fact-checkers, law enforcement officers, and human rights investigators to divide their investigations into one or more of the five macrotask types. We evaluated ORS through a semester-long deployment of the model and five study sessions where experts and the crowd conducted real-world investigations.

Our qualitative evaluation revealed that the macrotasks were relevant to expert work practices. Experts said that the sessions were productive, and mentioned strengths like speed, safety, quality and quantity of submissions, and the crowd's adaptability in response to expert feedback. The crowd enjoyed working with experts and felt that they successfully applied their OSINT skills. We also discuss experts' involvement during the sessions, the crowd's perceptions of their contributions, and how the type of tasks impacted the crowd's experience.

Our paper makes the following three contributions:

(1) We make a conceptual contribution by developing training modules for crowdsourcing diverse and complex OSINT tasks.
(2) Using design-based research (DBR), we present OSINT Research Studios (ORS), a sociotechnical framework that enables collaboration between investigators and a trained crowd. Experts from multiple domains can delegate OSINT tasks to a trained crowd and generate leads for their investigation.
(3) We evaluate ORS through multiple deployments during the OSINT lab course. We find ORS is effective for scaling and speeding up expert-led OSINT investigations and discuss the key design elements that make it successful.

2 BACKGROUND AND RELATED WORK

2.1 OSINT Investigations

OSINT investigations must deal with large amounts of digital content, including from social media platforms, search engines, online databases, among other sources. These investigations frequently employ a diverse set of tools and techniques to analyze digital traces, augmenting investigations in domains like journalism [75], law enforcement [59, 100], and human rights advocacy [56, 80]. Numerous applications are possible because of the vast quantities of detailed data available online alongside tools used to gather and analyze this data[92].

OSINT can be divided into four stages: information discovery, verification, archival and reporting [133]. Researchers have developed systems to support each of these processes. CrowdTangle [62] and Algorithm Tips [53] help in automating discovery; Hoaxy [115] and DejaVu [90] provide structure for verification tasks, such as verifying Twitter bot networks and accounts; The Web Archive Workbench [76] supports digital archiving; and Hunchly [4] is used for documenting and reporting. The number of open source OSINT tools is growing quickly with citizen journalism organizations like Bellingcat [37] collaborating with and funding software developers and data scientists [1]. To map this space, aggregators like OSINT Explorer collate these resources and provide guides on which OSINT tools to use [23]. However, most current tools only support *individual* steps of the OSINT process. These tools are also frequently rendered obsolete by changes in the underlying information architecture controlled by large online platforms (e.g., social media platforms, search engines). In contrast, we leverage the adaptability of a trained crowd to approach multiple tasks using a conceptual approach rather than relying on specific or customized tools.

OSINT investigations are carried out by both novices and experts. The OSINT community has grown rapidly in recent years due to its low barrier to entry and gained attention with influential investigations of the Ukraine-Russia conflict [69] and the storming of the U.S. Capitol in January 2021 [45]. Analysts in the OSINT community volunteer for organizations like Bellingcat [7], Trace Labs [50], and the Syrian Archive [19]. These organizations run collaborative projects to analyze open information from global events including wars, missing person investigations [6], human rights violations [32], and election irregularities [113]. There have been successful instances of this form of collaboration, but there is little structure to how they are performed [48]. Lack of intelligence training and coherence in preparing policy options for decision-makers among OSINT enthusiasts emerged as challenges for collaboration between the OSINT community and law enforcement agencies [48]. Our work addresses this gap in structuring crowdsourced OSINT investigations and augmenting expert work practices.

OSINT investigations can have wide variation in scope and depth. Existing collaborations with OSINT analysts for investigative journalism and human rights advocacy predominantly tackle long-running investigations that require a deep understanding of the context and its evolution [2, 97].

These investigations require high levels of involvement, communication and training. Instead, our work here seeks to support rapid, more targeted tasks, scaling up and speeding up larger, more complex investigations led by experts.

Finally, OSINT, though popular in its application across many domains, has not received much research attention in the form of frameworks and systems that help to scale up and speed up these investigations [27, 36, 79]. Our work contributes to the development of OSINT Research Studios (ORS), a collaborative crowdsourcing framework that can support experts with multiple steps within complex OSINT investigations. The work also demonstrates that OSINT can be a valuable domain of study for the CSCW and HCI community.

2.2 Collaborative Sensemaking and Crowdsourced Investigations

2.2.1 Collaborative sensemaking. Investigations are a type of sensemaking task, as they involve collecting and analyzing large amounts of information to reach a conclusion [26, 52, 126]. OSINT investigations, when broken down into the steps of the OSINT cycle [36, 133] — discover, verify, preserve, and publish — follow the sensemaking process closely. Collaborations also play a crucial role in facilitating sensemaking by dividing tasks related to discovery and verification and incorporating diverse perspectives during data analysis [61].

Previous studies on supporting collaborative sensemaking have primarily focused on co-located teams working synchronously [124, 127], distributed crowd working asynchronously [51, 61, 88], and even a distributed crowd working synchronously [125]. In contrast, this paper studies a new, hybrid setting, with remote experts collaborating synchronously with co-located crowds distributed across two locations. We present a semester-long deployment with 30 university students in a classroom setting, accommodating experts from across the United States through remote participation.

Extensive research has studied the development of collaborative systems to assist fact-checkers and journalists. Various collaborative tools have been designed specifically for fact-checking news articles [114], videos [44], and visual disinformation [90, 126]. Additionally, systems like Newstrition [14] and Checkdesk [11] apply crowdsourcing approaches to verify information. The Datavoidant tool [63] facilitated human-AI collaboration to empower journalists in addressing data voids through information discovery and verification. The CAPER tool [27] aids law enforcement agents in collaborating for sensemaking tasks to prevent organized crime. Garcia et al. [30, 31] studied the exploration of social media data for human rights investigations and public safety. Most closely related to our work, Venkatagiri et al. [125] enabled a trained crowd to be effective in debunking online misinformation using OSINT techniques by introducing collaboration in a competitive Capture-the-Flag environment. We contribute to this line of work by developing a flexible expert-crowd collaborative framework where investigators can enlist crowdsourced support for a wide variety of tasks within broader, more complex OSINT investigations.

Studies investigating the effectiveness of crowdsourced sensemaking and fact-checking are highly relevant to our work. For instance, Arif et al. [33] and Dailey et al. [51] demonstrated that distributed crowds could effectively debunk rumors online, while Saeed et al. [112] revealed that crowdsourced fact-checking on Twitter often performs as well as professional fact-checkers. Experimental investigations into the efficacy of crowdsourced fact-checking by Pennycook and Rand [104] and Allen et al. [28] found that crowdsourced trustworthiness ratings can distinguish between authentic and fake news sources. However, Godel et al. [66] discovered that real-time crowdsourced veracity ratings performed worse than those generated by professional fact-checkers. While prior crowdsourcing approaches predominantly focused on studying online crowds operating independently without expert supervision, we show that a trained crowd can augment investigations of online information when led by experts.

2.2.2 Crowdsourced investigations. CSCW literature presents three types of crowdsourced investigations: top-down [25, 107], bottom-up [58, 77], and hybrid investigations [124]. Bottom-up investigations are driven by non-professional crowds and tend to move through the sharing, validation and analysis stages of an investigation in an online setting. Investigations have been studied in the context of collective sensemaking on social media during crisis events [51], e.g., analysis of photos related to 2013 Boston Marathon bombings [77], as well as correcting online information on social media [33]. Though potentially effective, these investigations have resulted in harmful behavior like misidenfications [86], doxxing [98], and perpetuating conspiracy theories [94].

Other related investigations demonstrate that crowds can collaborate under the guidance of experts to augment investigations [88, 124, 126]. Among hybrid investigations, GroundTruth [126] demonstrated crowd-augmented expert work using a novice crowd to reduce the search area for geolocating images, an important OSINT task. Venkatagiri et al.[124] described an expert-led crowdsourcing model through CrowdSolve, which is characterized by experts (law enforcement officers in this case) leading investigations by providing resources, training and feedback; and the crowd performing analytical tasks to generate leads. Our work has two key differences compared with CrowdSolve: 1) here, OSINT investigations are performed without the use of restricted information and led by experts from *multiple* domains; and 2) the hybrid sessions we studied are mediated by online collaboration and not restricted to working within a co-located setting involving physical paper case files. Our work provides additional flexibility to the concept of expert-led crowdsourcing [124] for diverse and complex OSINT investigations involving experts from multiple domains.

2.3 Decomposition and Training for Crowdsourcing Complex Work

OSINT analysis involves complex problem-solving tasks. One common approach to crowdsourcing is to decompose complex tasks into smaller subtasks that are easier to handle [46, 51, 83, 103]. Among related complex tasks, entire sensemaking processes have been decomposed into microtasks to solve fictional murder mysteries and terrorist plots [88, 89] and generating text content from journalism to how-to guides [35, 38, 70]. While microtask-based crowdsourcing offers scalability and efficiency by distributing small tasks among a large crowd, it may not be suitable for complex, creative OSINT investigations that require retaining contextual information. Macrotasks, on the other hand, allow for deeper engagement, contextual comprehension, and complex problem-solving [54]. In our work, we contribute a decomposition of the OSINT tasks of discovery and verification into macrotasks. We describe the steps and skills required for five types of macrotasks that are transferable across multiple domains.

Even with task decomposition, OSINT exhibits *complex problem-solving tasks* [54] that suggests multiple possible strategies to the crowd; workers can arrive at solutions, but not without relevant skills [54]. Decomposition also creates added coordination challenges that can be addressed through appropriate workflows [119]. We investigate how to crowdsource complex OSINT tasks, which are relatively less studied in crowdsourcing.

Previous studies have developed effective ways of training crowdworkers for complex crowdsourcing tasks [67, 82, 99]. Training provides more agency to the crowd and enables them to perform investigations without rigid roles and constrained workflows — strategies which have been effective for other types of creative and complex work [26, 109]. There are multiple ways to conduct training [96, 102, 138]. First, crowdworkers can gain experience by solving problems that are relevant to the task [24]. Second, reviewing expert and peer solutions can improve crowd performance [118]. Third, crowdworkers can improve their performance based on self-assessment and expert feedback [55]. Doroudi et al. [54] compared multiple such training strategies for a complex web search problem, which is closely related to OSINT tasks that involve consulting

multiple sources of information and arriving at a conclusion [34]. They found that all training strategies improved performance compared to the no training condition. Training through expert examples improved crowdworker accuracy the most. Wang et al. [128] explored training based on analytical thinking skills for historical analysis, which also closely relates to our strategy of imparting skills to help the crowd approach challenging OSINT tasks instead of relying on any particular tool. They found that crowdworkers developed domain expertise and performed at least as well as other training strategies mentioned above. Unlike prior work, we develop a semester-long training process focused on OSINT tasks.

Previous examples show that only minimal active training is provided to volunteers within real-world OSINT investigations [48, 71]. In our work, we contribute an OSINT crowd training module that combines multiple training strategies. Training includes demonstrations for acquiring OSINT skills, honing them during practice and expert sessions, as well as feedback and self-evaluations. We also illuminate how training can be effective within crowdsourced OSINT investigations.

3 DESIGN CHALLENGES AND OPPORTUNITIES FOR CROWDSOURCING COMPLEX OSINT INVESTIGATIONS

Based on prior work, we discuss the four main types of challenges in designing crowdsourcing solutions to support OSINT investigations.

As previously mentioned, OSINT investigators are overwhelmed due to the large volume of information involved and the complex nature of investigative tasks [5]. Effectively scaling up OSINT investigations can help investigators improve the speed and/or accuracy of their work. As automation by itself is difficult to achieve for highly contextual and nuanced tasks, crowdsourcing provides a viable approach. Crowdsourcing can leverage humans' creative and sensemaking capabilities to augment ongoing investigations [124, 126]. However, there are major challenges faced by investigators from domains like journalism, human rights investigations, and law enforcement who seek crowdsourcing support for the analysis of open online information [48]. We identify four design *challenges* based on prior work and OSINT investigation reports, detailed below, along with *design goals* for a sociotechnical framework to address these challenges.

3.1 Delegation of Complex OSINT Tasks

Previous studies have shown that investigators spend most of their time in the discovery and verification phases of OSINT [95]. The increasing volume of digital data online presents a significant challenge for investigators to manage effectively [5, 65]. Additionally, the transient nature of open online information [56] and the widespread presence of deepfakes, mis-, and disinformation [56, 78] further complicate the tasks of discovery and verification, posing challenges for successful investigations. These tasks belong to the class of problems known as *complex problem-solving tasks* that have a number of potential strategies and are difficult to solve without acquiring relevant skills [54]. Experts do not have a way to crowdsource such tasks using existing workflows.

3.1.1 Discovery.

Challenges: Micallef et al. [95] specify that monitoring social media for interesting content and contextualizing it is one of the most time-consuming parts of fact-checking. Law enforcement involves collecting and analyzing open source information for corroboration as well, but in a way that is admissible in court. The focus is on referring to sources that are genuine and ethically collecting the information [64]. Human rights investigations are long-running and highlight transparency in data collection. The rapid pace of generation, multiple platforms, and recycling of content make the collection process more challenging [95]. The strategy to discover relevant information needs to adapt based on newly found information and pivot around them.

Design Goals: Investigators need to discover relevant information by understanding the evolution of a topic, gathering relevant hashtags, and identifying actors who spread the information [15, 95]. These topics can be broad, like COVID misinformation, anti-vaccine protests, conspiracy theories, and many more. Or they can be more specific topics like a calamity, investigating claims by a local government representative, or reporting on local phenomena like crime and illegal activities. These activities require mining information tied to a particular geographical location and time range. Investigators also need to look for potential mis/disinformation around hot topics.

3.1.2 Verification.

Challenges: Verification of online information is crucial in human rights, law enforcement investigations and fact-checking. Human rights investigations involve documenting events within specific regions, often utilizing Geospatial Information Systems (GIS) and performing geolocation [139]. However, the lack of metadata and geotags for social media content poses challenges in geolocation efforts [139]. To ensure the credibility of information, various professionals such as law enforcement officers, journalists, and fact-checkers examine the background of the account that generated the content, while being mindful of bot accounts and coordinated campaigns that propagate disinformation [60, 129]. Image analysis becomes a vital component of the verification process in journalism, addressing the difficult task of identifying manipulated or fake visual information [95]. In addition, fact-checking plays a crucial role by gathering trusted information to assess the veracity of claims [95]. However, the complexity of these tasks, combined with limitations in time, personnel, and analysis skills among experts, presents challenges in conducting successful investigations [91].

Design Goals: Investigators need to verify any information that is collected from publicly available sources for their use. According to Wardle [129], there are four main elements that need to be verified: 1) Provenance — is it original or has it been used before in a different context? 2) Source — What is the background of the account that created the content? Is it a bot? 3) Time — When was it created? 4) Location — Where is the place shown in it? These tasks, combined with fact-checking and image analysis, are essential crowd skills that meet the verification requirements of investigators.

3.2 Safety, Privacy, and Other Ethical Considerations

Challenges: Previous instances of crowdsourced OSINT investigations have led to biased results [29], misidentifications [86], and vigilante behavior like doxxing [98]. The high-profile failures are perhaps more prominent than successful investigations [77]. Investigators are also wary of leaks and sabotaging ongoing investigations during collaboration with crowds [121]. Investigators across domains benefit from careful consideration of ethics, safety and privacy in their work.

Design Goals: OSINT comes with its own ethos: prioritizing transparency, avoiding subterfuge, and limiting investigations to passive reconnaissance [36]. This ethos can be leveraged to enable more successful investigations and reduce ethical mishaps. Crowdsourced investigators can operationalize these values, e.g., 1) Transparency: documenting the process of investigation for reproducibility; 2) Avoiding subterfuge: prohibiting forms of hacking and impersonation to gather private information; and 3) Passive reconnaissance: ensuring investigators view the information but don't engage. We must ensure that experts are actively overseeing and guiding the investigation. Experts should be responsible for validating information and making final decisions based on the crowd's input.

3.3 Organizational Overheads of Synchronous Collaboration

Challenges: Collaboration and communication are key elements of OSINT investigations [36]. Investigators find it hard to collaborate with a crowd in real time. Generally, a single expert has to work with a group of volunteers and the overheads of coordination and communication impact the effectiveness of the investigation [124]. Information silos and duplication of effort within the crowd are known issues with both competitive and collaborative OSINT investigations [124].

Design Goals: Both expert and crowd investigators need technological solutions to orchestrate resources that allow them to document and present their findings. Information has to be easily accessible and open to all participants during investigations. Having a robust infrastructure can mitigate risks of data loss and usability issues.

3.4 Maintaining the quality of investigation

Challenges: Investigators need to insist on highest standards to establish the legitimacy of their reports and be confident in facing public scrutiny [97]. In previous microtask-based crowdsourcing models for investigative work, experts cannot make interventions or provide feedback to improve the quality of crowd results [88, 126]. The success of the particular task is overly dependent on the initial design and how results from the microtasks are aggregated.

Design Goals: Based on previous crowdsourcing studies for complex work, feedback from experts and self-evaluation can be helpful in improving the results [55]. The crowd needs to improve their work based on expert feedback. Crowd submissions need to meet these characteristics: 1) Relevant: the claim is relevant to the topic they selected; 2) Specific: the claim can be attributed to a specific statement or piece of content, such as a tweet, photo, video, or quote in an article; and 3) Verifiable: the claim has the potential to be verified, so it must be a factual statement (i.e., not an opinion) that can be shown to be true or false. Relevance and verifiability have been used earlier in crowdsourced OSINT capture-the-flag events for scoring submissions [18].

4 A DESIGN-BASED RESEARCH APPROACH TO EXPERT-LED OSINT INVESTIGATIONS

4.1 Our Approach

Collaboration and crowdsourcing have been embraced by investigators as a way to increase the scale and speed of their work [61]. Previous studies [82, 96, 99, 102] have shown that training workers to prepare them with sophisticated domain knowledge can be effective in completing complex tasks. Based on the challenges faced by investigators from multiple domains while dealing with publicly available information, we argue that collaboration between the investigators and a crowd trained in OSINT can be helpful. The first part of the problem deals with developing a comprehensive training module that can enable the crowd to perform common OSINT tasks used by investigators, understand the ethical considerations, apply relevant skills in different contexts, and present information that meets the expert requirements. The second part involves getting the trained crowd to collaborate with experts synchronously.

In this work, we sought to address the challenges we identified in Section 3 through the development of OSINT Research Studios (ORS), where we empower a group of students to apply OSINT analysis to augment real-world expert investigations. This group of students serves as the crowd in our crowdsourcing framework for the study. Previous research has shown that a crowd of students can effectively test new forms of crowdsourcing and generate recommendations for the process [135, 136]. Students have worked carefully and safely on real-world OSINT investigations for cyber vulnerability assessments and human rights as part of experiential learning previously [2, 101].

Here, we design a sociotechnical framework where experts can get valuable crowdsourced support on diverse, complex OSINT tasks by addressing the design challenges.

4.2 Methods

We take a design-based research (DBR) approach to develop the ORS model for overcoming the challenges with current crowdsourced OSINT investigations. DBR [47, 57, 137] is characterized by iterative cycles of design and evaluation to develop insights into learning experiences. DBR also provides a way for researchers to iterate on and study complex models simultaneously [57, 106, 137]. We leverage that to evaluate the collaborative framework as it evolves with deployments in a classroom setting. There are constraints of experimentation in such education settings where the learning goals of students and their experience with the study sessions are also important considerations. DBR allows us to identify failures quickly, make changes to the design and evaluate the resulting system. We can evaluate the performance of the trained crowd across multiple deployments without comparing it against baselines.

Iteration is an essential part of the DBR process and we use reflection assignments after each study session to capture the crowd's feedback on three experiences: 1) coordination with experts; 2) teamwork; and 3) overall difficulty and enjoyability of the session. We gather insights into experts' experience through post-session interviews. We iterate based on these perspectives and observations to finalize how tasks can be assigned and carried out by crowd teams. Feedback from practice and study sessions is useful for optimizing team structure and communication between the investigator and the crowd. We highlight the iterations for each of the challenges in Section 4.4. We then evaluated the ORS model through a case study of the OSINT lab course, a semester-long university course taught by the third author.

Our overall approach is also heavily inspired by Agile Research Studios [137]. This work explored a classroom-based approach to collaborative, real-world HCI research with students. We adapt this approach for the context of OSINT investigations.

4.2.1 Data collection and analysis. As a part of the OSINT lab course, we conducted five study sessions where the trained crowd worked with investigators to augment their ongoing investigations. Data was not collected for the first, practice session, a pilot study aimed at developing coordination between the crowd and expert and iterating on the script for the interview study. The first study session was conducted during week 7 of the semester, and the final one during week 15, with the other deployments roughly every two weeks.

We collected qualitative and quantitative data from the study sessions to understand the experts' and crowd's attitudes and performance. The different data sources were: 1) observations during the study sessions by the first and third authors; 2) reflection surveys submitted by the crowd; 3) spreadsheets containing crowd submissions and expert feedback; 4) semi-structured interviews with each expert after sessions; and 5) separate focus group interviews with members of the crowd. We present the reflection survey form details, a snapshot of a spreadsheet containing crowd submissions and corresponding expert feedback, and interview scripts for experts and the crowd in Appendix A. We had pre-session meetings with 3 of the 5 investigators to decide on investigation topics and break down the investigation into tasks. We worked asynchronously through email with the other two experts.

We conducted a total of 10 interviews, consisting of 5 semi-structured interviews with investigators and 5 focus group interviews with 14 students. The first author transcribed all recorded interviews. In collaboration with the rest of the research team, the first author conducted a deductive thematic analysis [41] of the transcripts. The themes were informed by prior work (Section 3) and aligned with our interview guide. They were aimed at capturing the user experience and evaluating

our design arguments (Section 4.4). We extended the thematic analysis to include the crowd's reflection survey responses.

For experts, we identified the following themes: 1) the usefulness and effectiveness of OSINT macrotasks; 2) criteria for successful completion of tasks; 3) planning versus actual activities during sessions; 4) interaction and communication with the crowd; 5) assessment of submitted information in terms of quality and quantity; 6) comparison of trained crowd's effectiveness with the general crowd; 7) crowd's self-evaluation of submissions; 8) suggestions for improvements; and 9) willingness to work with the crowd again. For crowdworkers we identified the following themes: 1) team formation and evolution of teamwork; 2) usefulness of practice sessions; 3) perceived change of performance over time; 4) tools and techniques used in tasks; 5) challenges faced with tasks; 6) positive and negative aspects of sessions; 7) use of self-evaluation; and 8) suggestions for improvement. We took multiple steps to analyze the transcripts and organize the findings. First, we coded each transcript based on the established themes. After this, we engaged in detailed discussions to refine these themes. We sought insights into each design argument and goal, deepening our understanding of the collaboration process. We compared the similarities and differences across codes and themes to form higher-level themes. These themes helped organize our results and are finally presented in our findings (Section 5).

4.2.2 Participants recruitment and demographics. This study was approved by our university's IRB. The first set of participants were students in the course, who were junior and senior students in the Computer Science departments of two universities (U1 and U2). There were a total of 30 students in the course, 20 from U1 and 10 from U2. The first author recruited the students during an in-class lecture and their participation was voluntary. The consenting participants received $20 after completing a post-course completion interview.

18 out of 20 students from U1 consented to data collection. 2 out of those 18 students identified as female, while others identified as male. 10 out of those 18 students, all of them identifying as male, participated in focus group interviews conducted by the first author. 8 out of the 10 students at U2 consented to the study and 4 students participated in a focus group interview. All students from U2 identified as male. We refer to the 14 crowdworkers who participated in the focus group interviews as CW1 – CW14.

Our study consisted of 6 investigators as expert participants. The first pilot session, which we omitted from our data analysis, was conducted by a journalist recruited through an ad on Upwork and was compensated with $100. We estimated a total time commitment of 2 hours and 15 minutes for the investigators. There were three phases: a 30-minute pre-session training meeting, a 75-minute study session, and a 30-minute post-session interview. We decided on a rate of $45/hour based on the hourly rates of freelance journalists on Upwork and the compensation reported in previous studies [36, 95, 126] Details about the next set of 5 experts (referenced as E1 – E5; we use the terms "expert" and "investigator" interchangeably) who led the study sessions, including the topics and tasks are presented in Table 1. These investigators were invited to participate in the study through email and social media advertisements. All of the expert participants were based in North America and identified as male, reflecting the demographic trends in OSINT described above. There was a wide range of experts' professional experience, from 3–5 years to 11+ years. 2 out of the 5 investigators did not have previous experience with crowdsourcing. Only one of the investigators accepted the offered compensation of $100 for their participation.

4.2.3 Limitations. We did not set explicit learning goals and tests to measure performance of the crowd for each task/deployment. Instead, our evaluation is based on expert feedback and the session's utility for augmenting ongoing OSINT investigations. To reduce the potential for biased results in crowd feedback, as they are part of a for-credit course, student's participation in the

Table 1. Details about experts, topics, and tasks for our study sessions

Session / Participant Identifier	Expert profession	Years of Experience	Topic	Tasks Involved
1 / E1	Fact-checker	5+ years	Verify origins and content of video about Dr. Anthony Fauci	Verification: Source analysis, fact-checking, image analysis
2 / E2	Law enforcement officer	11+ years	Collect social media images and videos for a particular mountain range within certain dates that contain people or vehicles	Discovery; Verification: Geolocation, source analysis
3 / E3	Human rights investigator	3–5 years	Identify whereabouts of leader in Ukraine prior to their death; Collect evidence from social media of European immigrants being used as political pawns	Discovery; Verification: Geolocation, source analysis, image analysis
4 / E4	Investigative journalist	3–5 years	Identify discourse around anti-vaccine protests occurring throughout Europe as well as groups involved	Discovery; Verification: Geolocation, source analysis
5 / E5	Local news journalist	3–5 years	Identify local U.S. politician's public appearances; Report on the discourse around deer hunting within local city limits	Discovery; Verification: fact-checking, geolocation

study was completely voluntary, and interviews were conducted after the conclusion of the course. Another constraint was the organization of single sessions for each domain, which does not fully explore the potential of different investigations within each domain. In future work, replicating sessions for each domain could offer more varied insights and contribute to a more robust and comprehensive understanding of the crowdsourcing framework.

Another limitation of our study was that most of the crowd participants and all of the expert participants were male. Despite efforts by the instructor and universities to recruit more women, these predominantly male participant demographics are representative of broader gender gaps identified both within the university departments and the broader OSINT [37] and Computer Science [122] fields. This limitation may introduce gender bias into the research findings, as the perspectives and experiences of female participants are not represented. Future work should focus on approaches to broaden gender diversity in the OSINT field and courses such as this one.

4.3 OSINT Lab Course

The OSINT lab course was taught simultaneously with students from two universities in the Fall of 2021. Students in the program learned about the OSINT lab course through word of mouth, recruitment emails, and course catalogs. They received credit for completing the course but were not required to stay in the course. The collaboration with investigators from multiple domains was designed to provide an authentic learning experience.

Participants spent the first half of a semester learning about the entire OSINT analysis process. The training covered technical and ethical aspects of OSINT investigations. The first two weeks were spent on introducing the field of OSINT with examples of impactful investigations and its key elements of a culture of transparency, an adversarial mindset, and collaboration among individuals. Each subsequent week focused on one of the identified OSINT macrotasks. The third author demonstrated tools and techniques associated with that task. Details about these skills are provided in Table 2. All the authors participated in designing and implementing practice sessions where students formed teams and solved demo tasks that required the application of relevant OSINT skills. Based on previous work demonstrating the benefit of goal setting in training [108], we asked each team to submit at least 3 high-quality submissions during practice sessions.

The crowd conducted real-world investigations based on expert prompts during the second half of the semester. After recruiting each expert, we scheduled a 30-minute pre-session training meeting
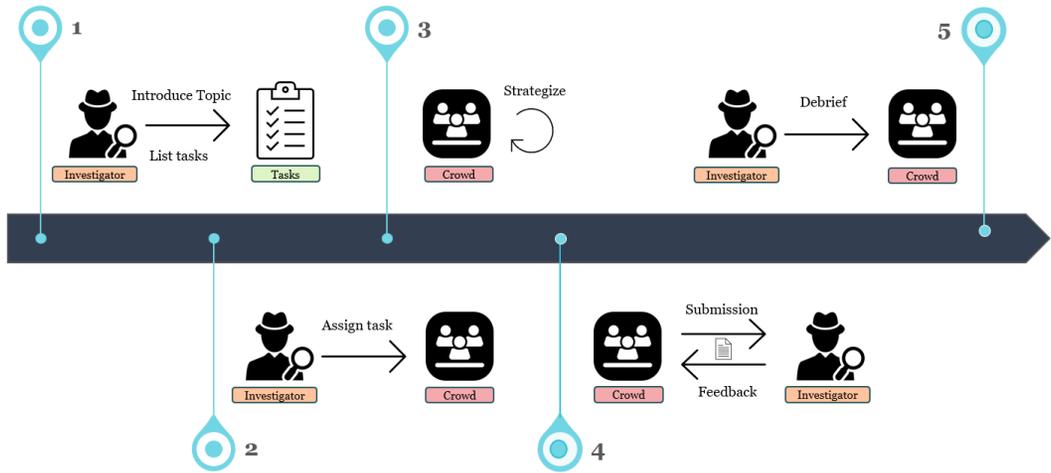
Fig. 1. Phases of our study. ORS connects experts with a trained crowd to perform real-world OSINT investigations described in Section 4.3. The study sessions facilitate synchronous collaboration through 4 chronological phases: (1) The investigator presents the investigation topic and lists tasks that relate to the 5 OSINT macrotasks (presented in Table 2). (2) The authors collaborate with the investigator to assign tasks to teams and provide links for submission. (3) The crowd strategizes the division of work within their teams. (4) The crowd conducts investigations for the rest of the session, submitting their findings through a tailored Google Form. Responses are aggregated in a spreadsheet, allowing real-time expert feedback and guidance. (5) The expert debriefs the crowd by sharing their insights and high-level feedback.

a few days before the session. First, we presented a short slide deck that gave an overview of the study and our expectations for the expert and described the five crowd macrotasks in detail. Second, the expert brainstormed an appropriate investigation topic and we discussed how to decompose it and map it onto one or more macrotasks. Third, we answered any questions the expert had about the forthcoming session.

The investigations varied in terms of topics and contexts as they were based on real ongoing investigations of experts. One common thread was the use of publicly available information (i.e., OSINT), predominantly social media content, and involved a combination of discovery and verification tasks. OSINT investigations ranged from fact-checking videos, documenting human rights violations, finding traces of homicide suspects, and investigating the whereabouts of public servants.

There were five investigation sessions. Each session was 75 minutes long and conducted during the class. Experts joined remotely through a video call and the crowd could either join remotely or be co-located in a classroom. The student crowd consisted of eight teams of three to four members for every session. Each session had four phases. First, the investigator gave a short presentation on what they wanted the crowd to investigate. Second, the authors worked with the expert to assign teams to different subtasks and pointed to resources for submission. Third, the crowd spent the next five minutes strategizing the division of labor within each team. Fourth, the crowd investigated for an hour. Teams made submissions to a Google Form tailored for each session. A spreadsheet aggregated these responses and experts reviewed them in real-time to provide feedback and guidance to the crowd. Finally, at the conclusion, the expert engaged in a debriefing process,

Table 2. Discovery and verification tasks for OSINT investigations

Task	Skills	Tools and techniques
Discovery [15]	Gather hashtags, identify interesting social media accounts, and tailor search for global and local events [15]	Advanced searches on social media platforms based on keywords and location; Follow digital trail to identify key actors and their associates; CrowdTangle [62]; Archive information through archive.is[8]
Verification: Source analysis [21]	Look at provenance of content; research the person, organization (or bot) behind posts [116]	Reverse image search through engines like Yandex[22], TinEye [20], Google[13]; Search on social media platforms including closed ones like Reddit and Telegram; Hoaxy[115], BotSentinel[9] to analyze sources
Verification: Fact-checking [60]	Find prior fact-checks or conduct original research to fact-check a text-based claim [95]	Look at previously tagged fact-checks [73]; Go through official documents
Verification: Image analysis [60]	Identify visual clues, metadata, potential manipulation/editing [21]	Use Exif metadata[12]; Reverse image search; Look at shadows, weather during the time at the location
Verification: Image geolocation [116]	Identify where on Earth a photo or video was taken [74]	Reverse image search; Navigate through satellite imagery (Google earth and Google street view)

discussing what they learned and giving some high-level feedback to the crowd. An overview of the structure of the investigation sessions and the roles of experts and crowds is presented in Figure 1.

4.4 Design Arguments

To scale and speed up OSINT investigations carried out by journalists, fact-checkers, law enforcement officers and human rights investigators, our work sought to develop a collaborative crowdsourcing framework that: (a) trained the crowd in OSINT analysis skills both in terms of tasks and the ethical aspects; (b) divided real-life expert investigations into constituent discovery and verification tasks; and (c) enabled synchronous collaboration between experts and a trained crowd. OSINT Research Studios (ORS) provides a sociotechnical approach that orchestrates training and deployment of a crowd for performing efficient and ethical investigations involving OSINT.

ORS addresses the design challenges for crowdsourcing complex OSINT tasks in the following ways:

4.4.1 Delegation of complex OSINT tasks. To cater to the needs of analysis of open source information across multiple domains, ORS uses the OSINT framework [133] to divide larger investigations into tasks. Prior work [36, 133] divided OSINT investigations into four steps through the OSINT cycle: (1) discover relevant information; (2) verify its provenance and look at the veracity of claims around that content; (3) archive them for future reference; and (4) report on the findings of the investigation. ORS dives deeper and further divides the discovery and verification phases into five parallelizable macrotasks. Each of these tasks is associated with distinct OSINT tools and techniques. These tasks can contribute to rapid, focused investigations as they do not require extensive background knowledge from the crowd and generate results quickly. We provide details about these tasks in Table 2. Work on journalistic processes of information discovery and verification by Brands et al. [40] and practitioner resources from First Draft News [15, 21], Bellingcat [74], and

Poynter [116] guide the design of these tasks. Micallef et al. [95] include most of these resources in their review of computational tools used by fact-checkers.

The five independent tasks serve two purposes in ORS: they (1) form the basis for training the crowd; and (2) break down expert investigations into one or more tasks. These tasks are designed to generate rapid, focused output that can augment ongoing expert investigations. As the crowd is trained to perform these tasks efficiently, they should be able to complete them when prompted by the experts. Following the ethos of the field of OSINT, the tasks discourage over-reliance on custom tools, and training is based on techniques for working directly with data sources. The contexts vary widely based on the domain and the particular investigation, but are circumscribed by OSINT analysis. Therefore, these tasks map the expert investigations to the crowd's skill set and provide a way to garner the required support for investigations.

Iteration: We iterated on making the tasks easy for experts to relate to and use. Initially, we had two discovery tasks for local and global events. However, we realized that the crowd applied the same tools and techniques to both, and experts did not need to specify the scope beforehand. They could delegate the task as discovery, and the available information would determine the scope.

4.4.2   Safety, privacy, and ethical considerations. ORS elevates expert supervision and provides the crowd with a solid understanding of the ethical considerations of investigations. ORS operationalizes the ethos of OSINT, including prioritizing transparency and avoiding the use of subterfuge, in the following ways:

- Prioritizing transparency: ORS encourages each piece of information submitted to be archived to counter the ephemeral nature of the information. Source identification and documenting all relevant details are also part of the process.
- Avoiding the use of subterfuge: ORS prohibits hacking and, given a classroom setting, the honor code is applied to curb any such attempts.
- Limiting investigations to passive reconnaissance: ORS promotes the use of sock-puppet accounts that help to anonymize the identity of the investigator. This is to ensure that ongoing investigations are not interfered with and investigators are not putting themselves at risk, in case organized crime groups apply countermeasures. Crowdworkers are provided with access to virtual machines and VPNs to access unsecured information without jeopardizing their computing systems.
- Fostering accountability: Having a dedicated, accountable crowd of known participants (i.e., the students) also helps preserve privacy and reduce the risks of leaks and sabotage.

As a part of the training process, we ensured that crowdworkers were well-versed in the guidelines and ethical considerations of OSINT investigations. We used real-life scenarios and case studies to help the crowd understand the potential risks and challenges. We also informed the crowd about the potential harm to individuals and communities and the responsibility they held in conducting investigations ethically.

ORS enables experts to stay in control of the investigation. We helped experts outline the scope, objectives, and boundaries of the OSINT investigation during the planning phase by providing examples of previous investigations and explaining the scope of the tasks. We also encouraged experts to balance providing context and disclosing sensitive information to protect the integrity of their investigation.

During each investigation, we established mechanisms for feedback and accountability. We encourage crowdworkers to report any issues or concerns and plan to address them promptly. We ensured that experienced investigators actively guided and oversaw the investigation. Investigators utilized their expertise to validate information, make informed decisions, and navigate complex

situations. They provided feedback to crowdworkers to ensure conclusions and findings were based on accurate, reliable, and verified information. We also conducted a post-investigation review with each expert to evaluate the process and outcomes.

Iteration: We iterated on the crowd submission forms to enforce the requirement of archiving any information put on the spreadsheet for transparency. A required field was added to provide archived links for investigated social media content. We also added a notes field to promote further analysis from the crowd and highlight interesting findings for the expert.

4.4.3 Organizational overheads of synchronous collaboration. ORS employs piggyback prototyping [68] approach to develop a workflow for synchronous collaboration between the expert leading the session and the crowd. The goal was to make the crowd submissions accessible to the expert and present all relevant details in one place. Google Forms are used to specify the requirements for the information requested. These submissions were collated in a Google Sheet. This sheet was accessible to all the participants of the session including the experts, the crowd, and the researchers. Experts provided qualitative feedback for the submissions on the sheet. This piggybacking approach proved beneficial for three reasons: (1) the crowd and experts had previous experience with the interfaces [40], (2) the established applications are robust, and (3) data can be easily exported or read from external applications for further processing.

Another major issue in such crowdsourced settings is duplication of effort [125]. ORS tries to circumvent this issue through these two aspects of team dynamics. First, the crowd works in teams, and each team is assigned an exclusive task. The division of labor can be on the basis of social media platforms, location, subtopics, and/or OSINT subtasks. For example, if one team is searching for information in France, the other teams are looking at countries other than France. Second, each team is responsible for submitting unique entries. This requires collaboration within the team and being aware of what other team members are working on.

Iteration: We initially created a team leader role who was responsible for dividing up and delegating tasks as well as communicating directly with the expert. However, the team leader requirement was relaxed based on crowd feedback that teams did not find that role useful. Without the formal leader role, we dedicated 5 minutes at the start of each session for teams to discuss their strategy.

4.4.4 Maintaining the Quality of Investigations. Previous studies have found that feedback from experts and self-evaluation both improve the quality of crowdwork [55]. ORS seeks to leverage these feedback benefits as follows. First, experts are asked to look through the submissions during the session and provide qualitative feedback that can help crowdworkers improve their performance. Second, the crowd rates their own submissions based on three measures. For a piece of content to fulfill the requirements, it must be specific, verifiable, and relevant.

Experts are also encouraged to communicate verbally with the crowd and make interventions to direct the crowd in productive directions. For example, if the expert finds the information coming in to be of a certain type which is not helpful, they can qualify their requirements further and help the crowd generate more appropriate information. Experts can also move teams from one task to another based on progress across tasks at any time during the session.

Iteration: Based on feedback from the crowd, we solicited more involvement from experts. We asked experts to provide feedback quickly. We also prompted experts to "think aloud" (in the virtual meeting) any information that they found helpful and discuss any high-level feedback that could help the crowd improve further during their debrief at the end of the session.

5 FINDINGS

5.1 Role of OSINT Macrotasks in the Crowdsourcing Setup

5.1.1 Experts found OSINT macrotasks to be relevant. Experts mentioned that they spend a lot of time on the 5 OSINT macrotasks in their typical investigations. We asked experts to rate the tasks on a scale of 1 (not useful) to 3 (very useful) based on how relevant they are to their investigations. Overall the tasks received an average rating of 2.76 out of 3. The verification tasks were rated more favorably compared to the discovery task.

Discovery. The discovery task received an average rating of 2.4 out of 3. E2 mentioned that discovery is their first step in any investigation and any help there would be beneficial. E5 stated that the task is important in news discovery. E1 and E3 found it to be less relevant as the discovery task can become overwhelming and collecting too much unverifiable information is counterproductive. Experts also thought that the task was suitable for crowds. E5 thought that the crowd has "far more sort of computer or digital literacy" than he does to perform well-directed advanced searches and mine information from a wide range of social media platforms.

Verification. Verification tasks had an aggregate rating of 2.85 out of 3. Each sub-task was found to be relevant to the verification process that is integral to OSINT investigations and the work practices of the investigators.

Source analysis enabled experts to dig deeper into interesting leads and possibly reach out to them for further investigation. E1 thought that the crowd would be especially strong in this task, and he could imagine delegating this work of finding background information on creators of social media posts as a part of his workflow.

Fact-checking was thought to be a form of deep research and termed "fairly straightforward." E1 and E4 thought that this task could be performed well by a group of trained people. E1 mentioned that it is an essential step and he would "still feel the need to do it [himself]." However, assistance from the crowd would be helpful to speed up the process and potentially find evidence that might get missed.

Experts found image analysis to be a hard task, especially detecting manipulation, but thought it would be useful to have members of the crowd look at it independently and come to a conclusion.

Geolocation was a part of all the investigators' work practices and received a maximum usefulness rating of 3. Experts explained the importance of the task in identifying recycled information. For example, it is the next step for most evidence in human rights investigations, placing the content in a particular geographical location. E4 had previous experience with crowdsourcing geolocation tasks and thought it worked very well.

5.1.2 Crowd was confident about applying OSINT skills that they were trained on. Crowdworkers felt that they understood the requirements clearly and could work towards the solution. CW6 recalled, "The submissions that I gave kind of knew exactly like what kind of information they're looking for and how much detail they wanted it to like go into. And what kind of information wouldn't be too useful to submit to the Google Doc." CW1 mentioned that the practice sessions were useful to learn the skills, especially geolocation. Practice using tools like reverse image search, Google Street View, and looking up the language of signs were all useful during expert sessions. He felt it was "like a game but [investigators] use that in real life."

The crowd had a positive experience being able to apply the learned skills. They appreciated how the tools and techniques learned during training could be applied to impactful real-life investigations. For example, CW2 mentioned, "...taking all of the techniques that we learned in class and going in seeing people who actually use those on their day to day it was really interesting."

5.2 Session Planning

5.2.1 Investigations were decomposed into OSINT macrotasks. We worked with experts to break down their ongoing investigations into prompts that each apply one of these five tasks described in Table 2: discovery, source analysis, image analysis, fact-checking, and geolocation. Experts came up with the prompts for students after we described the scope, tools and techniques involved and examples of previous practice and expert investigations relating to each task. In all 5 study sessions, experts chose a mix of discovery and verification tasks for the crowd as described in Table 1. The tasks were framed as questions that looked for detailed answers and supporting links and documents. For example, in the session with E1, five different questions were asked based on a viral Instagram video clip of Dr. Fauci speaking at an event:

- Is this a legitimate video, not one doctored to make it look like Dr. Fauci speaking?
- What is the context of the video — where/when/why did Dr. Fauci speak?
- What is the context of Dr. Fauci's remarks? He says something like, "you take an infectious agent and you introduce it into a population," making it seem like he is behind the HIV/AIDS epidemic, but what was he addressing with his remarks? What are some of the other related facts for the epidemic mentioned?
- What is the background of the Instagram user?
- Have any news articles or fact-checks been published about this particular video, or about Dr. Fauci's remarks in the video? Where else has this video been used?

Experts divided larger investigations into tasks for crowd teams across several dimensions, including social media platforms, location, subtopics, and OSINT subtasks. For example, in session 4, teams were looking into anti-vaccine protests across countries in Europe, and each team had a specific country to look at. This form of task assignment helped in the non-duplication of effort across teams. The investigator assigned teams to the scoped tasks randomly, as the crowdworkers were assumed to have the same skill level.

5.2.2 Experts sifted through information around the topic before session. Experts found planning before the session essential to the process. Investigators mentioned that they spent substantial time trying to acclimatize themselves to the information surrounding the discourse of the topics. For example, E3 said, "I gave [the crowd] a bunch of information on the location of the first test" and thought "that was very helpful for them that allowed them to provide geospatial data on trails." Investigators thought sifting through information around the topic could help ensure the right level of difficulty for the sessions. Discussing how his preparation would change if he were to do it again, E1 said, "I would definitely do my own research" ahead of time. The preparation helped experts respond faster and more reliably to the submissions.

5.2.3 Experts had different foci based on quantity and quality of submissions.

Focus on quantity. E2 and E4 looked to gather a bunch of information that could then serve as a starting point for their investigation and future reporting. For example, E4 prioritized quantity and mentioned that his organization's investigation was at a stage where they would be more concerned with quantity over quality. He reasoned that for "taking this data and turning it into a project, I think we want to err on the side of letting us decide [later] what is useful or not." He wanted the crowd to submit all relevant and interesting content without second-guessing.

Focus on quality. E1, E3 and E5 looked for verified information and had a focus on the quality of submissions. They wanted additional relevant information for the discovered content. This information would come from follow-up verification tasks that identify visual elements like buildings, cars, flags, and groups involved.

5.3 Collaboration Within Crowd Teams

5.3.1 Team formation. Team formation was initially based on proximity in terms of seating. Teams generally stuck together once they were formed during the first study session. There were very few changes to most of the teams across the sessions, typically caused by members being absent or leaving an existing team.

5.3.2 Division of work within teams. Within their teams, the crowdworkers divided up the work keeping in mind the requirements set by the investigators. The team members worked individually after choosing non-overlapping search spaces for online information. Sometimes each member had a preferred social media platform to investigate and they tried to divide the work equitably. Most of the coordination was to make sure that there were no duplicate posts from the same team. CW12 explained, "if one of us found something, we let the other person know so we don't find the same thing twice." They also got a sense of which social media platforms had more relevant content and pivoted their searches based on content discovered by other team members.

Only two teams reported having particular strategies in place to divide up the work which did not change over time. CW12 described this common pattern for the division of work, paraphrased as follows. First, get together and divide up the work individually, for example take up different social media platforms like Twitter, Reddit, Facebook, etc. Second, double check posts for duplicate when they find something relevant, especially if multiple members were working on the same platform. Third, add relevant verification details. Fourth, repeat this for all sessions, as they had the same team throughout. Other teams had members working individually on tasks without a particular strategy, but following a similar workflow.

5.3.3 Benefits of teamwork. The major benefit of teamwork was observed when getting relevant information for the prompts was difficult. For example CW8 mentioned that he relied on his team when he felt it was "just kind of daunting" to take up one platform and "find all the information" individually. As discovery got harder, their team wanted to move from one platform to another as a team. The members of the team bounced off ideas with each other when they hit a dead end with one platform.

Team members discussed the feedback that they received from the investigators. They generally found the feedback to be helpful and wanted all the members to be aware of the pointers or corrections provided. The teams communicated verbally and talked about the posts they were working on to ensure that others knew. We also observed communication within teams when they got together to figure out the next steps after getting stuck with their individual investigations. Some of the crowdworkers wished for better communication as a team and but thought that they could not successfully enforce any organizational structure. No leadership roles were established within the crowd teams.

5.4 Collaboration Between Experts and the Crowd

5.4.1 Collaboration through expert feedback. The primary mode of collaboration between experts and the crowd involved the submission of tasks by the crowd and the subsequent feedback provided by the experts. Across all the sessions, investigators provided substantial feedback on the spreadsheets containing crowd submissions. Out of 196 submissions across the sessions, 112 submissions received long-form feedback, generally 1–2 sentences. E2 mentioned that they did not leave any comments if they thought that the submissions "didn't have any relevant information that would help us one way or the other for sure." The feedback was used for two main purposes: getting the crowd to dig deeper into the discovered content and course-correcting the investigations.

Feedback asking crowd to dig deeper. Experts pointed out specific parts of submissions and asked them follow-up questions to bolster the evidence. Some examples from written feedback on spreadsheets include, "Is there any quick verification you can do to confirm that this is actually Brussels or at least Belgium?", "22 minutes of driving video. This is the kind of stuff that could be really helpful. How do we know when the video was shot? Spreadsheet says 10/9 but youtube says it was posted on 10/15." In one case, a crowd team submitted another social media post made by the same poster and indicated that the source had a history of spreading unverified information. The investigator asked the team to find more such historical posts that could be potential misinformation.

Feedback asking crowd to course-correct. Experts reiterated the requirements if they found submitted information to be less relevant. Feedback on the sheet clarified the date ranges, helped submissions to focus on primary sources and called out the lack of details for the investigations. Some examples include, "Posted on the Oct 5th but the page says the footage was shot on 9/25. Outside our range", "Good. But best if we can find non-mass media sources from individuals on the ground", "Hard to tell if this is a covid protest– getting more footage from the protest on this date would help to verify."

5.4.2 Experts made effective interventions. Experts could make announcements during the session to influence the crowd as a whole. The crowd responded well to high level feedback and tailored their investigation to the needs of the expert. E3 asked for more details about the source and location of the media submissions and the crowd provided that. E5 prompted the crowd to discover information about deer killings in and around a city in the US. He found that the crowd took some time, but pivoted based on his feedback of avoiding trophy pictures. He thought that the delay was reasonable and attributed the latency to rabbit holes that investigators might have gotten into and the time needed to reformulate the searches.

5.4.3 Crowdworkers found feedback from the investigators to be helpful. Feedback helped the crowd in navigating the information space around the topic. The crowdworkers felt that it was important to receive quick feedback from experts to improve their performance. CW6 recalled session 4 and said, "[E4] gave feedback right away, which is, which is a big deal for us because we're able to understand Okay, this is exactly what [the expert] is looking for . . . now we can kind of start tailoring our searches to that." Feedback from the experts helped the crowd improve their subsequent submissions. Quick feedback motivated the crowd to stay engaged and keep looking for new information. On the other hand, slow feedback from some of the experts made CW10 feel that his submissions were less useful.

5.4.4 Self-evaluation.

Experts speculated about the usefulness of ORS. Investigators did not use the self-evaluation scores from the crowd during the sessions. However, they thought it could be useful for sifting through information like in the case of geolocation, where the ratings could reflect their confidence in the result. One of the experts mentioned that having them can benefit longer running investigations with a huge number of submissions. E3 and E5 mentioned that they found self-evaluation ratings from students to be "fairly accurate." Likewise, there tended to be positive expert feedback on submissions with high self-evaluation scores, whereas submissions that did not meet requirements as pointed out by experts had low self-evaluation scores in the relevance and verifiability measures. Others like E4, who had a hard time figuring out the self-evaluation part of the sheet, mentioned that those measures were not clearly communicated to him.

It helped the crowd to reflect but did not improve quality of submissions. The crowdworkers had to rate each of their form submissions based on the three metrics of specificity, relevance and verifiability across all the expert sessions. The crowd thought self-evaluations reinforced important aspects of what the expert investigators were looking for into the submissions. CW13 felt that "it was really useful to evaluate ourselves just to kind of keep up the quality in our submissions and make sure the experts that were kind of like looking over our submissions actually got something out of like coming to the sessions." Crowdworkers did not feel that self-evaluations improved their performance, as it was after the submission that they added the scores. But they were able to reflect on their submissions and sometimes admitted that their results were lacking. The average scores on a scale of 0–2 across the three measures were high, and ranged between 1.67 and 1.93 across sessions. Submissions with all 2s ranged from 57.7% to 81.5% of total submissions during sessions. There was no noticeable improvement or significant changes in these scores over time.

5.5 Factors Affecting the Crowd's Performance During the Investigations

Each team was assigned a particular prompt and they submitted information to cater to that. The number of Google Form submissions varied widely across the sessions based on the difficulty and type of task. For example, discovery tasks generated more submissions, whereas verification tasks required more information and research which reduced the quantity. The highest number of submissions (71) were made during session 4, which was about documenting anti-vaccine protests across Europe. Other, more geographically specific and time-bound investigations generated leads ranging from 25 to 38 in number. Expert feedback, applicability of OSINT skills, difficulty and context of tasks, and the topic of investigation all may have influenced their performance.

5.5.1 Applicability of OSINT skills. Some of the crowdworkers felt that their performance was highly dependent on the expert prompt and how well relevant information could be mined using OSINT tools and techniques. CW5 talked specifically about session 3 which he thought was the hardest because information was difficult to find and the task was not suited for their skills. He found sessions to be easier when they featured the use of geolocation and searching on Twitter and Facebook. Another compounding factor that negatively impacted the performance of the crowd was explained by CW6 as the lack of clarity about the requirement: "... if [experts] were a little too broad and the subject matter expert didn't actively kind of talk about what they were looking for, as far as topics for mis or disinformation, it was kind of hard to hit the mark in some instances."

5.5.2 Difficulty of task. The crowd mentioned that the difficulty of tasks impacted productivity during sessions with experts. They wanted the tasks to be challenging but where they could make progress and have submissions to show for the session. CW12 mentioned that his favorite expert session, "...was a good balance between not being too easy or a little too hard, so I thought that was a good like middle ground where we had enough to work with ... there's like meat on the bone to work with." High difficulty significantly contributed to the least favorite sessions for the crowdworkers. This was because the crowd did not feel that they could contribute to the investigations meaningfully and some reached a dead end even before the end of the session. But a few of the crowdworkers enjoyed a challenging thread of investigation and were fueled by competition while trying to gather information on a hard task.

5.5.3 Context for task. Crowdworkers felt more engaged and had a positive experience if they understood the context for the investigations and got how the information they submitted could be used effectively. This also helped them provide more relevant information and dig deeper into interesting pieces of information. CW3 described one scenario where this was not the case: "I was able to find the picture and geo-located it, which was fun, but I just wasn't you know overly sure

of how it was actually helping," so this made the session his least favorite. CW10 found that the expert in the first session kept reiterating that the crowd submissions matched his own findings on the topic; this made him feel that he did not contribute to the investigation. The crowd was motivated by the context around the investigation and needed a clear objective to be specified right at the beginning of the session.

5.5.4 Topic of investigation. CW4 disliked finding COVID vaccine misinformation because it was "kind of boring" and "in the news all the time, I read about [it] all the time." The crowd responded well to topics that had close physical proximity; for example, discovering information about a tornado that hit the town where the crowd's university was located, or verifying claims about a local leader of a nearby city. CW2 wanted the topic to have "...a balance between like it'd be interesting topic for us and, like an important topic to do."

5.6 Reflecting on Our Synchronous Collaborative Crowdsourcing Setup

5.6.1 Collaborative workflow during session.

Information access. The proposed model of working matched how experts themselves organize some of the tasks in terms of the use of Google Docs and Sheets. Investigators were able to access the information seamlessly and were able to stay on top of the information. E5 mentioned, "I liked the ease of the process. You know sort of I had all the information at my fingertips." Talking about the experience of monitoring submissions, E1 thought a spreadsheet was efficient for him to provide quick feedback.

Hybrid setting of sessions. Investigators liked the hybrid setup of getting them to join remotely and work with two groups of colocated crowds. This was explained by E4 when he mentioned the shift to remote work which was exacerbated by the pandemic and how this model gets at that problem. He thought working on investigations is "not the same as doing group activities on zoom. It just doesn't work." This model was helpful for him to collaborate with 30 participants on Zoom and perform OSINT analysis. However, he mentioned that the same level of engagement and work atmosphere that can be achieved by having the investigators and the crowd in the same physical space cannot be emulated in such a hybrid setup.

During hybrid sessions where experts joined remotely, experts spent most of their time going through the submissions and providing feedback. In the only session where the investigator was present in the classroom, he spent the majority of their time helping the students think through the issues while walking around, and the remaining time on evaluating online submissions.

Crowd liked the setting but wanted to learn more from experts. The crowdworkers appreciated the remote involvement of expert investigators and how they were able to collaborate through forms and sheets. The crowd felt they understood the requirements of the ongoing investigation based on the brief of the investigation and how the tasks were carved out. Having more detailed requirements like date ranges and locations on the task helped the crowd to engage more efficiently with the available information. However, some of them wanted lessons and demonstrations from experts. For example, CW5 wished, "If experts could talk a little bit about what tools they would use for some things, just give us a little more insight on how they would solve it if they were you know in our seat." Such discussions could enrich their skill set and possibly help them discover new and efficient methods.

5.6.2 Experts had varied experiences while keeping up with crowd submissions. 4 of the 5 investigators mentioned that they were able to keep up with the information that was coming in and were able

to provide timely feedback. E5 talked about his professional experience in such a role and thought he was able to put his ability to peruse information to use and "didn't need to catch up."

Experts pointed out how submissions picked up pace during the later part of the session as the crowd got a hang of the topic. E4 talked about the difficulty of providing feedback to each submission with a quick turnaround time. He explained this problem, "...when I'm working by myself, I'm working at the speed of me, right, so I've got the link and I'm going to archive it. But if I've got seven groups of people submitting things at the same time. Then suddenly, not only am I working, but I have to work faster because I've got all this data coming in, but then I'm also not in a single thread in my head." This issue was more prominent in session 4 due to the high volume of submissions (71 form submissions). He also shared his thought about the time limit for such deployments. He thought these sessions could be extended to 2 hours without much negative impact. However, for a normal work day which is around 8 hours, the setup would get unsustainably taxing for the expert.

5.6.3 Lack of interaction. Experts mentioned a lack of interaction with the crowd; the teams did not reach out to the experts with any clarifications or concerns. E4 did not have any interaction other than providing feedback on submissions. He thought of this interaction as a trade-off and said "I'm not sure how you would do that and still get done what you get done during the class." Experts felt that the sessions could be more interactive through questions and verbal feedback, but acknowledged the challenges with the size of the crowd and time constraints. E4 thought having an ice-breaker pre-session with at least one member of the team members, possibly the team leader, could be helpful for the crowd to reach out to them.

Some of the crowdworkers also felt the lack of interaction during investigations. They mentioned possible advantages of having the experts join in person as that could add more communication channels and help gather quick verbal feedback.

5.7 Reflections on the Quality and Quantity of Submissions

5.7.1 Experts reflected positively on the efficiency of crowd. Investigators appreciated the contributions of the crowd investigations. E2 described the challenge with their task as "...people are actually probably posting stuff all day long and there are probably thousands of entries every day, you probably be super overwhelmed ...we just don't have that kind of manpower." He mentioned that the results from crowd investigation were important to them and the information would provide "more avenues of investigation because now [they] can potentially go back to some of these users and ask them for more data." Reiterating the lack of manpower and describing how the session can speed up their work, E5 mentioned, "we have some great talented freelancers occasionally at the [local newspaper], but it is mostly just me ...so having you know people who can help in this sort of process, particularly in some of these cases, you know investigative work it's time-consuming and looking through things would be super useful."

E1 speculated about the efficiency of the crowd in terms of speed and accuracy, "...for the most part, [the information submitted] was very strong, particularly how quickly they were able to respond. I would say that they pretty much grasped what I was asking for and provided good answers." E2 added, "The type of work, you guys did in an hour would take us you know, certainly all day with one person doing it, if not longer." The investigators found the information to be good in terms of both quality and quantity. They also mentioned particular pieces of information that they found to be very promising and could lead to breakthroughs in their investigations. E3 recalled two such posts and said, "...that was something that I had been searching for but couldn't find so that was really great." Experts could successfully crowdsource parts of their ongoing investigations to gather rapid, focused results.

5.7.2 Experts found trained crowd more suitable than the general public. Crowdsourced investigations with the general public have been plagued by issues of sabotage, low output, and leaks [29, 98, 121]. E4 talked about the challenges faced with their investigation of the US Capitol insurrection on January 6, 2021 which involved collecting pictures of the riot from mostly Twitter followers of an OSINT organization. He mentioned that they got a lot of "junk links" which was probably to slow them down. They had to be mindful of bad actors who want to "feed garbage." They used virtual machines to access links from unknown social media platforms. E5 mentioned that it might be risky to involve an online crowd, for example, by asking the Twitter followers of the journalist to find out more about a lead. They do not want to tip off competition and are mindful of the sensitive nature of the information that they are trying to present.

E2 and E4 talked about the comparative advantage of such a trained crowd. Compared to the leads from the general public through investigations that seek the same information from the general public, experts thought the current crowd was able to perform significantly better in terms of the quality and quantity of relevant information generated. E2 described it as "...to finish with 30-plus entries and I'm very happy with that because we nearly doubled what we had from the public in a really short span." To be precise, there were 29 submissions during the session compared to 15 leads that the expert received through an open call to the public.

Investigators appreciated the crowd's effort to archive and document the steps of their investigation as it is fundamental to the transparency of OSINT analysis. The archived links helped investigators to have access to the social media posts readily and use it for future reference. The crowd added relevant verification details for the discovered information and added archived links to peripheral information, map coordinates for geolocation, and found original sources for recycled information. Investigators trusted the process and thought they were more confident about the crowd's abilities than they started out with.

E5 thought that working with a crowd like the one during the session reduces the chances of a leak or creating potential misinformation about the subject of the investigation. Regarding the issues of safety and relevance, E4 felt that "with a group of students you know who are not deliberately trying to sabotage the investigation, those concerns are not there."

5.7.3 Experts identified weaknesses in crowd submissions. Some of the expert expectations were not met during the sessions. The investigators reported that only a fraction of the submissions hit the exact target. Characterizing the low-quality submissions, investigators talked about the inconsistency among teams in filling out the required details about the social media links. For example, E3 talks about a particular example, "for the notes and the visual cues section, some people are, I think, correctly saying that this is in front of the Trump hotel, which is super useful. But other people would write things like 'signs' or like, 'a building,' which is less useful."

During interviews, the experts shared issues related to the specificity, relevance, and verifiability of crowd submissions. These issues led to the lower quality of a part of the submissions. E3 pointed out a lack of understanding of the context behind the investigation and limited time during sessions for information that missed the required details. Some information was found to be only "tangentially relevant." Some submissions were news articles that were copied and pasted, without looking at the veracity of those sources.

Investigators reflected on what they could have done differently to avoid this. E5 thought useful clarifications could be provided before starting off the crowd, such as specifying the type of content and social media platforms to be mined. The prompts and the little time that investigators had while introducing the topic were crucial. Interventions were successful, but investigators felt giving the form of content that would be most valuable at the beginning could have improved the quality

of submissions. For example, E2 reflected, "I don't know if I'd mentioned the GoPro cameras or not, but I wish I would have, if I didn't I wish I would have said it."

Some of the experts mentioned that the tasks were not fully completed. E2 said, "I wouldn't say that what we did today is the end of it because we're still asking the public to send this information to but I'm certain what you guys produced today will be helpful." E3 stated that the information collected from their first task would "basically reduce the time that [he] would need in finding other subsequent information." It was therefore hard to complete investigations based on the results of a single 75-minute session.

5.7.4 The crowd perceived that their performance improved over time. Crowdworkers thought that they got faster with their searches and application of tools. CW3 listed the factors of "repetition of like practicing [techniques] over and over again" and "some decent feedback [from experts] where I'd be like, 'Okay, so I need to do like some of this more'" as helpful for improving their performance. Other crowdworkers mentioned factors like better teamwork as a result of getting to know their team members, being more independent and dividing up work efficiently. The crowd had a better understanding of how to discover content on heavily used social media platforms like Twitter, Facebook, Instagram and Reddit as they remained common across sessions.

Crowdworkers mentioned improvement in both the quality and quantity of submissions over the sessions. In terms of quantity, they attributed faster searches to practice with the tools and techniques and increasing ease of navigating social media platforms for discovery. CW7 thought "everything got a lot easier" and improved on all major tasks including archival, discovery through tools like Hoaxy [115] and verification including geolocation. Crowdworkers satisfied the major requirement of not generating duplicate submissions and there were no duplicates after the first session (session 1 had 2 duplicate entries among 38 submissions). CW9 talked about how he got more comfortable with this requirement by checking for duplicates and putting "a different spin on [submission], make sure it was coming from a different angle."

More importantly, the crowdworkers thought that the quality of their submissions improved. CW6 said, "And maybe not so much the quantity, I mean I guess it went up from the beginning, just because things were quicker knowing how to do things but definitely, was able to I felt like I found one post that was like very relevant to the topic every time." Combined with ongoing training and a better understanding of the requirements, the crowd was able to provide relevant details for the submissions as the sessions went on. They talked about making more refined searches and leveraging previously successful strategies for verification.

However, CW8 found the sessions to be repetitive in terms of the tasks and for him "performance over the different expert sessions didn't really change or improve or anything like that."

5.8 Future Engagement

All experts reported that it was an enjoyable experience for them. E3 felt that "...this is really promising and interesting and I think it was fruitful. We did this for an hour and I think it was fun." E5 said, "I mean when [course instructor] said, you know, like five minutes up, I was like, oh wow we're already done, so time flies when you're having fun."

Investigators thought the positive experience of working collaboratively could translate into longer term engagement with such a trained crowd in the future. They talked about how such a crowdsourcing framework can be incorporated into their regular work. E1 thought this could be a way to break away from the solo activity of investigating online information and delegate work to the crowd reliably. E1 thought identifying the crowd's "areas of expertise" could be helpful as those could be leveraged while planning a new investigation. E1 mentioned that for him, "there's always something available, that is not super time-sensitive and those would be the ones that would allow

Table 3. Addressing the challenges faced by investigators for crowdsourced OSINT investigations through ORS.
(Section numbers are provided in parentheses in Expert After column to refer back to the Findings.)

Design Challenge	Expert Before	Design Argument	Expert After
Delegation of diverse, complex OSINT tasks in ways that crowds can meaningfully help	Experts avoid crowdsourcing complex work altogether or delegate very narrow microtasks.	We help experts to delegate complex tasks to trained crowdworkers using the OSINT framework they already know.	Experts found the individual tasks to be relevant to their work practices (5.1.1). Experts got valuable information and leads to augment their ongoing investigations from the crowd (5.7.1).
Safety, privacy, and ethical considerations for investigations performed by crowds	Experts can't delegate work to any online crowd due to privacy and ethical concerns	Experts are in control over the information and benefit from the crowd's training in transparency and privacy. The ethos of the field are imparted to the crowd and upheld during sessions.	Experts had better experience compared to a novice online crowd (5.7.2). They were optimistic about future engagement with the setup (5.8).
Logistical challenges of synchronous collaboration	Experts don't work synchronously with crowds to generate useful leads	We run multiple sessions to bring the crowd and expert together and provide technical support using Google Forms and sheets. The crowd works in teams and submits relevant information without duplication of effort and are driven by quick feedback.	Experts found the information from crowd investigation to be easily accessible (5.6.1). The setup was convenient to provide feedback for the submissions (5.4.1).
Control over the quality and direction of investigation	Experts have limited communication with the crowd and cannot provide feedback to improve crowd performance	Experts provide detailed qualitative feedback to crowd submissions. Experts use interventions to address any unfavorable patterns in the submissions.	Experts influenced the crowd investigation positively and drilled down on interesting leads (5.4.2, 5.7.4). Expert guidance improved the overall performance of the crowd for the tasks (5.4.3).

for me to have the pre-meeting, and then develop questions for the students, so I think it can be replicated." E1 also talked about how the crowd could perform peripheral tasks like looking at the background of the person/bot who made the post while he "might be digging into more of the central questions about the fact check."

Investigators suggested longer and deeper projects as a good fit for the trained crowd, as they were able to perform these rapid and focused investigations. That would allow the crowd to take up investigations with a larger scope and hone their skills further.

6 DISCUSSION

We designed and evaluated OSINT Research Studios (ORS), a sociotechnical framework that enables collaboration between investigators and a trained crowd. Through ORS, we address design challenges in crowdsourced OSINT investigations including the delegation of diverse, complex OSINT tasks; safety, privacy, and ethical considerations; organizational overheads of synchronous collaboration; and maintaining the quality of investigation. Table 3 summarizes how the challenges faced by investigators while performing crowdsourced OSINT investigations are addressed through ORS.

The OSINT lab course was a semester-long deployment of ORS, where the first half involved training a group of 30 undergraduate students. During the latter half, this trained crowd collaborated with professionals, including a journalist, a fact-checker, law enforcement investigator, and a human rights analyst. The crowd performed time-boxed and highly targeted investigations based on prompts from the expert, that alluded to one or more information discovery and verification tasks. Experts said that the results from these investigations were useful for solving parts of their broader investigations, finding new leads for subsequent investigations, and validating some of

their own findings, thereby helping scale up and speed up OSINT investigations. We revisit how the challenges were addressed and identify opportunities for future research.

6.1 Delegation of Diverse and Complex OSINT Tasks

6.1.1 Effectiveness of macrotasks. ORS focused on developing crowd expertise in five macrotasks — discovery, source analysis, image analysis, fact-checking, and geolocation. The tasks helped decompose crowdsourced OSINT investigations and gather results from the crowd, providing a structure that has been lacking in previous investigations [48, 124]. Similar to CrowdForge [84], we found that high-level initial decomposition of complex work by experts is effective in helping crowdworkers complete assigned tasks. ORS contributes to an expanding body of research that demonstrates how crowds can effectively tackle complex tasks, given they possess adequate motivation, support, training, and autonomy [72, 109, 124, 125]. Our exploration of training based on how to approach different tasks without overdependency on tools matched Wang et al. [128]'s results of training based on analytical thinking skills. In both cases, crowdworkers developed domain expertise and applied their knowledge to solve complex tasks. All expert tasks leveraged the crowd's capabilities to discover of information from multiple online platforms and identify provenance and location information required for verification. The crowd's results were viewed as potential leads and valued by experts for their speed and quantity.

6.1.2 Role of training in ORS. In this work, we argue that training can enable crowdworkers to augment investigations carried out by journalists, fact-checkers, human rights investigators, and law enforcement officers. Doroudi et al. [54] mention challenges like the unavailability and unwillingness of experts for training in a crowdsourcing framework. Experts might sometimes lack understanding about how to ensure the successful completion of tasks, which makes it harder for them to train. In the context of OSINT, these challenges are met by the field's unique characteristics. Belghith et al. [36] described OSINT as a community of practice with legitimate peripheral participation [87]. This involves training novice practitioners through low-risk tasks as they grow into the roles of experts in the community. Experts in the community participate in this model and strike a balance between practice and training endeavors. In terms of skills, OSINT has a wide range of tasks and applications and experts tend to be generalists. Experts learn from each other by sharing how they performed challenging investigations. Training is also available through online certification [10, 17] and MOOC programs [16] for skills related to investigation of people, online information and websites [17]. Training involving multiple OSINT skills is time-consuming, but just-in-time training and developing shorter modules that can enable novice crowdworkers to contribute to particular investigations can be an effective alternative, as shown here. With additional available resources and further modification, trained crowds can be employed by experts outside of our classroom study setting.

6.1.3 Extending ORS to other domains. In the current work, investigations involved discovering social media posts and verifying their content. OSINT investigations ranged from fact-checking videos, documenting human rights violations, finding traces of homicide suspects, and investigating the whereabouts of public servants (presented in Table 1). Given the parallelizable nature of the tasks that we chose, this setup can be scaled up to contribute to rapid response scenarios that are crucial to fight misinformation [130] and respond to crisis events [51]. ORS has the potential for seamless adaptation across domains beyond investigations of online information. For instance, it can be employed to facilitate the coordination of physical search-and-rescue operations for missing individuals or animals [6, 132], as well as to evaluate the extent of damage following both natural and man-made disasters [39]. The parallelizable and targeted characteristics combined with the

inherently adversarial nature of investigations also make the ORS framework applicable to the domains of cybersecurity [59] and finance [110, 111].

6.1.4 Extending ORS to other types of crowds. To scale up the ORS framework further, future work can consider different crowds outside of a classroom setting. Organizations like Bellingcat [7], Trace Labs [50], and the Syrian Archive [19] engage thousands of volunteers in their efforts to collect and verify open source information. In order to incorporate individuals possessing relevant skills and diverse backgrounds, volunteering efforts should collaborate closely with such established communities of practice [131] in OSINT. However, there are open challenges involved with engaging workers of different skill levels, motivating them for sustained participation, having the right mechanism for quality control and aggregating results [49, 123].

6.2 Safety, Privacy, and Ethical Considerations

Based on experts' responses, our findings show that the crowd could conduct OSINT macrotasks and contribute safely and meaningfully to experts' investigations. We achieved this by operationalizing the OSINT ethos: prioritizing transparency, avoiding subterfuge, and limiting investigations to passive reconnaissance [36]. Training was essential for the crowd and experts to implement strategies ensuring ethical investigations. Detailed results and descriptions of the steps involved made the investigations transparent and reproducible. The use of virtual machines, sock puppet accounts, and VPNs helped the crowd mitigate any risks of retaliation and safeguard their systems. We found these measures to be essential for investigators' safety and sufficient for the tasks. The student crowd's accountability ensured private investigations, free from information leaks and vigilante behavior.

6.2.1 Safeguarding against harms caused by crowd inaccuracy. One major concern with crowd-sourced investigations is the accuracy of the information generated and its implications for results of an investigation. In our work, the investigations are scoped to tasks meant to generate leads for the experts to follow up on. Experts play an important role in providing the right context that enables the crowd to generate useful results. For example, in study session 2, E2 (law enforcement officer) was working on finding digital traces of a homicide suspect. Instead of revealing the suspect's identity and asking the crowd to research them, E2 tasked the crowd with collecting social media images and videos of a particular mountain range within certain dates that contained any people or vehicles. This ensured that the crowd could augment the investigation without risks of misidentification and leaking sensitive information. Similarly, GroundTruth used a diagram-drawing technique to allow investigators to crowdsource image geolocation tasks without sharing the target photo or video, which might contain confidential details or disturbing imagery [126]. We advocate for pervasive expert oversight, including quality assurance processes where experts review and verify the information gathered by the crowd. Given that crowd submissions may not always meet the requirements set by experts, systems should provide experts the control to make conclusions and findings based on accurate, reliable, and verified information.

Experts must often make critical decisions, especially within sensitive or high-risk situations. They can leverage their experience and judgment to avoid potential pitfalls and ensure the investigation stays on track. Investigators can conduct a comprehensive risk assessment before initiating the crowdsourced investigation to identify potential risks, vulnerabilities, and threats. Researchers and domain experts should engage in scenario-planning to anticipate and prepare for possible challenges and develop strategies to mitigate them. Collaboration with legal authorities can help to clearly define what is permissible and what is not, ensuring alignment with legal and ethical standards.

6.2.2   Protecting crowd investigators and students from harm. Operational Security (OPSEC) in the context of OSINT investigations refers to the process and strategies used to protect sensitive information, ensure personal and organizational security, and maintain the effectiveness and integrity of the investigation. As mentioned in Section 4.4.2, we apply key OPSEC practices in our deployment of ORS to ensure crowdworkers can conduct research without revealing their identity or affiliation. These include using Virtual Private Networks (VPNs), secure communication platforms for messaging and email, and virtual machines. Crowdworkers used sock puppet accounts to limit the exposure of personal and sensitive information on social media platforms.

In our work, we were careful about the appropriateness of investigations in terms of the topic and the general nature of information surrounding it to make it more suitable for the classroom setting. To further broaden the scope of investigations, future work should consider potential risks, such as issues related to secondary trauma and exposure to sensitive content and triggers [48]. Secondary trauma refers to the emotional stress experienced by individuals as a result of witnessing or being exposed to traumatic events indirectly. To address this, it is crucial to provide psychological support and resources for investigators, including access to counseling services and regular debriefings [117, 133]. Training programs designed to help crowdworkers deal with secondary trauma through self-care, establishing boundaries, seeking peer support, and recognizing warning signs can be beneficial in managing the emotional demands of their work and maintaining their mental well-being [17, 48, 133].

When recruiting OSINT crowds in a classroom context, as we did, it is especially important to consider the unique needs of students. Experiential learning is valued in higher education because it provides authentic, real-world learning experiences [85]. However, in investigative or adversarial contexts, it can also expose students to some risk. For example, some universities teach courses where students conduct real-world OSINT investigations in human rights [2] and cybersecurity contexts [101]. This work may require students to engage with disturbing or illegal material and investigate bad actors. But it can also be personally meaningful and societally impactful, and prepares students for successful careers as professionals in these fields. By employing the safeguards above, OSINT students can learn investigative work in a relatively safer environment compared to those requiring direct interaction with persons of interest. Nevertheless, it is important for instructors to communicate these risks and trade-offs to students and to help experts provide appropriate levels of exposure during investigations that match students' developing skill levels.

6.3   Organizational Overheads of Synchronous Collaboration

6.3.1   Session planning. Experts played an active role in breaking down ongoing investigation(s) and came up with prompts that suited the crowd's expertise. Interestingly, the three major factors for the overall crowd experience - topic, difficulty, and context of investigation (section 5.5) - were heavily influenced by planning. Experts were able to guide the crowd better by doing prior research and providing actionable feedback. Based on crowd feedback, session planning can be improved by specifying how crowd contributions fit in with the larger investigation, thereby setting clear expectations about the type and level of details for submissions. Future research can look at systems that conduct surveys among the crowd, enabling them to communicate their motivations to the expert [105]. Based on this feedback, the expert can make appropriate modifications to their tasks.

The crowd collaborated with each expert for 75 minutes. This setup helped scope out tasks related to discovery and quick verification of social media content. Future research can explore longer-running and more in-depth OSINT investigations as seen in investigative journalism and human rights advocacy programs [2, 97]. Such collaboration and training will enable the workers to dive deeper, apply advanced skills, and learn new ones while solving complex OSINT tasks.

Based on prior work showing the benefits of competition in crowdsourcing [36, 125], gamifying the collaborative process can make the sessions more productive.

The skill level of workers was assumed to be the same as they go through the same training and start with no background in OSINT investigations. There were also 8 teams participating in the sessions throughout. This allowed the process of assigning tasks to crowd teams to be random and manual. To scale up collaboration and accommodate crowdworkers of varying skill levels, AI-mediated crowdsourcing shows promise in automating task assignment and skill assessment [49]. Future efforts can focus on efficient task allocation and result aggregation based on expert need [84, 118], particularly for large-scale deployments involving participants from crowd marketplaces like Amazon Mechanical Turk.

6.3.2 Teamwork and communication. The crowd took up tasks in teams of three to four members. Crowdworkers worked individually after dividing up work based on different social media platforms, geographical locations, and OSINT sub-tasks. Communication within teams grew during difficult tasks as they discussed individual findings and feedback from experts to decide on the next steps. Participants felt that there was a lack of interaction between the crowd and the expert due to a hybrid setup (with remote experts collaborating synchronously with co-located crowds distributed across two locations), which, on the other hand, enabled having experts from across the country. Participants acknowledged that increasing interaction also comes with the cost of slowing down investigations. Hybrid intelligence involving crowd-AI interaction [49] can be explored to provide relevant context while maintaining the fast and focused nature of investigations.

6.4 Maintaining the Quality of Investigations

The major form of collaboration between the experts and the crowd was through crowd submissions and expert feedback on them. The feedback helped the crowd direct their efforts in the right direction and improve their performance. Experts suggested, based on their prior experience, that the trained crowd performed better than the working with the general public. Experts wanted to engage with the crowd in the future based on other investigative tasks like source analysis, discovery, and geolocation. ORS also addressed our design goal to limit duplication of effort, which has been an issue across crowdsourcing solutions [36, 124]. Adding flexibility to the expert-led crowdsourcing framework [124, 126], a trained crowd quickly responded to feedback and changed course to meet the requirements on a wider range of OSINT tasks. Based on crowd feedback, training strategies like solving more tasks and setting goals (explored through practice sessions), and expert feedback (explored through expert sessions) were found effective as seen in [55, 138]. However, self-evaluations did not have an impact on our model as compared to prior work [55]. Self-evaluation effectiveness can be improved through automated quality checks by flagging unverified sources and missing archival links. To improve the quality of submissions further, peer review, which is effective as an adversarial training strategy [118], can be implemented to elicit feedback from other teams.

6.4.1 Automating expert feedback. Leveraging insights from expert feedback in the current study, future crowdsourced OSINT investigations could automate specific feedback mechanisms using large language models (LLMs) [43]. Automating specific feedback saves time and effort for both experts and crowdworkers, as the system can provide tailored feedback to crowdworkers [120] while allowing experts to focus on nuanced assessments or more complex tasks. While automation benefits efficiency, a balanced approach is needed to consider experts' unique perspectives [81]. There is a need for combining automated and human feedback to ensure comprehensive evaluation and boost the productivity of OSINT investigations.

7 CONCLUSION

In our work, we supported the need to scale up and speed up OSINT investigations across multiple domains. We addressed the practical challenges of crowdsourcing OSINT investigations through crowd training and synchronous collaboration. The training was based on the technical and ethical aspects of OSINT and contributed to the successful completion of a wide range of tasks. Collaboration was centered around feedback that experts said improved the overall quality of their investigations. Taking a design-based research (DBR) approach, we iteratively designed OSINT Research Studios (ORS), a sociotechnical system that facilitated rapid and focused OSINT investigations. Through the OSINT lab course, we had a semester-long deployment of ORS, including evaluation sessions with investigators from the domains of journalism, fact-checking, law enforcement, and human rights investigation to evaluate the system. Experts found the sessions to be useful and mentioned strengths like speed, safety, high quality and quantity of submissions across tasks, and the crowd's adaptability to feedback. The crowd enjoyed working with experts and successfully applied their OSINT skills. In conclusion, ORS enabled ethical and effective crowdsourced OSINT investigations.

ACKNOWLEDGMENTS

REFERENCES

[1] 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-policymaking/

[2] 2021. *J298 OSINT Seminar — Open Source Investigations.* https://journalism.berkeley.edu/course-section/j298-human-rights-center-seminar-f21/

[3] 2021. WhatsApp can be a black box of misinformation, but Maldita may have opened a window. https://www.poynter.org/fact-checking/2021/whatsapp-can-be-a-black-box-of-misinformation-but-maldita-may-have-opened-a-window/

[4] 2022. *Hunchly - OSINT Software for Cybersecurity, Law Enforcement, Journalists, Private Investigators, and more.* https://www.hunch.ly/

[5] 2022. Russia is losing so much equipment in Ukraine that weapons monitors can't keep up. https://www.independent.co.uk/news/world/europe/russia-ukraine-military-equipment-losses-b2049613.html Section: News.

[6] 2022. TraceLabs Twitter Account. https://twitter.com/tracelabs/status/1558831625986777088

[7] 2023. *About Bellingcat.* https://www.bellingcat.com/about/

[8] 2023. archive.is. https://archive.is/

[9] 2023. *Bot Sentinel - Dashboard.* https://botsentinel.com/

[10] 2023. *Certified in Open Source Intelligence (C|OSINT) from McAfee Institute | NICCS.* https://niccs.cisa.gov/education-training/catalog/mcafee-institute/certified-open-source-intelligence-cosint

[11] 2023. Check (@checkdesk) / Twitter. https://twitter.com/checkdesk

[12] 2023. *EXIF Data Viewer.* https://exifdata.com/

[13] 2023. Google Images. https://images.google.com/

[14] 2023. Join us in pushing back on misinformation. https://our.news/

[15] 2023. Newsgathering and Monitoring on the Social Web. https://firstdraftnews.org:443/long-form-article/newsgathering-and-monitoring-on-the-social-web/

[16] 2023. *OSINT: Open-Source Intelligence.* https://www.udemy.com/course/osint-open-source-intelligence/

[17] 2023. *Practical Open-Source Intelligence (OSINT) Training — SANS SEC497.* https://www.sans.org/cyber-security-courses/practical-open-source-intelligence/

[18] 2023. Search Party Rules. https://www.tracelabs.org/about/search-party-rules

[19] 2023. Syrian Archive | Syrian Archive. https://syrianarchive.org/

[20] 2023. *TinEye Reverse Image Search.* https://tineye.com/

[21] 2023. *Verifying Online Information.* https://firstdraftnews.org:443/long-form-article/verifying-online-information/

[22] 2023. *Yandex*. https://yandex.com/

[23] 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

[24] Elena Agapie, Jaime Teevan, and Andrés Monroy-Hernández. 2015. Crowdsourcing in the field: A case study using local crowds for event reporting. In *Third AAAI Conference on Human Computation and Crowdsourcing*. Association for the Advancement of Artificial Intelligence, 11. https://www.microsoft.com/en-us/research/publication/crowdsourcing-in-the-field-a-case-study-using-local-crowds-for-event-reporting/

[25] Joelle Alcaidinho, Larry Freil, Taylor Kelly, Kayla Marland, Chunhui Wu, Bradley Wittenbrook, Giancarlo Valentin, and Melody Jackson. 2017. Mobile Collaboration for Human and Canine Police Explosive Detection Teams. 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, 925–933. https://doi.org/10.1145/2998181.2998271

[26] Sultan A. Alharthi, Nicolas James LaLone, Hitesh Nidhi Sharma, Igor Dolgov, and Z O. Toups. 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

[27] 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.

[28] 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.

[29] 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*. PsyArXiv, 1–19. https://doi.org/10.31234/osf.io/57e3q

[30] Adriana Alvarado Garcia, Matthew J. Britton, Dhairya Manish Doshi, Munmun De Choudhury, and Christopher A. Le Dantec. 2021. Data Migrations: Exploring the Use of Social Media Data as Evidence for Human Rights Advocacy. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 268 (jan 2021), 25 pages. https://doi.org/10.1145/3434177

[31] Adriana Alvarado Garcia and Christopher A. Le Dantec. 2018. Quotidian Report: Grassroots Data Practices to Address Public Safety. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 17 (nov 2018), 18 pages. https://doi.org/10.1145/3274286

[32] Amnesty International. 2020. *Syria: 'Nowhere is Safe for Us': Unlawful Attacks and Mass Displacement in North-West Syria*. Technical Report. https://www.amnesty.org/en/documents/document/?indexNumber=MDE24%2f2089%2f2020&language=en

[33] 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 (CSCW '17)*. ACM, Portland Oregon USA, 155–168. https://doi.org/10.1145/2998181.2998294

[34] Anne Aula and Daniel M Russell. 2008. Complex and exploratory web search. In *Information Seeking Support Systems Workshop (ISSS 2008), Chapel Hill, NC, USA*. Citeseer.

[35] Charlie Beckett. 2017. Wikitribune: can crowd-sourced journalism solve the crisis of trust in news? *POLIS: journalism and society at the LSE* (2017).

[36] 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

[37] Bellingcat Investigation Team. 2015. Diversifying OSINT: Women Experts. https://www.bellingcat.com/resources/articles/2015/12/08/women-in-osint-diversifying-the-field/

[38] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 313–322.

[39] 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.

[40] Bert Jan Brands, Todd Graham, and Marcel Broersma. 2018. Social media sourcing practices: How Dutch newspapers use tweets in political news coverage. *Managing democracy in the digital age: Internet regulation, social media use, and online civic engagement* (2018), 159–178.

[41] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[42] Andrea Broughton, Beth Foley, Stefanie Ledermaier, and Annette Cox. 2014. The use of social media in the recruitment process. (2014). 81 (2014).

[43] 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).

[44] Guilherme Carneiro, Miguel Nacenta, Alice Toniolo, Gonzalo Mendez, and Aaron J Quigley. 2019. Deb8: A tool for collaborative analysis of video. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*. 47–58.

[45] Southern Poverty Law Center. 2021. The Road to Jan. 6: A Year of Extremist Mobilization. https://www.splcenter.org/news/2021/12/30/road-jan-6-year-extremist-mobilization

[46] 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

[47] Paul Cobb, Jere Confrey, Andrea DiSessa, Richard Lehrer, and Leona Schauble. 2003. Design experiments in educational research. *Educational researcher* 32, 1 (2003), 9–13.

[48] Josie Cochrane. 2022. Citizen OSINT Analysts: Motivations of Open-Source Intelligence Volunteers.

[49] António Correia, Andrea Grover, Daniel Schneider, Ana Paula Pimentel, Ramon Chaves, Marcos Antonio De Almeida, and Benjamim Fonseca. 2023. Designing for Hybrid Intelligence: A Taxonomy and Survey of Crowd-Machine Interaction. *Applied Sciences* 13, 4 (2023), 2198.

[50] 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

[51] Dharma Dailey and Kate Starbird. 2014. Journalists as Crowdsourcerers: 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

[52] Dharma Dailey and Kate Starbird. 2015. "It's Raining Dispersants": Collective Sensemaking of Complex Information in Crisis Contexts. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work and; Social Computing* (Vancouver, BC, Canada) *(CSCW'15 Companion)*. Association for Computing Machinery, New York, NY, USA, 155–158. https://doi.org/10.1145/2685553.2698995

[53] 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

[54] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. 2016. Toward a Learning Science for Complex Crowdsourcing Tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2623–2634. https://doi.org/10.1145/2858036.2858268

[55] 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

[56] Sam Dubberley, Alexa Koenig, and Daragh Murray (Eds.). 2020. *Digital Witness: Using Open Source Information for Human Rights Investigation, Documentation, and Accountability*. Oxford University Press, Oxford, New York.

[57] Matthew W Easterday, Daniel Rees Lewis, and Elizabeth M Gerber. 2014. Design-based research process: Problems, phases, and applications. Boulder, CO: International Society of the Learning Sciences.

[58] Sheena L. Erete. 2015. Engaging Around Neighborhood Issues: How Online Communication Affects Offline Behavior. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 1590–1601. https://doi.org/10.1145/2675133.2675182

[59] Esteban Borges. 2019. SecurityTrails | OSINT Framework: The Perfect Cybersecurity Intel Gathering Tool. https://securitytrails.com/blog/osint-framework

[60] Giancarlo Fiorella. 2021. *First Steps to Getting Started in Open Source Research*. https://www.bellingcat.com/resources/2021/11/09/first-steps-to-getting-started-in-open-source-research/

[61] 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, New York, NY, USA, 247–256. https://doi.org/10.1145/2207676.2207711

[62] 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).

[63] Claudia Flores-Saviaga, Shangbin Feng, and Saiph Savage. 2022. Datavoidant: An AI System for Addressing Political Data Voids on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–29.

[64] Riccardo Ghioni, Mariarosaria Taddeo, and Luciano Floridi. 2022. Open Source Intelligence and AI: a Systematic Review of the GELSI Literature. *AI & Society* (2022).

[65] 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 (March 2012), 673–682. https://doi.org/10.1016/j.chb.2011.11.014

[66] 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).

[67] Miaomiao Gong, Yuling Sun, and Liang He. 2019. A Social Network Engaged Crowdsourcing Framework for Expert Tasks. In *2019 IEEE 23rd International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 249–254. https://doi.org/10.1109/CSCWD.2019.8791923

[68] Catherine Grevet 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*. 4047–4056.

[69] Pablo Gutierrez and Paul Torpey. 2015. How Digital Detectives Say They Proved Ukraine Attacks Came from Russia. *The Guardian* (2015).

[70] 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

[71] Melissa Hanham and Jaewoo Shin. 2020. Ethics in the Age of OSINT Innocence. *Stanley Center for Peace and Security* (May 2020), 6. https://stanleycenter.org/publications/ethics-osint-innocence/

[72] Alexa M Harris, Diego Gómez-Zará, 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.

[73] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. ClaimBuster: the first-ever end-to-end fact-checking system. *Proceedings of the VLDB Endowment* 10, 12 (Aug. 2017), 1945–1948. https://doi.org/10.14778/3137765.3137815

[74] Annique Mossou Higgins, Ross. 2021. *A Beginner's Guide to Social Media Verification.* https://www.bellingcat.com/resources/2021/11/01/a-beginners-guide-to-social-media-verification/

[75] Eliot Higgins. 2021. *We Are Bellingcat : An Intelligence Agency for the People.* Bloomsbury Publishing, London.

[76] 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.

[77] Y. Linlin Huang, Kate Starbird, Mania Orand, Stephanie A. Stanek, and Heather T. Pedersen. 2015. Connected Through Crisis: Emotional Proximity and the Spread of Misinformation Online. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, New York, NY, USA, 969–980. https://doi.org/10.1145/2675133.2675202

[78] Arthur S. Hulnick. 2002. The Downside of Open Source Intelligence. *International Journal of Intelligence and CounterIntelligence* 15, 4 (Nov. 2002), 565–579. https://doi.org/10.1080/08850600290101767

[79] 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.

[80] Lachlan Kermode, Jan Freyberg, Alican Akturk, Robert Trafford, Denis Kochetkov, Rafael Pardinas, Eyal Weizman, and Julien Cornebise. 2020. Objects of violence: synthetic data for practical ML in human rights investigations. *arXiv:2004.01030 [cs]* (April 2020), 12. http://arxiv.org/abs/2004.01030 arXiv: 2004.01030.

[81] 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.

[82] Yongsung Kim, Darren Gergle, and Haoqi Zhang. 2018. Hit-or-wait: Coordinating opportunistic low-effort contributions to achieve global outcomes in on-the-go crowdsourcing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[83] 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

[84] Aniket Kittur, Boris Smus, Susheel 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

[85] Alice Kolb and David Kolb. 2017. Experiential Learning Theory as a Guide for Experiential Educators in Higher Education. *Experiential Learning & Teaching in Higher Education* 1, 1 (June 2017), 7–44. https://nsuworks.nova.edu/elthe/vol1/iss1/7

[86] Meryl Kornfield. 2021. *The wrong ID: Retired firefighter, comedian and Chuck Norris falsely accused of being Capitol rioters*. Washington Post. https://www.washingtonpost.com/technology/2021/01/16/sleuths-falsely-identify-rioters/

[87] Jean Lave and Etienne Wenger. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge University Press, Cambridge, UK. Google-Books-ID: CAVIOrW3vYAC.

[88] 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

[89] 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.

[90] Hana Matatov, Adina Bechhofer, Lora Aroyo, Ofra Amir, and Mor Naaman. 2018. DejaVu: A System for Journalists to Collaboratively Address Visual Misinformation.

[91] 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

[92] Sean McKeown, David Maxwell, Leif Azzopardi, and William Bradley Glisson. 2014. Investigating people: a qualitative analysis of the search behaviours of open-source intelligence analysts. In *Proceedings of the 5th Information Interaction in Context Symposium (IIiX '14)*. Association for Computing Machinery, New York, NY, USA, 175–184. https://doi.org/10.1145/2637002.2637023

[93] Stephen C. Mercado. 2004. *Sailing the sea of OSINT in the information age. Technical Report*. dataset, American Psychological Association. type. https://doi.org/10.1037/e741272011-005

[94] Panagiotis Metaxas and Samantha T Finn. 2017. The infamous# Pizzagate conspiracy theory: Insight from a Twitter-Trails investigation. (2017).

[95] Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. 2022. True or False: Studying the Work Practices of Professional Fact-Checkers. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–44.

[96] Tanushree Mitra, Clayton J Hutto, and Eric Gilbert. 2015. Comparing person-and process-centric strategies for obtaining quality data on amazon mechanical turk. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1345–1354.

[97] Nina C. Müller and Jenny Wiik. 2021. From Gatekeeper to Gate-opener: Open-Source Spaces in Investigative Journalism. *Journalism Practice* 0, 0 (May 2021), 1–20. https://doi.org/10.1080/17512786.2021.1919543 Publisher: Routledge _eprint: https://doi.org/10.1080/17512786.2021.1919543.

[98] Johnny Nhan, Laura Huey, and Ryan Broll. 2017. Digilantism: An analysis of crowdsourcing and the Boston marathon bombings. *The British journal of criminology* 57, 2 (March 2017), 341–361. https://doi.org/10.1093/bjc/azv118

[99] Jon Noronha, Eric Hysen, Haoqi Zhang, and Krzysztof Z Gajos. 2011. Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 1–12.

[100] Department of Homeland Security. 2010. (U//FOUO//LES) DHS Terrorist Use of Social Networking Facebook Case Study | Public Intelligence. https://publicintelligence.net/ufouoles-dhs-terrorist-use-of-social-networking-facebook-case-study/

[101] Massachusetts Institute of Technology. 2023. Urban Cyber Defense: Cybersecurity Clinic. http://urbancyberdefense.mit.edu/cybersecurityclinic Accessed: 2023-10-02.

[102] David Oleson, Alexander Sorokin, Greg Laughlin, Vaughn Hester, John Le, Christopher R Van Pelt, and Lukas A Biewald. 2013. Evaluating a worker in performing crowd sourced tasks and providing in-task training through programmatically generated test tasks. US Patent 8,554,605.

[103] 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

[104] 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.

[105] 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.

[106] Tjeerd Plomp et al. 2013. Educational design research: An introduction. *Educational design research* (2013), 11–50.

[107] Ronald Poelman, Oytun Akman, Stephan Lukosch, and Pieter Jonker. 2012. As if being there: mediated reality for crime scene investigation. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work (CSCW '12)*. Association for Computing Machinery, New York, NY, USA, 1267–1276. https://doi.org/10.1145/2145204.2145394

[108] Amy Rechkemmer and Ming Yin. 2020. Motivating Novice Crowd Workers through Goal Setting: An Investigation into the Effects on Complex Crowdsourcing Task Training. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 8 (Oct. 2020), 122–131. https://doi.org/10.1609/hcomp.v8i1.7470

[109] 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

[110] Daniel Rouach and Patrice Santi. 2001. Competitive Intelligence Adds Value:: Five Intelligence Attitudes. *European Management Journal* 19, 5 (Oct. 2001), 552–559. https://doi.org/10.1016/S0263-2373(01)00069-X

[111] Ryan Hunt. 2012. Thirty-Seven Percent of Companies Use Social Networks to Research Potential Job Candidates, According to New CareerBuilder Survey - Apr 18, 2012. http://press.careerbuilder.com/2012-04-18-Thirty-Seven-Percent-of-Companies-Use-Social-Networks-to-Research-Potential-Job-Candidates-According-to-New-CareerBuilder-Survey

[112] 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.

[113] Edward Schumacher-Matos. 2012. Election 1: Fact Checking The NPR Fact Checkers. *NPR* (Oct. 2012). https://www.npr.org/sections/publiceditor/2012/10/28/161839145/election-1-fact-checking-the-npr-fact-checkers

[114] Ricky Sethi and Raghuram Rangaraju. 2018. Extinguishing the backfire effect: using emotions in online social collaborative argumentation for fact checking. In *2018 IEEE International conference on web services (ICWS)*. IEEE, 363–366.

[115] 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

[116] Craig Silverman. 2013. Verification Handbook: A Definitive Guide to Verifying Digital Content for Emergency Coverage. http://verificationhandbook.com/

[117] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease. 2021. The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support. 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 341, 14 pages. https://doi.org/10.1145/3411764.3445092

[118] Ryo Suzuki, Niloufar Salehi, Michelle S. Lam, Juan C. Marroquin, and Michael S. Bernstein. 2016. Atelier: Repurposing Expert Crowdsourcing Tasks As Micro-internships. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2645–2656. https://doi.org/10.1145/2858036.2858121

[119] 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

[120] 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.

[121] Daniel Trottier. 2017. Digital vigilantism as weaponisation of visibility. *Philosophy & Technology* 30, 1 (March 2017), 55–72. https://doi.org/10.1007/s13347-016-0216-4

[122] Roli Varma. 2010. Why so few women enroll in computing? Gender and ethnic differences in students' perception. *Computer Science Education* 20, 4 (2010), 301–316.

[123] Sukrit Venkatagiri. 2022. *Supporting and Transforming High-Stakes Investigations with Expert-Led Crowdsourcing*. Ph. D. Dissertation. Virginia Tech.

[124] 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), 30 pages. https://doi.org/10.1145/3449192

[125] Sukrit Venkatagiri, Anirban Mukhopadhyay, David Hicks, Aaron Brantly, and Kurt Luther. 2023. CoSINT: Designing a Collaborative Capture the Flag Competition to Investigate Misinformation. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (Pittsburgh, PA, USA) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 2551–2572. https://doi.org/10.1145/3563657.3595997

[126] 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

[127] 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.

[128] Nai-Ching Wang, David Hicks, and Kurt Luther. 2018. Exploring Trade-Offs Between Learning and Productivity in Crowdsourced History. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 178:1–178:24. https://doi.org/10.1145/3274447

[129] Claire Wardle. 2014. Verifying user-generated content. *Verification handbook: A definitive guide to verifying digital content for emergency coverage* (2014), 24–33.

[130] Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policymaking.

[131] Etienne Wenger et al. 1998. Communities of practice: Learning as a social system. *Systems thinker* 9, 5 (1998), 2–3.

[132] 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.

[133] 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

[134] Heather J. Williams and Ilana Blum. 2018. *Defining second generation open source intelligence (OSINT) for the defense enterprise. Technical Report*. RAND Corporation Santa Monica United States.

[135] Joseph Jay Williams, Juho Kim, Anna Rafferty, Samuel Maldonado, Krzysztof Z Gajos, Walter S Lasecki, and Neil Heffernan. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*. 379–388.

[136] Anbang Xu, Huaming Rao, Steven P Dow, and Brian P Bailey. 2015. A classroom study of using crowd feedback in the iterative design process. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 1637–1648.

[137] 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*. 220–232.

[138] Haiyi Zhu, Steven P Dow, Robert E Kraut, and Aniket Kittur. 2014. Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1445–1455.

[139] H. Akın Ünver. 2018. *Digital Open Source Intelligence and International Security: A Primer*. Technical Report. Centre for Economics and Foreign Policy Studies. https://www.jstor.org/stable/resrep21048

A APPENDIX

A.1 Reflection Survey

Task

(1) What is your name? *
(2) What was your team name? *
(3) What task was assigned to your team? *

Coordination with Expert

(1) How was your overall experience working with the expert investigator? *
(2) What did you like about the expert collaboration? What could have been better?

Teamwork

(1) How successful was your team overall for the investigative task(s)? *
    On a scale of 1-5, where 1 is not successful and 5 is very successful.
(2) How did your team work together (or not) on the task(s)? How effective was this?
(3) What did you specifically do to help your team solve the task(s)?

Overall

Fig. 2. Snapshot of the spreadsheet containing crowd submissions and corresponding expert feedback for session 4. This session was led by an investigative journalist (E4). The goal of investigation was to identify discourse around anti-vaccine protests occurring throughout Europe as well as the groups involved. The investigation involved the discovery task and verification tasks like geolocation and source analysis.

(1) How confident were you about applying the practised skill in real investigations? *
On a scale of 1-5, where 1 is not confident at all and 5 is very confident.
(2) How difficult were the task(s) overall? *
On a scale of 1-5, where 1 is very easy at all and 5 is very difficult.
(3) How enjoyable was the session?
On a scale of 1-5, where 1 is not enjoyable at all and 5 is very enjoyable.

A.2 Interview Guide for Crowdworkers

Focus group interviews with 3-4 students at a time. [45 minutes]

(1) How did you form your teams and how did the relationship with teammates evolve over time? What were some issues that you faced as a part of the team?
(2) What changes did you notice over time in your performance during expert sessions? Quality/quantity?
(3) Can you tell me about your favorite task among the ones used during practice or expert sessions? [prompt - list of tasks slide]
  (a) What tools and techniques did you use for the task?
  (b) Why did you like it? How is it different from other tasks?
(4) Can you tell me about your hardest task among the ones used during practice or expert sessions? [prompt - list of tasks slideshare]
  (a) What tools and techniques did you use for the task?
  (b) Why did you find it to be hard? How is it different from other tasks?
  (c) What helped you continue to work on the tasks when relevant information was hard to find?
(5) Can you tell me about your favorite expert session? What were the best parts of the session?

(6) Can you tell me about your least favorite expert session? What were the parts you did not like?

(7) Did the self-evaluation measures help you cover the requirements of the tasks? Why/why not?

(8) What else did you enjoy during this experience?

(9) What could be improved during this experience?

A.3 Interview guide for Experts

Post-session Individual interview with expert [35-40 minutes]

(1) What techniques do you use for discovery and verification in your investigation? How do you collect the required information?

(2) How useful are the defined tasks (OSINT macrotasks) in your own investigations? Please rate each task. Follow-ups:
   - Can the student crowd perform the task in the same way? Why/why not?
   - How does this setup act differently?

(3) What are your thoughts about the investigative tasks used in the session? Can you share examples of where you can incorporate these tasks into your work?

(4) Why did you choose the particular tasks from the list of tasks for working with the crowd?

(5) Only if not answered previously - When would you say that these tasks are successfully completed?

(6) What information did you decide to give the crowd before the session? Is there other information you wish you had given them?

(7) How did you plan to spend your time during the session, and what did you end up doing?

(8) How (if at all) did you interact/communicate with the crowd during the session? Any major communication issues?

(9) How well did the students respond to your interventions during the session?

(10) Any blockers in the collaborative setup?

(11) Overall, what did you think about the information submitted by the class in your investigation in terms of quality and quantity?

(12) Will you use it for your investigation? If yes, how? If not, why not?

(13) Do you have any experience with crowdsourcing for reporting? If yes, can you describe your experiences? If not, any particular reason?

(14) How would you describe the effectiveness of the students (compared to the general crowd)?

(15) What did you think about the crowd's self-evaluation of the submissions?

(16) What else did you enjoy during this experience?

(17) What could be improved during this experience?

(18) Would you want to work with a crowd again this way in the future? Why / why not?

(19) Can you provide an example where you could have used the crowd in your investigation