Dropping the Baton? Understanding Errors and Bottlenecks in a Crowdsourced Sensemaking Pipeline

TIANYI LI, Virginia Tech, USA
CHANDLER J. MANNS, Virginia Tech, USA
CHRIS NORTH, Virginia Tech, USA
KURT LUTHER, Virginia Tech, USA

Crowdsourced sensemaking has shown great potential for enabling scalable analysis of complex data sets, from planning trips, to designing products, to solving crimes. Yet, most crowd sensemaking approaches still require expert intervention because of worker errors and bottlenecks that would otherwise harm the output quality. Mitigating these errors and bottlenecks would significantly reduce the burden on experts, yet little is known about the types of mistakes crowds make with sensemaking micro-tasks and how they propagate in the sensemaking loop. In this paper, we conduct a series of studies with 325 crowd workers using a crowd sensemaking pipeline to solve a fictional terrorist plot, focusing on understanding why errors and bottlenecks happen and how they propagate. We classify types of crowd errors and show how the amount and quality of input data influence worker performance. We conclude by suggesting design recommendations for integrated crowdsourcing systems and speculating how a complementary top-down path of the pipeline could refine crowd analyses.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; *Computer supported cooperative work*; *Empirical studies in collaborative and social computing*.

Additional Key Words and Phrases: Sensemaking; Text analytics; Intelligence analysis; Mysteries; Investigations; Crowdsourcing

ACM Reference Format:
Tianyi Li, Chandler J. Manns, Chris North, and Kurt Luther. 2019. Dropping the Baton? Understanding Errors and Bottlenecks in a Crowdsourced Sensemaking Pipeline. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 136 (November 2019), 26 pages. https://doi.org/10.1145/3359238

1 INTRODUCTION

Modern technologies such as social media and mobile devices produce a growing wealth of data. Such data offers an unprecedented opportunity to develop a deeper and more global view of the world, but also poses the risk of spreading misinformation and exacerbating biases [37]. Failing to make sense of this data to prevent terrorist attacks or solve crimes could also harm national security.

Sensemaking offers great potential to understand the meaning and patterns contained within large quantities of unstructured, noisy source materials. Sensemaking is used in many domains, from intelligence analysis to investigative journalism. Pirolli and Card modeled the expert sensemaking

Authors' addresses: Tianyi Li, tianyili@vt.edu, Virginia Tech, Blacksburg, VA, 24061, USA; Chandler J. Manns, chandm8@vt.edu, Virginia Tech, Blacksburg, VA, 24061, USA; Chris North, north@cs.vt.edu, Virginia Tech, Blacksburg, VA, 24061, USA; Kurt Luther, Virginia Tech, Arlington, VA, 22203, USA, kluther@vt.edu.

2573-0142/2019/11-ART136 $15.00
https://doi.org/10.1145/3359238

process as an iterative loop with multiple interdependent steps [42]. Managing this complex process is cognitively demanding and often requires significant person-power [52]. The increasing volume and complexity of data makes it even more challenging, given a limited number of experts.

One way to address these challenges is to involve novice crowds in the sensemaking process. Crowdsourced sensemaking has shown great potential for enabling scalable data analysis and achieving complex goals. For example, crowds can label and create taxonomies of online discussions [10], or perform bottom-up qualitative content analysis [1]. However, most novice crowd sensemaking solutions focus on well-defined sub-problems (e.g., schematizing text data [35]), provide crowds with ideal input data (e.g., raw documents manually broken down by researchers into smaller text items [10]), or require facilitation by experts [7]. The crowd results are perceived as useful and a good starting point, but usually require additional work by requesters [55]. Non-decomposable macrotasks are generally limited to expert crowds [44, 49].

To overcome limitations of requiring experts, we focus on enabling novice crowds to perform the entire sensemaking process, without expert intervention, via a pipeline of microtasks. However, unsupervised crowd sensemaking, where crowd analyses are directly handed off to another group of workers for the next step of analysis, does not always succeed [32, 34]. There are two main challenges: (1) interconnecting the inputs and outputs in a pipelined series of crowdsourcing microtasks, and (2) slicing the large data for microtasks and re-aggregating results. These challenges introduce a level of complexity that is subject to errors and bottlenecks, which could compound when propagated down the pipeline, potentially causing incorrect results. Understanding these effects could enable designers to produce more robust crowdsourced sensemaking pipelines.

In this paper, we probe the errors and bottlenecks that occur in a crowdsourced sensemaking pipeline that connects multiple intermediate crowdsourcing processes to achieve holistic problem-solving without expert intervention. Previous work [34] has shown that such a pipeline enables unsupervised crowd collaboration to solve simple and moderate mysteries, but is challenged by more difficult datasets. Here, we use a similar pipeline to investigate why and how the crowd collaboration is challenged.

Specifically, we aim to answer the following research questions:

RQ1 What are the errors (type and frequency) workers make in a crowdsourced sensemaking pipeline, both within each step and across steps?

RQ2 How does the amount of local data context affect the errors within and across steps in a crowdsourced sensemaking pipeline?

To answer the research questions, we conducted a series of mixed-method studies with 325 crowd workers to solve the difficult fictional terrorist plot in [34]. We first investigated how crowd performance is influenced when given either crowd-generated input or gold-standard input (RQ1). We then examined how the amount of local data context influences worker performance and error propagation (RQ2). We evaluated the crowd performance by comparing their analysis to a gold-standard analysis adapted from the dataset's answer sheet. We classified the types of errors that occurred in each intermediate step, specifically focusing on the source and impact of errors, and how the amount of local context influence the error propagation in the pipeline.

Our analysis indicates that while errors happen in each step and propagate to later steps, surprisingly, both false positive and false negative errors were mitigated to some extent in later steps. In addition, our results suggest that the appropriate amount of local context is different depending on the data formats and the type of tasks. We offer design implications to improve the current pipeline and recommend optimal data context slice sizes in each step.

Our main contributions are:

- a study of the errors and bottlenecks that occur in a crowdsourced sensemaking pipeline in holistic, unsupervised problem solving;
- an analysis of crowd performance in each of the sensemaking steps and the handoffs between them;
- an analysis identifying the trade-offs of the amount of local data context in microtasks.

Our primary intention of this work is to further our understanding of the opportunities and limitations of incorporating crowdsourcing efforts into complex problem-solving in a more scalable fashion. We aim to demonstrate the impact of sensemaking challenges in the context of a holistic sensemaking process, especially the asynchronous analysis handoffs among crowds between connected steps. Rather than artifact creation, we focus on data analysis and sensemaking with multiple distinct data transformation steps that require reusing previous analysis outputs among a sequence of workers. Based on the lessons learned, we draw design recommendations for integrated crowdsourcing systems and speculate how a complementary reverse path of the pipeline could refine the existing crowd analysis.

2 RELATED WORK

In this section, we first discuss the errors and challenges encountered by experts and small-group collaboration in traditional sensemaking. We then discuss the shared and unique errors and challenges for crowds. We review the role of experts in the success of crowdsourcing processes, as well as the role of crowd performance and influencing factors, and distill types of crowd errors seen in prior work.

2.1 Challenges in Sensemaking for Experts and Groups

Traditional intelligence analysts faces the ongoing challenges of parsing, marshaling and synthesizing large quantities of evidence. Analysts need to distinguish pertinent information from noise, deal with incomplete pieces, find potential suspects, to eventually identify the criminal or suspect [18, 52]. The expert sensemaking process is modeled by Pirolli et al. [42] as an iterative loop composed of information foraging and sensemaking (synthesizing). Typical errors of individual expert analysts include wrong or missing information due to inaccurate memory, misinterpreting evidence due to cognitive fatigue, and biases due to perception constraints [11, 22].

Collaboration in the sensemaking processes can help mitigate many individual errors. Analysts from different organizations may have access to different documents, and more readers can sift through larger amounts of data and generate more diverse perspectives to identify alternative patterns. On the other hand, collaborative sensemaking does not eliminate all possible errors made by individuals. Below, we detail the challenges for collaborative sensemaking in small groups.

2.1.1 Additional requirements on shared artifacts and common ground. Collaborating on sensemaking tasks requires analysts to externalize their mental models and represent insights in an understandable way to each other. Research and tool development in collaborative sensemaking aims to support multiple analysts to explicitly work together. Large displays where analysts can annotate, link, and spatially organize documents were proved to establish an efficient visual common ground that facilitate collaborative sensemaking [6]. Small groups tend to rely on shared interfaces and visual metaphors (such as node-link graphs) to co-create concept maps [13]. Such shared artifacts and metaphors are important for a group of analysts to collaborate synchronously on foraging for information, identifying topics and planning more in-depth analysis [9, 17]. However, synchronous collaboration can be constrained by expert availability and does not scale well with a bigger number of analysts. It might also lead to additional errors due to groupthink [24] that produces irrational or dysfunctional decision-making outcomes.

2.1.2 Hand-off timing and instruments. Asynchronous collaboration, however, faces the key challenge of handing off intermediate results between analysts. The efficacy of hand-off heavily depends on timing. If not happening as early transfer or late referral [45], the hand-off is rarely successful. In addition, the instruments of hand-off are important to establish a shared understanding among analysts. Goyal et al. [20] found that visualization of data links is more effective as an intermediate analysis artifact than a notepad of annotations. Schema and visual layout of the information [4, 47] is usually designed to best suit the mental models of previous analysts and are hard to understand without sufficient context and a detailed walk-through. To address this challenge, Zhao et al. [56] developed Knowledge-Transfer Graphs to support hand-off of partial findings during analysis. However, this introduces a new risk of sharing a premature focus on wrong suspects and can derail the overall investigation trajectory.

2.1.3 Teammate inaccuracy blindness and reluctance to share information. Handing-off intermediate analyses can amplify biases and error propagation among analysts. Group biases might be caused by similar backgrounds of analysts or by individuals misleading the group. Kang et al. coined the term "teammate inaccuracy blindness" [25] to describe the phenomenon where previous work from a partner is assumed valid and useful without sufficient quality checks. Inaccurate information can be reused and premature focuses can be built upon by other analysts. On the other hand, analysis may fear their own analysis is wrong and hesitate to exchange information and insights [22]. Goyal et al. [19] proposed a social translucence interface to balance the visibility and quality of analysis between distributed collaborative pairs, but it is unclear how well such approaches would scale to a large number of analysts.

Some of the above-mentioned challenges for experts can be alleviated in a crowdsourcing context. For example, the crowds can delve into significantly larger amounts of information with less fatigue and more diverse perspectives. It is also easier to require use of a certain artifact to promote sharing information with novice crowds. However, the novice crowd's lack of expertise and variability on different tasks can cause crowd-specific challenges and errors.

2.2 Challenges in Crowdsourced Sensemaking

Crowdsourcing has been successfully applied to many complex sensemaking problems. Crowds can identify unknown individuals from old photos [38], provide reliable annotations on named entities in multimedia Twitter data [16], and contribute "outside-the-box" thinking for innovative problem-solving [54]. Below we review the expert intervention to prepare and guide crowd tasks and the crowd performance in current crowdsourcing solutions.

2.2.1 Expert intervention to prepare and guide the crowd tasks. Many evaluations of crowd systems provide crowds with ideal input and detailed task specifications to illustrate best-case scenario results. In Mobi [55], the crowds were given very detailed background information and bulleted lists of traveling goals to plan an itinerary. In Cascade [10], researchers manually break down original Quora responses into smaller text items. When providing analysis and explanations on social data [51], crowds are given nicely visualized and carefully selected charts, with hints and examples relevant to the tasks. In a hidden profile task [48], crowd workers were given well-written profiles with no typos or errors. During some open-ended crowd processes, expert also need to provide real-time guidance [7, 33] or heavy-duty centralized coordination [49]. While crowds showed potential to subcontract existing microtasks [39], it is unclear how subcontracting can be applied successfully in more complex problem solving efforts with multiple interdependent steps. Novice facilitators [7] and crowds [32] are shown to be inadequate to adapt a given workflow and produce unsuccessful results as a consequence. Chaining multiple crowdsourcing processes without the above-mentioned expert facilitation could cause unexpected errors and problems.

2.2.2 Crowd performance and requester decisions. After the crowds complete micro-tasks, experts often need to curate the mixed-quality results [21] and solve the remaining problems. The accuracy in crowd work depends on the task, context, and the baseline condition. Reported accuracy is often around 60% [1, 16, 26] and sometimes can be as good as above 90%. For example, crowds can create a global taxonomy of online question datasets with quality 80-90% of that of experts [10]. Willet et al. [51] proposed seven strategies to improve crowd performance and achieved 63% useful responses in the best results. CRICTO [12] reports that 73.98% of crowdsourced links in a sensemaking exercise were rated valid by authors. In some mixed-initiative systems [8, 14], no standalone crowd performance was reported. Many papers focus on indirect quality measures such as the number of responses [5], or subjective ratings of the tasks [40], rather than comparing crowd results to a gold standard. Crowds have demonstrated the promising capability of solving complex problems, but even successful systems cannot completely eliminate errors in the analysis. It remains unknown how imperfect parts of the analysis may influence later analysis.

Various requester decisions beyond poor task design also influence crowd performance. Lack of workflow transparency [27] can decrease quality and volunteerism, and a higher number of perceived co-workers can induce social loafing [40]. In addition, US-only workers tend to outperform non-US workers [14, 51], and a qualification test [14] can improve task performance. Some studies recruited expert crowds [49] or volunteers from social media [5], who tend to have higher quality performance than those from paid platforms like Amazon Mechanical Turk (MTurk). In this work, we chose a low recruiting requirement (acceptance >90% without enforcing US-only crowds) to investigate errors made by a broad range of crowd workers. Meanwhile, we draw on the findings in previous work to eliminate errors caused by poor requester decisions.

2.2.3 Crowd challenges and errors. Crowdsourcing as a paradigm applied to sensemaking problems is challenged by the tension between the microtask local view and the global goal, optimal decomposition of the process and the data into hierarchical workflows and task assignments, as well as management and quality control of a large-scale workforce.

Fragmented and distributed local data can cause irrelevant [51], missing, or incorrect judgements [10]. Crowd analysis can also be focused on only a fraction of the given information due to unevenly distributed data. While devoted analysts have access to the entire data set to gain a rich understanding of global themes, paid crowd workers usually commit only a short period of time, and thus are only able to work with a small portion of the data. Decomposing the data into local microtasks makes it difficult for workers to accomplish high-level synthesis tasks, like identifying emergent global categories in the data. State-of-art solutions include increasing the amount of local data [35, 48], re-representing and condensing the raw data [1, 50], or iteratively revisiting the previous results [10].

Parallel analysis by many workers may lead to multiple interpretations of the same data. To avoid falling into an infinite loop of "categorizing the categorization", hybrid systems are introduced to recognize duplicates and conflicts in the analysis [21, 29] and reassign the edge cases to crowds [8] to consolidate the analysis.

The mechanisms of MTurk and similar platforms have been criticized for incentivizing low-quality work, such as random guesses [16]. Some crowd workers might not pay attention to the given input [26]. Other low-effort errors include unclear or speculative responses, inattention to details, or focusing on superficial facts [51]. Requesters can improve worker engagement with more formative instruction language [3], peer-evaluation [53] or even mutual reward dependency [23]. There are also visual analytics tools that support monitoring worker's task status and managing the overall workflow [28]. Reviewing other people's work can help workers improve their own

results [31]. On social media platforms, people tend to engage in self-correcting rumors when encountering information conflicts [2].

High-quality results in previous works provide proofs-of-concept of crowd capabilities and the efficacy of proposed methods. However, little research focuses on understanding good-faith reasons why workers struggle, make errors, and fail. Our research addresses this gap and frames the findings within the broader sensemaking loop to make them relevant to many types of crowdsourced sensemaking and data analysis systems.

3 EXPERIMENT DESIGN

In this section, we first describe the experiment setup and rationale, and then outline the methods and details of the experiment.

3.1 Problem and dataset: solving mysteries

We focus on the problem of solving mysteries as an example sensemaking process in our experiment. It is also an important real-world task for which crowdsourcing is increasingly used [12, 34]. It contains all the tasks and stages in the sensemaking loop [42] and thus, represents a good coverage of crowd-powered sensemaking processes.

In our study, the mystery is the same as the one that crowds failed to solve using CrowdIA [34]. It is adapted from a real-world professional training exercise for intelligence analysts. For publication purposes, we have changed some of the names and places used in the dataset. The scenario is about a fictional terrorist attack and the goal is to identify the target location of the attack. The following known facts are shared as global context among all participating crowd workers:

> A C-4 plastic explosive bomb will be detonated at 0900hrs on 30 April 2003, by a group of terrorists: Harvey Wulfen, Cedric Whappadder, Joed Shearper, Irving Sprunkiddle. Where is their target location?

The known facts seem to be abundant, yet the mystery is actually difficult to solve. The C-4 explosive bomb is masked, stored, and transferred among multiple places. Three of the terrorists have aliases and forged documents to cover their activities. The data set contains many phone calls among anonymous numbers, voice messages with code words, with the phone number holder information in separate, distributed documents. Previous work [34, 46] indicated that the mystery is difficult for one committed analyst.

The dataset is composed of 15 fictional report documents from intelligence agencies. Ten have key information relevant to the attack but also contains noise (irrelevant information) within the documents. The 5 completely irrelevant documents are intentionally misleading, with similar terrorist activities, timing, and weapons. The lengths of the documents are mostly around 400 words, but there is one with 193 words and one with 1189 words. The longest document contains about two-thirds noise.

3.2 Instrument: a crowdsourced sensemaking pipeline

To understand the crowd's capability in different sensemaking tasks and the error propagation in interconnected steps, we conducted the experiment using a system similar to CrowdIA [34], which was adapted from the expert sensemaking loop [42].

The pipeline (Fig.1) guides the crowd workers through five sensemaking stages from the bottom up. *Step 1. Search and Filter* takes all the raw external documents and rates their relevance. The output is a subset of documents considered as more relevant. *Step 2. Read and Extract* takes the relevant documents and extracts important evidence information. The output is a list of information pieces structured as simple sentences. *Step 3. Schematize* takes the information pieces and organizes

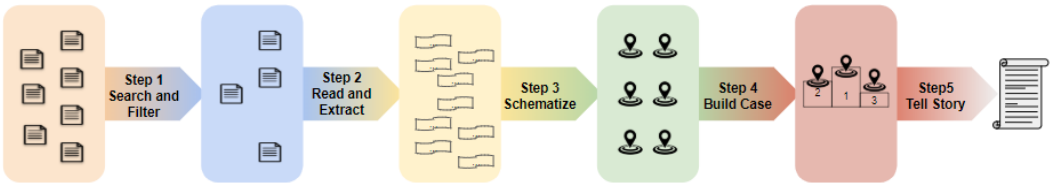

Fig. 1. The crowdsourced sensemaking pipeline. There are 5 steps connected by their inputs and outputs: Step 1 *search and filter* relevant documents; Step 2 *read and extract* important information pieces; Step 3 *schematize* information pieces into profiles of candidate locations; Step 4 compare candidates to *hypothesize* on the most likely one; Step 5 *present* the final conclusion as a narrative *story*.

them into profiles of supporting evidence for possible solutions. *Step 4. Build Case* takes the profiles and rank the likelihood of each. The output is the most likely answer. *Step 5. Tell Story* takes the most likely answer with its profile and outputs a final presentation that expounds how the answer fits into the known facts and solves the mystery. Within each step, the system slices the step input into small and contextually relevant pieces, *context slices*, and distributes them among crowd workers as microtasks. The system automatically generates microtasks and aggregates crowd analyses in each step sequentially following the pipeline.

In our implementation, we draw on previous works to eliminate errors caused by poor requester decisions. In the instructions, we explain to crowd workers that they are analyzing the results of previous workers and their analysis will be used by future crowds, to provide workflow transparency [27]. In addition, we followed the task designs in the previous successful deployments of a similar pipeline [34].

3.3 Participants

We deployed the pipeline on Amazon Mechanical Turk (MTurk) with workers of higher than 90% approval rate. This is a relatively lower requirement compared to most of the similar crowdsourcing research. For example, Crowd Synthesis [1] used 95% approval rate on MTurk with additional training, and flash teams used crowds of experts from Upwork [49]. Our goal is to involve crowd workers at a larger scale and make our results more generalizable to different real-world problem-solving situations. We estimated the time needed for each microtask based on pilot studies and paid a fixed amount for each Human Intelligence Task (HIT) with the minimum wage of our location ($7.25 per hour).

3.4 Task and procedure

Our experiment aims to investigate the crowd competency in different sensemaking tasks, and probe the source of errors by manipulating the quality of step input (gold-standard or crowd-generated) and the size of context slices (1, 3, or all items in the step inputs) (Fig. 2).

3.4.1 Context slicing methods and choices. In the uni-item condition, each item (documents, info pieces, profiles) in the step input is considered a context slice, so the number of context slices equals to the number of items in the step inputs. In the all-item condition, the entire step input is considered as one context slice, so the number of context slices is always one.

In t theriple-item condition, the ways the items are distributed among context slices of size 3 is more complicated. Taking Step 1 as an example, there are $\binom{15}{3} = 455$ possible combinations to distribute the 15 raw input documents into context slices of size 3. As a proof-of-concept, we implemented a context slicing method that favors item similarity defined by entity overlap and does

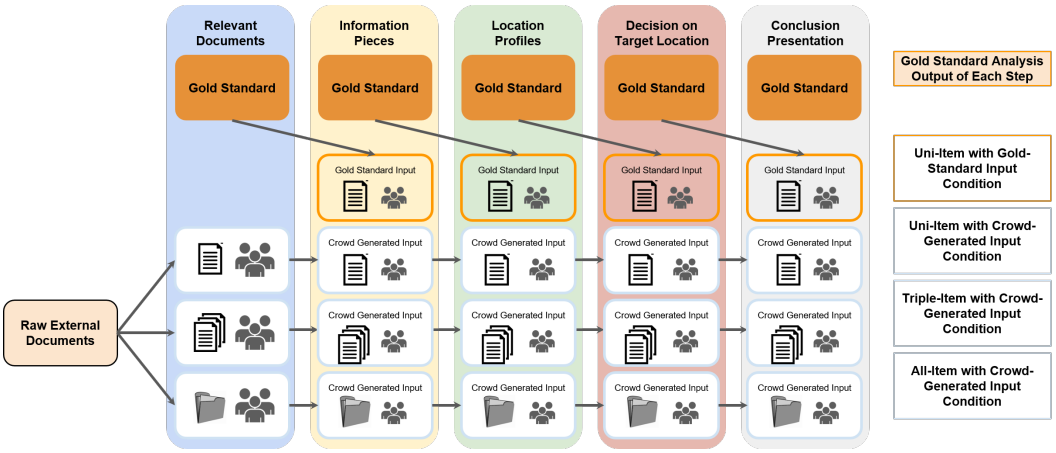

Fig. 2. Experiment Design: Besides the five different steps in the pipeline, we manipulate the quality of step input (gold-standard or crowd-generated) and the size of context slices (1, 3, or all items in the step inputs). There are 4 conditions in total.

not allow overlapping items between context slices. This is to guarantee that the information is evenly distributed among workers and each input item is analyzed exactly once. If some documents appear in more context slices, they will be analyzed in more microtasks and by more workers. As found by Willett et al. [51], this can introduce biases favoring the information in more heavily analyzed documents.

3.4.2 Pipeline execution walkthrough. One execution of the pipeline results in one batch of mystery-solving analysis from steps 1 through 5. For each batch, the system starts with the 15 documents and executes the five steps sequentially.

Step 1: rate document relevance. The step input is always the 15 raw documents in all conditions. For each context slice, the system assigns 3 crowd workers to rate the relevance from 0 (completely irrelevant) to 100 (completely relevant) of each document in the slice. Ratings above 50 (neutral) are considered as positive. Each worker is also required to briefly explain their rating rationales in a text box. Documents with a majority vote for positive relevance are considered as relevant.

Step 2: extract important info pieces from relevant documents. For each context slice, the system first assigns 1 crowd worker to extract all the important info pieces from the documents in the slice. The crowd was instructed to format info pieces as simple sentences structured as "who, what, where, when" as much as possible. After that, documents in each context slice and the extracted info pieces are assigned to a second worker for review. Reviewers are instructed to correct errors in the existing info pieces, delete any bad or useless ones, and add new ones to include any missing information.

Step 3: tagging info pieces. For each context slice, the system assigns 3 crowd workers to tag the possible target locations mentioned (if any) and the related known facts in each info piece. Each worker is also required to briefly explain their tagging rationales in a text box. For each info piece, workers will select one or more tags. Tags with a majority vote will be attached to the info piece. Info pieces that are tagged the same location names are all put together to form a *location profile*, and organized by their evidence types.

Step 4: evaluate location profiles. For each context slice, the system assigns 3 workers to evaluate the profiles in it. When context slice size is 1, each worker rates the likelihood of the location to be

the attack target, from 0 (completely unlikely) to 100 (completely likely). Ratings above 50 (neutral) are considered as positive. Each worker is also required to briefly explain their rating rationales in a text box. The location with the highest average rating is considered as the most likely. Otherwise, the workers pick the most likely location in a given context slice and explain the rationale in a text box. Locations with the majority vote are considered as more likely than the others in each context slices. The process repeats until only one most likely location is left.

Step 5: write a narrative presentation to summarize the conclusion. Step 5 only works on the most likely answer thus is the same across conditions. For the most likely location, the system first assigns 1 crowd worker to write a narrative story that explains why the given location is most likely the target of the attack. After that, the story and the winner profile are reviewed by a second worker. In reality, Step 5 only needs to run once with two workers (one writer and one reviewer), but for the purposes of this paper, we ran it 4 more times (4×2 = 8 workers) to gather more data and be comparable to the other conditions (Table 1).

3.5 Data Analysis

The data that informs this analysis includes the microtask responses submitted by the workers; the step outputs aggregated by the system; the system log of workers previewing, abandoning, and submitting the tasks; and the login/logout time for each worker.

We first compared the crowd analysis to the gold-standard analysis. Except step 1, all other steps with crowd-generated input require a qualitative comparison between the crowd output to the gold-standard output. Specifically, for the info pieces extracted in step 2, the crowd might extract the same information in different ways. We coded for two levels of correctness: 1) *matching* the gold-standard info piece, and 2) *not matching but relevant* and useful to solving the mystery. Since the crowd might partially extract the info pieces, we also count the number of matched and relevant elements in the crowd results. An element is any one of the "who, what, where, when" items in the info piece. In step 3, we compare the resulting location profiles to the gold-standard analysis. We also qualitatively examine the tagging results and explanations by the crowd. In step 4, we manually rank the crowd-generated locations with the same criteria as used when ranking the gold-standard locations, then compare with the crowd rankings. In step 5, we examine the number of retrieved key evidence and qualitatively evaluate the writing by the crowd.

In addition, we open-coded and analyzed the crowd explanations from steps 1, 3, and 4, identifying common behaviors and speculating on crowd analysis rationales. Two authors first sampled around 10% of the data and analyzed separately, then compared the coding to agree on a set of codes with clear definitions. After that, author A focused on comparing the crowd results to the gold-standard analysis, while author B focused on coding the explanations provided in steps 1, 3, and 4. The two authors then reviewed and iterated on the analysis until reaching a consensus. From the task performance perspective, we categorize the source of error with respect to data quality and task behavior (Fig. 3). We classify the data correctness by comparing to gold-standard analysis. We define "the right thing" in task behavior by the following four possible levels of analyses:

- Accurate: true to the information source (directly copied from the document text)
- Focused: relevant to the investigation goal
- Interpretive: rephrase what the facts *mean* (not directly copying)
- Deductive: synthesize facts and develop hypotheses (including facts from multiple documents and hypotheses not directly mentioned in any document)

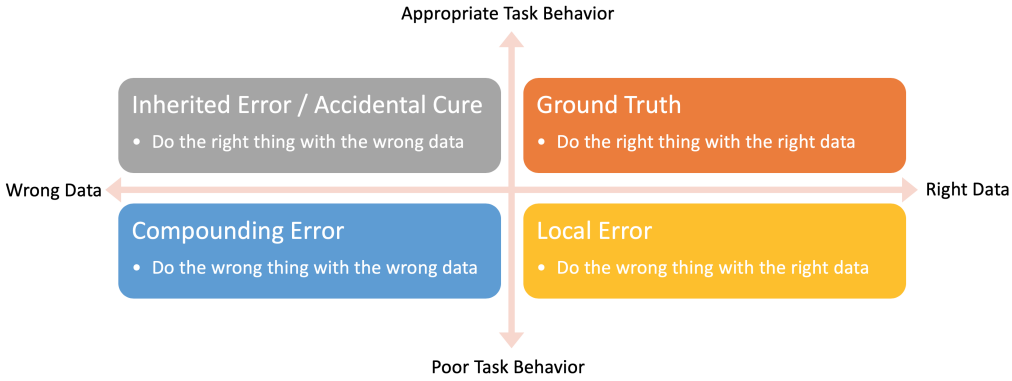

Fig. 3. Task performance measured by two sources of errors: data quality and task behaviors

Pipeline Components	Uni-item GI	Uni-item CI	Triple-item CI	All-item CI	Step Total
Step 1. Search and Filter	45	45	15	3	63
Step 2. Read and Extract	20	18	8	2	48
Step 3. Schematize	57	48	15	3	123
Step 4. Build Case	15	24	9	3	51
Step 5. Tell Story	2+8=10	2+8=10	2+8=10	2+8=10	8+32=40
Total	139+8=147	92+8=100	49+8=57	13+8=21	293+32=325

Table 1. Number of workers hired in each step and each condition, and the total number of workers in each step across conditions. While Step 5 only requires two workers (one writer and one reviewer), for the purposes of this paper, we ran it 4 more times (4×2 = 8 workers) to gather more data and be comparable to the other conditions.

3.6 Limitations

Our evaluation studies have several limitations. We focus on one specific pipeline and software, one crowdsourcing platform with one recruiting requirement, and one example data set. These choices might limit our generalizability. However, the pipeline modularizes the mystery solving process into representative subtasks, and results on local tasks resonate with prior work. Future research should explore the pipeline application in different types of problems and scenarios.

While the pipeline is adapted from the sensemaking loop that is widely used in sensemaking research, including crowdsourced sensemaking, there are alternatives, such as data-frame theory [30], that are also prevalent. It is possible that the crowd shows different analysis performance with a different underlying theory and framework.

Additionally, our participants are recruited from Amazon Mechanical Turk with a low requirement of 90% approval rate. Workers with different approval rate or from different platforms, such as expert crowds from Upwork or volunteers from Reddit, could have different reactions and behaviors to the tasks and collaborations.

4 RESULTS

The correct answer for the target location in the mystery is "NYSE" (New York Stock Exchange). On the one hand, no CI conditions found the exact correct answer. On the other hand, the uni-item

and triple-item CI conditions found evidentially and geographically close locations, and provided meaningful analysis provenance that explains why the crowd made certain correct or wrong decisions. This good performance is consistent with prior evaluations of a similar pipeline [34]. In this section, we focus on reporting the types and sources of the errors crowds make in each step (RQ1) and how crowd analysis in each step is different when given more local context (RQ2).

4.1 RQ1: Error types and propagation

We compare the crowd analysis under the GI uni-item and CI uni-item condition to investigate how the quality of input influences worker analysis performance. Counting the additional data collection in step 5, this includes 231+16=247 crowd workers in total: 139+8=147 workers for GI and 92+8=100 workers for CI (Table 1).

4.1.1 Different numbers of crowds were hired in GI and CI conditions. Passing on analysis results by previous crowd workers introduces uncertainty in the hiring process. Since the crowds selected 9 relevant documents in step 1, step 2 in CI condition only hired $9 \times 2 = 18$ workers to write and review info pieces. This resulted in fewer info pieces and thus, fewer workers were hired in CI step 3. In CI step 3, the crowd tagged additional irrelevant locations from the irrelevant information. Therefore, the CI step 4 hired more workers to evaluate the likelihood of all those locations.

4.1.2 Local error types and examples. Comparing the crowd analysis with the gold standard analysis, we found local errors in each of steps 1–4. Crowd did well in step 5 with gold-standard input, and the errors in the CI condition are due to error propagation. We first categorize the local errors as follows, then discuss error propagation in the following subsection.

Insufficient context errors. In step 1, two relevant documents refer to terrorists by their aliases and were left out by the crowd. A different document reveals that those names are terrorist aliases. Insufficient context also led workers to rate irrelevant documents as relevant, because there was not enough information to prove the document irrelevant, and the information might be "worth looking into". In step 2, the information about terrorist aliases were not extracted. In step 3, the info pieces about terrorist aliases were tagged as not containing any relevant evidence. In step 4, many related locations are rated as likely target locations. One worker pointed out that "it's also very possible that it [Empire State Vending Services (ESVS)] is somewhere that they're just using as part of their cover stories", but still rated ESVS as a likely target.

Misinterpretation errors. In step 1, one relevant document was rated irrelevant by a worker. The explanation was "report date and deposit is dated after the date in question" but the dates in the document are actually before the attack date. In step 2, a worker extracted an info piece "Cedric Whappadder announced he would pick up the carpet..." that misinterprets the information in the document; Cedric Whappadder is the carpet store owner, rather than the customer. In step 3, a worker tagged "Sudan and Afghanistan" as one candidate location in info piece "Joed Shearper recieved explosive training Sudan and Afghanistan". This indicates that the worker did not understand that 1) Sudan and Afghanistan are two different locations and 2) those locations are where the terrorists received training, not the targets of the attack. In step 4, a worker rated the New York Stock Exchange (NYSE) as irrelevant because "there is not a lot of mention about the stock exchange specifically". The worker only focused on the frequency of a location being mentioned, but did not interpret how this location is connected to the known facts of the attack.

Inattention to background knowledge. In step 1, one document directly mentioned a terrorist name, but workers rated it as irrelevant by the majority vote. Two of the workers did not mention anything about the terrorist name in the explanation. In another example, one irrelevant but misleading

document about an attempted bombing was rated relevant, but it involves a different time from the known attack time. This indicates that some crowd workers did not pay attention to the known facts (e.g., terrorist names and the attack time) given in the instructions.

In step 2, the important information about the attack weapon (C-4 explosive) was not extracted. The workers only wrote about a cigarette being tossed into a waste basket in a carpet shop, but did not mention that this resulted in a fire and led the firemen to discover several cartons of C-4 explosives, nor did anyone mention that the carpet shop belongs to one of the terrorists.

In step 3, one crowd worker tagged an info piece, "Cedric Whappadder has C-4 explosives in the basement of his carpet shop until April 26, 2003" to have candidate location "in the basement of his carpet shop". This indicates that the worker did not pay attention to the known attack time (April 30) given in the instructions. The explosives are moved before the attack thus the carpet shop cannot be the target location.

In step 4, one worker explained that "it was written in the page above that bomb attack will take place in new jersey april 26 2003." This, too, conflicts with the known facts that the attack was to take place on April 30, 2003. The worker might have mistaken the date for other dates mentioned in the task.

Failing the task goals. In step 1, a worker rated a relevant document as irrelevant, but pointed out the relevance of the document in the explanation: "Although the sentences describe how this attack may have been funded, there is nothing there that would make one aware of the location of the attack." The worker did not fulfill the task goal to rate documents as relevant if they contain information about the known facts of the terrorist attack.

In step 2, a worker extracted an info piece that reads, "I LISTED IN THE CITY NORTH BERGEN NJ ON APRIL 22,2003". This crowd worker put "I LISTED IN THE CITY" in the "what" field. This indicates that the worker did not follow the instructions to write complete sentences about the important information to solve the mystery.

In step 3, some workers put terrorist names as a candidate location tag. Some workers selected all the evidence tags, explaining, "*I chose the tags above because it was stated in the instructions that they knew the weapon, the time and date, as well as, the group of terrorists who are expected to detonate the weapon.*" The worker did not understand that the task is to find the info pieces that are relevant to the known facts.

In step 4, a worker explained, "However there is no details related to the attack location or target of attack. It is extremely difficult to extract details." The worker didn't understand the task is to rate the likelihood of the given location based on the available information in the profile.

Low effort errors. In step 1, one worker put "goode" in the explanation box. In step 2, several workers directly copied text from the documents to fill in the "who, what, where, when" fields that do not combine to read as a meaningful sentence. In step 3, some workers put "Available Material" as a candidate location tag, and put "good" as the explanation. In step 4, some workers put "Available Material", or "this is clear" as the explanation.

4.1.3 Error propagation. We analyze crowd error propagation by tracing the crowd analysis on previous errors (the left half of Fig. 3) and comparing it to the equivalent in the GI condition. The errors in earlier steps led to an increasing number of errors due to insufficient context and missing information (*inherited errors*), as well as more low effort and other local errors (*compounding errors*). On the other hand, some irrelevant information was filtered out in later steps (*accidental cure*). Below, we describe how each step is influenced by these types of error propagation.

Step 2 was influenced by inherited errors and compounding errors. Overall, step 2 only retrieved around half of the gold-standard info pieces. The CI condition missed all the information from the

3 missing relevant documents (*inherited errors*) and included additional irrelevant information from the 2 extra irrelevant documents (*compounding errors*). However, the CI condition extracted more matching info pieces than the GI condition, even though the input has fewer relevant documents. To further understand this surprising outcome, we analyzed the individual responses of the microtasks for the 7 relevant documents shared in both conditions. It turns out most errors (i.e., incomplete or irrelevant sentences) are due to local errors (*failing the task goals*). We speculate that the varied performance in the same task might be because the workers are overwhelmed or confused by the task and did not extract more info pieces than the minimum requirement. A follow-up experiment repeated step 2 for both the GI and CI conditions with the same microtasks but enforcing a minimum of 2 info pieces. The results confirmed this intuition. The new crowd ($N' = 20 + 18 = 38$) mostly extracted 2 info pieces, but the overall quality did not improve. Since the design choices are consistent with the previous successful deployment of a similar pipeline pipeline [34], we suggest that extracting and restructuring information (step 2) is the most challenging step of the pipeline with more challenging dataset and longer documents.

Step 3 was influenced by all types of error propagation. The crowd-generated info pieces are less understandable due to incomplete and poorly structured sentences, typos and grammar errors, and some are written in all capital letters. As a consequence, some important information was tagged as not containing relevant evidence and introduced additional false negative errors (*inherited errors*). In addition, misleading information continued to be tagged by evidence types and propagated strongly. The crowd generated 8 location profiles, of which 3 are from irrelevant documents (*compounding errors*). There was also an increased number and percentage of meaningless explanations (*low effort errors*) in step 3. Most of them occurred in info pieces about aliases, phone numbers, etc., that require more context to tag. We suggest that the previous poor-quality analysis provided less context and might have confused the workers about the task goals. On the other hand, some false positive errors from step 1 and 2 was cured by step 3 crowd, because they couldn't find any evidence related to the known facts (*accidental cure*).

Step 4 was influenced by all types of error propagation. The CI condition crowd rated a fake apartment address of terrorists in NYC as the most likely target location. The correct answer NYSE was not in the step 4 input since the step 1 crowd did not rate the corresponding document as relevant (*inherited false negative error*). *The USA*, a very low-resolution location that nevertheless encompasses the correct answer, received almost the same score and ranked second place. Rating the apartment address as a likely target location, whose profile contains irrelevant and wrong information, as well as missing some important relevant information, is a *compounding error*. The irrelevant profile, on the other hand, were rated as unlikely to be the target location (*accidental cure*). We further analyzed the explanations in CI step 4 to understand how the crowd was able to mitigate the previous errors. We found that the crowd compared the available information to the known facts, identified and excluded non-logical possibilities (e.g., locations not worth attacking), recognized cover-up activities (e.g., that the "carpet" to be picked up is actually C-4 explosives), and brought in their common sense for geographic proximity (e.g., identified explosive location in the carpet store in North Bergen, NJ and suggested the target is nearby).

Step 5 was influenced by all types of error propagation. While the GI presentation logically connects all 6 gold-standard facts (see Appendix A.1), the CI presentation only has 1 matched fact. The final CI presentation eliminated the two irrelevant pieces of information from the given location profile (*accidental cure*), revealed the connection between Cedric Whappadder with C-4 explosives, but reused the misinterpreted information (Cedric Whappadder picking up the carpet) from the wrong location profile (*inherited errors*), and connected the wrong information with the correct

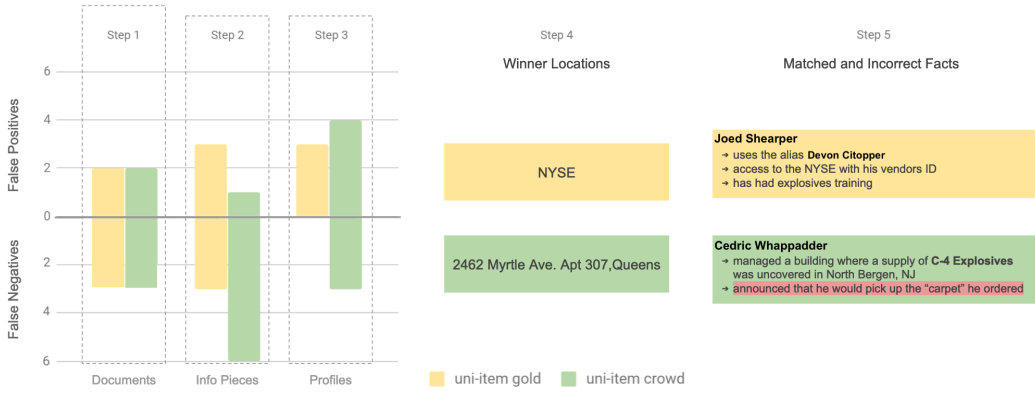

Fig. 4. RQ 1: Errors in GI (gold-standard input) and CI (crowd-generated input) conditions.

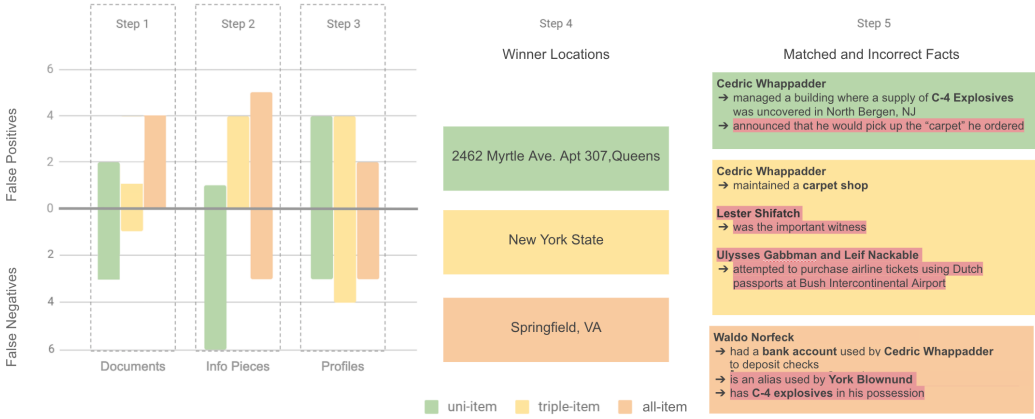

Fig. 5. RQ 2: Errors in uni-item, triple-item and all-item conditions.

information (*compounding errors*). The additional data collected in step 5 is consistent with this result. The GI condition presentations are all complete and cohesive. Three of the new CI condition presentations inherited the wrong information about Cedric Whappadder picking up the carpet, one of which included more false positive errors from the wrong location profile. The other one new CI presentation is very short: "all terrorist attack attempt in April month". We consider this as a local error (low effort).

Overall, while the compounding of the errors in the CI condition is problematic, the GI condition of Study 1 suggests that if each step can be improved (perhaps through review or refinement processes), the chaining of the steps in a pipeline can be a successful strategy for crowd sensemaking.

4.2 RQ2: Impact of context

We analyze the crowd analysis in CI uni-item (Fig. 7), triple-item (Fig. 8), and all-item (Fig. 9) conditions to investigate how the amount of local context influences the worker performance. Counting the additional data collection in step 5, this includes 154+24=178 crowd workers in total (Table 1).

4.2.1 Step 1: increasing context changes frequency of different local errors. The uni-item condition retrieved 7 (out of 10) relevant documents and 2 (out of 5) irrelevant documents; the triple-item condition retrieved 9 relevant documents and 1 irrelevant document; the all-item condition retrieved all 10 relevant documents but also 4 irrelevant documents (Fig. 5).

Increasing context reduces unforced errors due to insufficient context. Both triple-item and all-item conditions successfully retrieved the one document about NYSE, while the uni-item condition failed to. Increased local context enabled workers to use information from different documents and retrieved documents with hidden relevance. The all-item condition saw the most references to related documents in the explanations.

Increasing context introduces unforced errors due to misleading context. On the other hand, the relevant documents in each context slice did not help crowds eliminate the irrelevant document. Rather, more irrelevant documents were rated as relevant. Our analysis of the explanations indicates that workers also drew connections between entities such as the country names and date time mentioned in the irrelevant and relevant documents (*unforced errors* due to misleading context). For example, one worker put in their explanation that "previous phone call made out from the Netherlands, and the passports are Dutch in this document." The call from the Netherlands is related to the terrorists in the mystery, but the Dutch passport is linked to a different crime.

Increasing context increases frequency of inattention to the background knowledge. In addition, workers in bigger context slice conditions are generally more likely to rate documents as relevant since they might contain "potential clues" in their explanations, though we already listed the relevant terrorists in the instructions (*inattention to the background knowledge*). This might indicate that too much context prohibited workers from focusing on important information.

Increasing context increases frequency of low effort local errors. Compared to the 1 low effort explanation in uni-item (from 1/45 workers), there were 7 low effort explanations in triple-item (from 4/15 workers) and 30 low effort explanations in all-item (from 2/3 workers).

4.2.2 Step 2: increased context overwhelmed workers. Workers in the uni-item condition extracted 16 info pieces from 9 documents (9 context slices), the triple-item condition extracted 12 info pieces from the 12 documents (3 context slices); all-item workers extracted 7 info pieces from the 14 documents (1 big context slice). None of the conditions extracted info pieces about NYSE.

Increased local context helped eliminate compounding false positive errors from step 1. The uni-item condition extracted info pieces from every document retrieved in step 1, thus some false positive errors propagated in step 2. The triple-item condition eliminated 1 of the 3 irrelevant documents. The all-item condition eliminated 2 of the 4 irrelevant documents.

Increased local context reduced unforced errors. The triple-item and all-item condition synthesized information from more than one document, which is not possible in uni-item condition. One crowd worker from the triple-item condition connected the phone number addresses and the owner employment information. This reveals the identities involved in the mysterious phone calls reported in different documents.

Increased local context provoked higher frequency of "failing the task goals" local errors. A higher workload led to more relevant documents being missed. In the uni-item condition, no relevant document was completely ignored, even though the information in the documents was not fully extracted. Yet, in the triple-item condition, 3 relevant documents were completely ignored, and in all-item condition, 5 relevant documents were completely ignored (*failing the task goals*). Fewer gold-standard info pieces and elements were retrieved in triple-item and all-item conditions, despite

the increased availability of relevant documents. There were fewer relevant (although not gold-standard) info pieces and elements, as well.

4.2.3 Step 3: increased local context reduced local errors, but also suffered more from propagated errors. The uni-item condition generated 8 profiles from the 16 info pieces, of which 3 are from irrelevant documents/info pieces. The triple-item condition generated 7 profiles from the 12 info pieces, of which 3 were from irrelevant documents/info pieces. The all-item condition generated 4 profiles from the 7 info pieces, of which 2 are from irrelevant documents/info pieces. None of the conditions created a profile for NYSE.

More context reduced unforced errors but too much can lead to more low effort errors. In uni-item condition, 3 irrelevant info pieces were eliminated and 8 workers provided 8 low effort explanations. In triple-item condition, 1 irrelevant info piece was eliminated and 3 workers provided 5 low effort explanations. In all-item condition, 1 irrelevant info piece was eliminated and 2 workers provided 30 low effort explanations. We conjecture that providing explanations to every single info piece encourages workers to analyze the info pieces more carefully, but could be too arduous with big context slices.

The distribution of context could limit accidental cure and encourage inherited errors. With our design, the context slices do not overlap, so it is hard to bring together the most optimal context without reusing the same info piece in multiple context slices. Some info pieces from step 2, though incomplete, could still make sense when put together with other info pieces. However, the KNN context slicing algorithm might not successfully put them in the same microtask, thus preventing effective tagging of those info pieces.

4.2.4 Step 4: increased local context enhanced accidental cure with more in-depth analysis. Workers in the uni-item condition selected "2462 Myrtle Ave. Apt 307, Queens," the apartment address listed under two terrorist aliases, as the most likely target location. The triple-item condition selected "new york state" and the all-item condition selected "Springfield VA."

Increased context encouraged relative comparison and mitigated propagated errors. Despite the sparse information in each profile, the crowd was able to compare the relative likelihood of given locations with external knowledge and common sense. For example, one worker mentioned in the explanation that they recognized that "*the phone calls from Ramazi are originating from 703 area code — Virginia.*" Given insufficient information about each profile, some workers focused on eliminating unlikely profiles, rather than selecting likely ones. One explained, "*As for why I chose Springfield, it is the only unclear one. Two have no direct relation to the terrorist, and the third seems to be a home address.*" Although the final decision is farther from the correct location (NYSE), the analysis is more logical and accurate than workers who selected the apartment address of the terrorists. There were also zero low effort explanations in the triple-item and all-item conditions. We speculate this is because workers were less sure about their result and felt more obliged to explain their uncertainty and thought processes.

4.2.5 Step 5: workers had more compounding errors. The winning profiles fed to step 5 in all conditions consist of almost half irrelevant info pieces. The workers introduced compounding errors by connecting the relevant info pieces with the irrelevant ones with dates since all documents are reports collected in mid-April 2003. The uni-item condition presentation contains 1 matched fact and 1 irrelevant fact. The triple-item condition presentation did not have any matched facts. There were 3 relevant facts connected to irrelevant information with dates. The presentation mostly described "evidence led the police to investigate…" The all-item condition presentation has 1 matched facts and 3 relevant facts. The two crowd workers build on top of each other's

imagination and created a nice and long story with additional imagined information. Despite the false positive errors, the resulting prose is actually written in a similar fashion as the two-page long crime reconstruction given in the data set's answer sheet.

Crowds can also be distracted by recent news. Most of the presentations in the additional data collection are consistent with the original results. However, there are 3 presentations directly copied from recent news articles about the 2019 Sri Lanka Easter bombings. The original data was collected between February and March in 2019, but the additional data for step 5 was collected in June 2019. We suggest that the strong social impact of the real-world attack could have distracted crowds from the analysis of a fictional mystery.

Overall, while crowd performance is negatively affected by too little or too much local context, the triple-item and all-item conditions of Study 2 suggest that the crowd can reliably synthesize distributed information and deduce hidden evidence, given the right amount and segmented local context, which varies based on the data and task design.

5 DISCUSSION AND IMPLICATIONS

In this paper, we empirically investigated crowd errors and trade-offs of additional local context in different sensemaking stages. We categorized 5 major types of local errors, and inspected how they manifest in each step with different amounts of the local context. Below we first discuss how the error propagation in crowdsourced sensemaking resembles and differs from collaborative sensemaking among experts. We then draw the design recommendations for each sensemaking stage based on the crowd reaction to propagated errors and different amounts of local context. We also examine how the experiment setup could have influenced the crowd performance and the generalizability of our findings.

5.1 Error propagation among crowds: easier hand-off but more inaccuracy blindness.

The pipeline clearly defines the step inputs and outputs, which makes it easier to distribute and aggregate crowd analysis. More importantly, it enables analyses of one step to be directly handed off to another. The step inputs and outputs serve as shared artifacts that facilitate crowd collaboration and eliminate errors due to misunderstanding and miscommunication. This allows us to focus research efforts on the analytical errors.

To encourage volunteerism and avoid social loafing due to awareness of co-workers [27], we included workflow transparency in our task instructions. Workers were told that 1) the input they are given is from previous workers (except step 1 and the GI condition), and 2) their results will be used by future workers in later analysis. The workers were not told how many co-workers were working on the same part of the data. Despite the exposure to the pipeline workflow, however, the crowd still made low-effort errors. In addition, the crowd demonstrates strong team inaccuracy blindness [25]. Experts are cautious to re-use any given intermediate analysis and usually trace back to raw material to double check the if they agree with the given input [11]. In contrast, the majority of the crowd workers take the given analysis as correct, increasing inherited errors from previous steps.

5.2 Design recommendations for each step.

5.2.1 Searching and filtering tasks need more than one documents to better judge relevance, but smaller context slice sizes produce higher quality analysis in explanations (step 1). The optimal context slice size might differ by the investigation goals and the sizes of the data set. If the amount of raw materials is too big for experts to go through, the crowd can reliably handle 3–15 short documents (word count 1200–6000), if not more. 3 workers are enough to achieve reasonable accuracy via majority vote. On the other hand, if the experts aim to use crowds for more diverse perspectives

with smaller data sets, assigning a smaller amount of data (about one document or word count 400) would encourage more thoughtful analysis that may be worth incorporating in expert analysis. In addition, when there is more than one document in each microtask, it might be helpful to require one explanation per microtask rather than per document, to balance the amount of context and the workload. The disagreement in ratings is also an indicator of crowd uncertainty [41] and reveals the more difficult part of the data. If the raw materials contain data of multiple formats and granularity, it might be helpful to have a mixed design with some big slices and some small slices to balance the workload.

In terms of the task design, our pilot and actual studies demonstrated improved crowd analysis quality when setting an explicit threshold in a numerical rating task. For example, in our task design, the crowd was asked to give a rating from 0–100. We annotated the slider with 0 (completely irrelevant), 100 (completely relevant), with a box showing the selected value by the crowd and an indication of the value. (If it is above 50, the word "relevant" is shown next to the number, otherwise, it displays the word "irrelevant.") This helps to normalize the subjective rating preference of different individuals and supports more reliable result aggregation via majority vote.

5.2.2 Reading and extracting tasks might require further task decomposition and benefits from small, focused context slices (step 2). The current design of step 2 worked well in simpler data sets with shorter (word count 50–100) and easier documents [34], but cannot handle even one document written in report language (word count 400). For more difficult input data or when the crowd workers are not guaranteed to be native speakers, it is worth the effort to incorporate related research that focuses specifically on information extraction and hire more workers for each document. Successful examples include iterative re-representation [1] and the highlighting and clipping [21].

5.2.3 Schematizing tasks can handle big context slices and benefit from more effective hand-off with additional information (step 3). Restructuring information in the documents into simpler info pieces can support larger-scale information synthesis by increasing analytical power and efficiency. When working with data of more concise formats, microtask performance benefits from bigger context slices. The crowd handled context slices of size 3 to 15 fairly well in our case studies, but the explanations became a burden. We expect the crowd can take even more than 15 data points per microtask, but it is also recommended to reduce the requirement on explanations, perhaps to one explanation per microtask, rather than per info piece. Similar to step 2, step 3 is yet another complex component that may benefit from being further modularized into sub-workflows. Schematizing is a more challenging step that connects the information foraging and synthesizing in the pipeline. It also challenges the local task with a global view more than other steps, since the organization of the information is required to make global sense and lead to further hypotheses. Successful related research could be incorporated in the pipeline to support this need and improve the task performance, such as using machine learning to pre-process info pieces and extract global patterns, and then focus crowd intelligence on edge cases [8], or having multiple iterations on crowd tagging results [10], or more effective sub-workflows and task interfaces [35].

5.2.4 Hypothesizing tasks benefit the most from bigger context slices to judge the relative likelihood and mitigate propagated errors (step 4). The crowd performance was not negatively affected by an increased number of profiles. Thus, we would expect the microtasks could handle 3-4 profiles (word count 1000), if not more. The main bottleneck of step 4 was that the correct answer was not even one of the available options. Besides improving the design of the previous step, we suggest the most effective refinement would be from a top-down path of the pipeline, with feedback provided

based on a more global understanding of the data set and current analysis, to retrieve the missing information in previous steps and redo step 4.

5.2.5 Story-telling tasks might not benefit from additional context (step 5). In step 5, the crowd was given only one winner profile in all conditions, by design. However, the amount of information contained in the profiles ranged from 88 to 335 words. Thus, the crowd is able to handle at least this amount of information and write a reasonable report. When using the mixed-quality crowd analysis results as input, step 5 does not recover from errors and can inherit or even compound previous errors by connecting irrelevant information to the relevant evidence with coincidentally overlapping entities. In addition, our case study implemented a subtask in step 4 that allows workers to optionally merge profiles. This ended up providing richer information in step 5 that avoided losing precious relevant information and helped reveal the propagated errors. Thus, we also suggest that in similar pipelined crowdsourcing systems, the benefit of including more results from the previous step outweighs the potential disadvantages of introducing more false positive noise.

5.3 Generalizability and future work

5.3.1 Performance variance due to recruiting requirements and strategies. When the scale of collaboration increases, requesters must either spend more time recruiting more workers, or lower the recruiting requirements. Spending more time is not always possible, especially in time-constrained scenarios such as intelligence analysis. This paper focuses on the problems caused by lowering the recruiting requirements. Our experiment results show that this leads to higher variance in crowd efforts and more low-effort performance. While the variance of crowd efforts is influenced by the sensemaking tasks and the context slice sizes, it does not differ by the input source.

Using time as a proxy measure of crowd effort, we found there is no significant difference in the crowd effort with crowd-generated and gold-standard input across all steps: $p_{step2} = .079; p_{step3} = .98; p_{step4} = .65; p_{step5} = .26$. On the other hand, different sensemaking steps have different variances, and the variance is also influenced by the context slice size (Fig. 6). When using uni-item context slices, steps 2, 4, and 5 have higher variance than steps 1 and 3. Step 2 and 5 microtasks require free-text responses, which invokes more thinking and requires more time for some workers. Step 4 microtasks might be difficult to rate the likelihood depending on how much information the given candidate has. When using triple-item context slices, the difference in variance is smaller across the five steps, but step 5 is still the highest.

When using all-item context slices, step 1 and 3 took many crowd workers much more time and has a higher variance than other conditions. This might be because the microtasks for the two steps are more atomic (i.e., rating each document and tagging each information piece). Some workers might need to take breaks while finishing the microtask when the number of items increases. Step 2 and 5 variance is similar to other conditions, but the time required for all workers in step 2 is a lot higher than in other conditions, indicating that the microtasks get a lot more challenging for everyone. Step 4 takes a longer time but has a smaller variance, possibly because with all the candidates in sight, there is more to read, but picking one most likely candidate becomes easier.

Finally, microtasks for step 5 in all conditions are the same, and there is no statistically significant difference in the time spent ($p_{step5} = .31$). This result suggests that the crowd performance and errors can be generalized to a bigger pool of similar crowds.

One possible way to address the high variance and low effort problem is to raise the recruiting requirements. The extreme case would be expert collaboration [19]. Another strategy suitable for novice crowd workers would be to increase the number of workers per microtask. Our case studies used 3 workers for majority-vote tasks and 2 workers for create-review tasks. The results indicate that increasing the number of workers could help converge on better analysis since poor

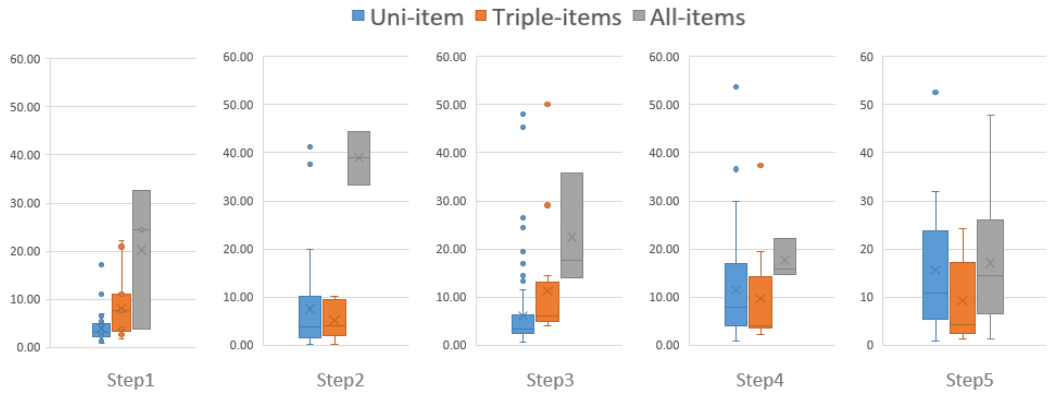

Fig. 6. Time spent on each step when given different amount of local context

performance is a smaller percentage than high-quality analysis. However, increasing the number of crowd workers might require changing the mechanism to aggregate crowd results. For example, in steps 2 and 5, an alternative would be hiring more crowd workers for the same context slice and then hire another group of workers to pick the best results. Future work is needed to investigate how the error frequency changes with different hiring requirements, and prioritize the design recommendations to eliminate the most influential errors.

5.3.2 Optimizing the analysis with task design and execution strategies. Besides our design recommendations to improve crowd analysis performance, additional subloops between connecting steps could also increase the quality of the intermediate step output. Our analysis of the explanations indicates that the crowd can sometimes recognize poor input and explain what is missing. This makes it promising to enable crowd-driven reporting of low-quality work (similar to the panic button proposed by Retelny et al. [43]) and redo the prior analysis. For example, if workers in a later step complain about the quality of input, they can be switched back to the previous step and fix the prior analysis. This forms sub-loops between connecting steps before reaching the final step. Designers might need to put a limit on the number of iterations allowed between steps, to prevent long, inefficient sub-loops from wasting crowd workers' effort. Future work is needed to further quantify the crowd's ability to identify and self-correct analysis errors. Unlike rumors on social media [2], the task domain of mystery solving may not require the same amount of incentive and background knowledge to critique each other's analyses. Showing high-quality results from other workers is effective for making workers reconsider their judgments [31], and also brings in additional local context. However, this approach will make each worker stay for a longer session in each microtask, and require a mechanism to automatically pick the high-quality results.

5.3.3 Top-down refinement with a global understanding of the data and crowd analysis. Given the unpredictable and challenging nature of exploratory sensemaking, we see the potential benefit of a top-down path of the pipeline to complement the bottom-up analysis. There is always a limit to how much context a local view of the data can synthesize, and even with a completely reasonable local analysis, the key evidence can be so well hidden that it cannot be revealed without iterative analysis. By the end of the pipeline execution, the crowd has produced a more detailed global understanding of the information available, which could help an expert prioritize and focus on the important evidence. Experts [15] or the crowd [36] could "debug" the pipeline and identify where mistakes were made during the sensemaking process. The pipeline structure provides built-in

provenance analysis that traces the information and insights between steps, and makes it easy to examine the initial crowd analysis, evaluate analysis quality, identify information holes or logic flaws, trace back to errors in a top-down manner, and guide the refinement of the previous analyses with feedback. In future work, we plan to continue to explore how the pipeline can also support debugging and refining previous imperfect analysis. The errors and bottlenecks we classified in this paper can serve as a checklist to identify errors in previous crowd analysis. More importantly, the analysis provenance that connects the intermediate results and traces information flow will be critical for obtaining a big picture of the mystery and applying the newly acquired knowledge from the previous analysis in the refinement process.

6 CONCLUSION

Crowdsourced sensemaking has demonstrated impressive potential across a range of complex tasks and domains, but most systems still require expert intervention because of crowd errors. This paper studies the errors and bottlenecks in a crowdsourced sensemaking pipeline that connects multiple sub-processes without expert intervention. Our analysis shows how chaining together mixed-quality crowd analyses can inherit or even compound previous errors. Wrongly retrieved irrelevant documents or useless information pieces can further pollute later analysis. Surprisingly, both false positive and false negative errors were mitigated to some extent without external mediation. We attribute this to the design of the pipeline that condenses information from documents to info pieces to profiles as the analysis progresses to high-level goals. In addition, our results also demonstrated that increasing the amount of local context can facilitate synthesizing information from distributed sources, but introduce additional workload that can overwhelm workers and harm the overall analysis. We also proposed design recommendations for supporting complex crowdsourced sensemaking, both within individual steps and across the broader pipeline.

ACKNOWLEDGMENTS

This research was supported in part by NSF under grants IIS-1527453, IIS-1651969, and IIS-1447416.

A APPENDICES

A.1 Gold standard analysis and decision rationales

We describe the final gold standard output for each step and the decision rationales in the generating process.

10 relevant documents. We trace the source documents of each piece of key evidence used in the Wigmore chart and mark them as the *gold-standard relevant documents*.

19 most important info pieces. The answer sheet listed 20 key evidence parsed from the documents that contribute to solving the mystery. The key evidence includes both direct evidence and supporting clues, some are inferences that cannot be directly generated from one document only. In order to develop a baseline performance, we focus on only the direct evidence and assume the condition where each microtask only has access to one document. We break down the direct evidence used in the Wigmore chart into simpler sentences that 1) can be generated from one single document, 2) are structured as "who, what, where, when" as much as possible. This is to match the baseline condition of the step 2 microtasks where each worker only has access to one document. The resulting 19 info pieces are *gold-standard info pieces*.

5 location profiles. The answer sheet organizes the key evidence vertically by deductive reasoning, but not horizontally into profile schemata. We manually tag the single-document gold standard info pieces by whether they contain information about any known "*Terrorist*", the C-4 explosive "*Weapon*", and the planned attack "*Time*". Some information pieces need to be put together with

others to reveal meaningful evidence; we mark these as hidden tags. For possible target locations, we extract all the locations mentioned in the information pieces. It's worth noting that the names and resolutions of the locations are tricky. For example, the correct answer "*NYSE (New York Stock Exchange)*" is also in New York City. Since both "NYC" and "New York City" appeared in the documents by themselves, we extracted New York City [NYC] as one of the possible target locations. Putting the info pieces about each location together is the *gold-standard location profiles*.

Likelihood ranking of profiles. The answer sheet also doesn't rank the likelihood of all locations mentioned in the dataset. We rank the locations by 1) their geographical distance to the real target location (NYSE), 2) the number of terrorist activities, and 3) the minimum depth of its mention in the Wigmore chart. We rank NYC, the lower resolution location containing the correct answer NYSE, both as the first place. The final ranking of likelihood is New York Stock Exchange [NYSE] = New York City [NYC] > Empire State Vending Services [ESVS] > carpet store in North Bergen, NJ > Springfield, VA.

Final presentation. The answer sheet contains a conclusion statement and almost two pages of an article detailing the process of terrorist coordinating the attack. The article connects the facts and develops hypotheses but also involves domain knowledge and speculative details (not mentioned nor derived from the given documents, e.g. "Several days before the delivery of the vending machine containing the bomb, Alwan goes to the NYSE to fill a coffee, tea, hot chocolate machine and, in the process, disrupts its functioning.") Thus, we define the *gold-standard final presentation* not as a pragraph but regulates the most important facts to mention, namely:

- Joed Shearper is 1) a terrorist 2) with explosive training and 3) has access (to the vending machines in) the New York Stock Exchange under 4) the alias Devon Citopper.
- Cedric Whappadder 5) stores buckets of C-4 explosives in his (cover-up) carpet shop (in North Bergen, New Jersey).
- Joed Shearper (alias Devon Citopper) and Irving Sprunkiddle (alias Virgil Sneworf) are both terrorists. They live in the same apartment. One of them 6) picked up C-4 explosives from Cedric Whappadder's carpet shop.

Crowd generated presentation with gold-standard input. With the gold-standard profile, the crowd workers were able to reconstruct the crime from the given profile into a cohesive narrative. Below is the final presentation by the crowd, with all 6 matched facts from our gold-standard analysis boldfaced:

> The New York Stock Exchange (NYSE) may be under imminent threat of attack by several known terror suspects. One suspect, a Saudi **explosives expert** from named Joed Shearper (**aka "Devon Citopper"**), possesses a **NYSE** vendor's ID through his employment as a **vending machine operator**, providing the group both clearance and expertise for planning and conducting an attack. Alwan lives at the **same address** (2462 Myrtle Ave, Apt 307, Queens) and works at the same vending machine company as a known Pakistani **Taliban member**, Irving Sprunkiddle (aka "Virgil Sneworf"). A third suspect, **Cedric Whappadder**, as recently as April 26, 2003, **had access to C-4 explosives** that could be employed in a terror incident. Hallak manages a **carpet store** in North Bergen, NJ; after several previous calls, on April 22 a caller from a number associated with Alwan and Albakri's Queen's address (718-352-8479) stated that he would **pick up a previously ordered carpet from the store's location**. Hallak has since vanished. Thus, this terror cell has the capability to attack the NYSE as believed.

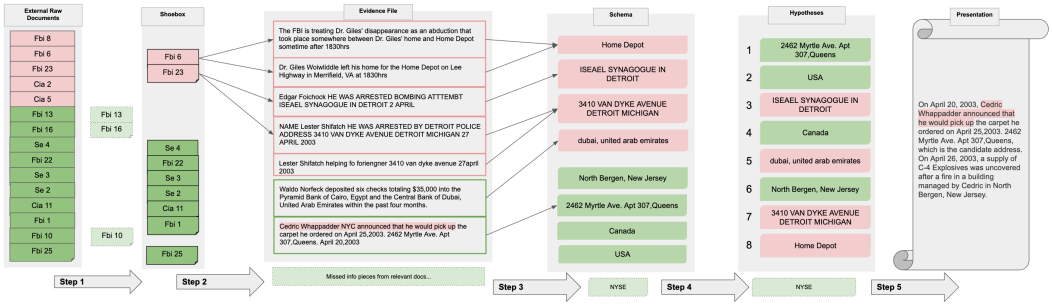

Fig. 7. Error Propagation in Uni-item Condition: the pink colored items are irrelevant documents, info pieces, and location profiles; the green colored ones are relevant to solving the mystery.

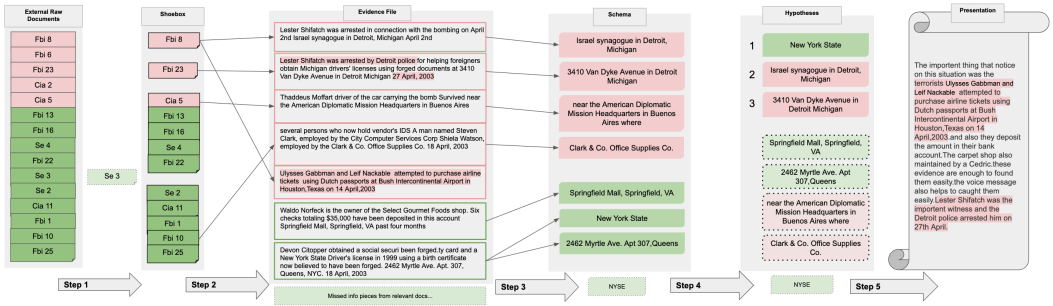

Fig. 8. Error propagation in triple-item condition

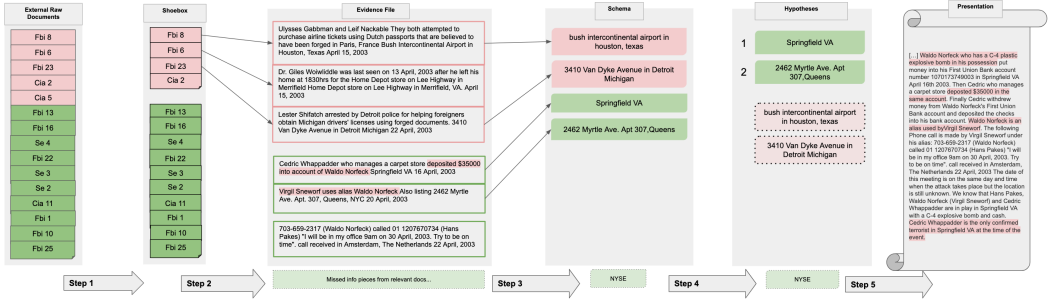

Fig. 9. Error propagation in all-item condition

A.2 Error propagation shown in diagrams

We present examples of error propagation in each step under the uni-item (Fig. 7), triple-item (Fig. 8), and all-item (Fig. 9).

REFERENCES

[1] Paul André, Aniket Kittur, and Steven P. Dow. 2014. Crowd Synthesis: Extracting Categories and Clusters from Complex Data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 989–998. https://doi.org/10.1145/2531602.2531653

[2] Ahmer Arif, John J Robinson, Stephanie A Stanek, Elodie S Fichet, Paul Townsend, Zena Worku, and Kate Starbird. 2017. A closer look at the self-correcting crowd: Examining corrections in online rumors. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 155–168.

[3] Tal August and Katharina Reinecke. 2019. Pay Attention, Please: Formal Language Improves Attention in Volunteer and Paid Online Experiments. (2019).

[4] Eric A Bier, Stuart K Card, and John W Bodnar. 2008. Entity-based collaboration tools for intelligence analysis. In *2008 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 99–106.

[5] Alessandro Bozzon, Marco Brambilla, and Stefano Ceri. 2012. Answering Search Queries with CrowdSearcher. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 1009–1018. https://doi.org/10.1145/2187836.2187971

[6] Lauren Bradel, Alex Endert, Kristen Koch, Christopher Andrews, and Chris North. 2013. Large high resolution displays for co-located collaborative sensemaking: Display usage and territoriality. *International Journal of Human-Computer Studies* 71, 11 (2013), 1078–1088.

[7] Joel Chan, Steven Dang, and Steven P. Dow. 2016. Improving Crowd Innovation with Expert Facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1223–1235. https://doi.org/10.1145/2818048.2820023

[8] Joseph Chee Chang, Aniket Kittur, and Nathan Hahn. 2016. Alloy: Clustering with Crowds and Computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3180–3191. https://doi.org/10.1145/2858036.2858411

[9] Wen-Huang Cheng and David Gotz. 2009. Context-based Page Unit Recommendation for Web-based Sensemaking Tasks. In *Proceedings of the 14th International Conference on Intelligent User Interfaces (IUI '09)*. ACM, New York, NY, USA, 107–116. https://doi.org/10.1145/1502650.1502668

[10] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. 2013. Cascade: Crowdsourcing Taxonomy Creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 1999–2008. https://doi.org/10.1145/2470654.2466265

[11] George Chin, Jr., Olga A. Kuchar, and Katherine E. Wolf. 2009. Exploring the Analytical Processes of Intelligence Analysts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 11–20. https://doi.org/10.1145/1518701.1518704

[12] H. Chung, S. P. Dasari, S. Nandhakumar, and C. Andrews. 2017. CRICTO: Supporting Sensemaking through Crowdsourced Information Schematization. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. 139–150. https://doi.org/10.1109/VAST.2017.8585484

[13] Haeyong Chung, Seungwon Yang, Naveed Massjouni, Christopher Andrews, Rahul Kanna, and Chris North. 2010. Vizcept: Supporting synchronous collaboration for constructing visualizations in intelligence analysis. In *2010 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 107–114.

[14] Gianluca Demartini, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. 2012. ZenCrowd: Leveraging Probabilistic Reasoning and Crowdsourcing Techniques for Large-scale Entity Linking. In *Proceedings of the 21st International Conference on World Wide Web (WWW '12)*. ACM, New York, NY, USA, 469–478. https://doi.org/10.1145/2187836.2187900

[15] 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 (CSCW '12)*. ACM, New York, NY, USA, 1013–1022. https://doi.org/10.1145/2145204.2145355

[16] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 80–88. http://dl.acm.org/citation.cfm?id=1866696.1866709

[17] 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 (CHI '12)*. ACM, New York, NY, USA, 247–256. https://doi.org/10.1145/2207676.2207711

[18] Steven Gottlieb, Sheldon I Arenberg, Raj Singh, et al. 1994. *Crime analysis: From first report to final arrest*. Alpha Publishing Montclair, CA.

[19] Nitesh Goyal and Susan R Fussell. 2016. Effects of sensemaking translucence on distributed collaborative analysis. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 288–302.

[20] Nitesh Goyal, Gilly Leshed, and Susan R. Fussell. 2013. Effects of Visualization and Note-taking on Sensemaking and Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2721–2724. https://doi.org/10.1145/2470654.2481376

[21] 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 (CHI '16)*. ACM, New York, NY, USA, 2258–2270. https://doi.org/10.1145/2858036.2858364

[22] Richards J Heuer. 1999. *Psychology of intelligence analysis*. Jeffrey Frank Jones.

[23] Shih-Wen Huang and Wai-Tat Fu. 2013. Don'T Hide in the Crowd!: Increasing Social Transparency Between Peer Workers Improves Crowdsourcing Outcomes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 621–630. https://doi.org/10.1145/2470654.2470743

[24] Irving Lester Janis and Irving Lester Janis. 1982. *Groupthink: Psychological studies of policy decisions and fiascoes*. Vol. 349. Houghton Mifflin Boston.

[25] Ruogu Kang, Aimee Kane, and Sara Kiesler. 2014. Teammate Inaccuracy Blindness: When Information Sharing Tools Hinder Collaborative Analysis. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '14)*. ACM, New York, NY, USA, 797–806. https://doi.org/10.1145/2531602.2531681

[26] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S Bernstein. 2017. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, 233–245.

[27] Peter Kinnaird, Laura Dabbish, and Sara Kiesler. 2012. Workflow transparency in a microtask marketplace. In *Proceedings of the 17th ACM international conference on Supporting group work*. ACM, 281–284.

[28] Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. 2012. CrowdWeaver: Visually Managing Complex Crowd Work. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work (CSCW '12)*. ACM, New York, NY, USA, 1033–1036. https://doi.org/10.1145/2145204.2145357

[29] Aniket Kittur, Lixiu Yu, Tom Hope, Joel Chan, Hila Lifshitz-Assaf, Karni Gilon, Felicia Ng, Robert E Kraut, and Dafna Shahaf. 2019. Scaling up analogical innovation with crowds and AI. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1870–1877.

[30] Gary Klein, Jennifer K Phillips, Erica L Rall, and Deborah A Peluso. 2007. A data-frame theory of sensemaking. In *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*. New York, NY, USA: Lawrence Erlbaum, 113–155.

[31] Masaki Kobayashi, Hiromi Morita, Masaki Matsubara, Nobuyuki Shimizu, and Atsuyuki Morishima. 2018. An empirical study on short-and long-term effects of self-correction in crowdsourced microtasks. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

[32] Anand Kulkarni, Matthew Can, and Björn Hartmann. 2012. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the acm 2012 conference on computer supported cooperative work*. ACM, 1003–1012.

[33] Walter S. Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P. Bigham, and Michael S. Bernstein. 2015. Apparition: Crowdsourced User Interfaces That Come to Life As You Sketch Them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1925–1934. https://doi.org/10.1145/2702123.2702565

[34] Tianyi Li, Kurt Luther, and Chris North. 2018. CrowdIA: Solving Mysteries with Crowdsourced Sensemaking. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 105 (Nov. 2018), 29 pages. https://doi.org/10.1145/3274374

[35] Kurt Luther, Nathan Hahn, Steven P Dow, and Aniket Kittur. 2015. Crowdlines: Supporting synthesis of diverse information sources through crowdsourced outlines. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

[36] Kurt Luther, Jari-Lee Tolentino, Wei Wu, Amy Pavel, Brian P. Bailey, Maneesh Agrawala, Björn Hartmann, and Steven P. Dow. 2015. Structuring, Aggregating, and Evaluating Crowdsourced Design Critique. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 473–485. https://doi.org/10.1145/2675133.2675283

[37] Filippo Menczer. 2016. The Spread of Misinformation in Social Media. 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, Switzerland, 717–717. https://doi.org/10.1145/2872518.2890092

[38] Vikram Mohanty, David Thames, and Kurt Luther. 2018. Photo Sleuth: Combining Collective Intelligence and Computer Vision to Identify Historical Portraits. In *ACM Conference on Collective Intelligence (CI 2018)*.

[39] Meredith Ringel Morris, Jeffrey P. Bigham, Robin Brewer, Jonathan Bragg, Anand Kulkarni, Jessie Li, and Saiph Savage. 2017. Subcontracting Microwork. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 1867–1876. https://doi.org/10.1145/3025453.3025687

[40] Meredith Ringel Morris, Jeffrey P Bigham, Robin Brewer, Jonathan Bragg, Anand Kulkarni, Jessie Li, and Saiph Savage. 2017. Subcontracting microwork. In *Proceedings of the 2017 CHI conference on human factors in computing systems*. ACM, 1867–1876.

[41] Susannah BF Paletz, Joel Chan, and Christian D Schunn. 2016. Uncovering uncertainty through disagreement. *Applied Cognitive Psychology* 30, 3 (2016), 387–400.

[42] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[43] 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

[44] Heinz Schmitz and Ioanna Lykourentzou. 2018. Online Sequencing of Non-Decomposable Macrotasks in Expert Crowdsourcing. *Trans. Soc. Comput.* 1, 1, Article 1 (Jan. 2018), 33 pages. https://doi.org/10.1145/3140459

[45] Nikhil Sharma. 2010. Sensemaking Handoffs: Why? How? and When? (2010).

[46] Maoyuan Sun, Lauren Bradel, Chris L. North, and Naren Ramakrishnan. 2014. The Role of Interactive Biclusters in Sensemaking. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 1559–1562. https://doi.org/10.1145/2556288.2557337

[47] Maoyuan Sun, Peng Mi, Chris North, and Naren Ramakrishnan. 2015. Biset: Semantic edge bundling with biclusters for sensemaking. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 310–319.

[48] 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*.

[49] Melissa A. Valentine, Daniela Retelny, Alexandra To, Negar Rahmati, Tulsee Doshi, and Michael S. Bernstein. 2017. Flash Organizations: Crowdsourcing Complex Work by Structuring Crowds As Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 3523–3537. https://doi.org/10.1145/3025453.3025811

[50] Vasilis Verroios and Michael S Bernstein. 2014. Context trees: Crowdsourcing global understanding from local views. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

[51] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2012. Strategies for Crowdsourcing Social Data Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 227–236. https://doi.org/10.1145/2207676.2207709

[52] William Wright, David Schroh, Pascale Proulx, Alex Skaburskis, and Brian Cort. 2006. The Sandbox for Analysis: Concepts and Methods. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 801–810. https://doi.org/10.1145/1124772.1124890

[53] H. Xie and J. C. S. Lui. 2018. Incentive Mechanism and Rating System Design for Crowdsourcing Systems: Analysis, Tradeoffs and Inference. *IEEE Transactions on Services Computing* 11, 1 (Jan 2018), 90–102. https://doi.org/10.1109/TSC.2016.2539954

[54] Lixiu Yu, Aniket Kittur, and Robert E. Kraut. 2016. Encouraging "Outside- The- Box" Thinking in Crowd Innovation Through Identifying Domains of Expertise. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 1214–1222. https://doi.org/10.1145/2818048.2820025

[55] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human Computation Tasks with Global Constraints. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 217–226. https://doi.org/10.1145/2207676.2207708

[56] Jian Zhao, Michael Glueck, Petra Isenberg, Fanny Chevalier, and Azam Khan. 2017. Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 340–350.

Received April 2019; revised June 2019; accepted August 2019