CrowdIA: Solving Mysteries with Crowdsourced Sensemaking

TIANYI LI, Virginia Tech, USA
KURT LUTHER, Virginia Tech, USA
CHRIS NORTH, Virginia Tech, USA

The increasing volume of text data is challenging the cognitive capabilities of expert analysts. Machine learning and crowdsourcing present new opportunities for large-scale sensemaking, but we must overcome the challenge of modeling the overall process so that many distributed agents can contribute to suitable components asynchronously and meaningfully. In this paper, we explore how to crowdsource the sensemaking process via a pipeline of modularized steps connected by clearly defined inputs and outputs. Our pipeline restructures and partitions information into "context slices" for individual workers. We implemented CrowdIA, a software platform to enable unsupervised crowd sensemaking using our pipeline. With CrowdIA, crowds successfully solved two mysteries, and were one step away from solving the third. The crowd's intermediate results revealed their reasoning process and provided evidence that justifies their conclusions. We suggest broader possibilities to optimize each component, as well as to evaluate and refine previous intermediate analyses to improve the final result.

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; Investigation; Crowdsourcing

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

1 INTRODUCTION

Intelligence analysts working to prevent terrorist attacks and preserve national security have access to an unprecedented wealth of data about persons of interest. Yet, events such as the September 11th, 2001 terrorist attacks and the miscall on weapons of mass destruction in Iraq — "the two major U.S. intelligence failures of this century" [14] — illustrate the difficulties that even experienced professionals face in analyzing this data, and the high-stakes consequences of failure. Traditional intelligence analysis faces the ongoing challenges of distinguishing crucial information from noise and dealing with incomplete pieces. Marshaling and synthesizing heaps of evidence is especially difficult. As observed by Wright et al., "To get the big picture by looking at many pages of text, the analyst relies heavily on memory to connect the dots" [64]. In this paper, we focus on a class of

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

2573-0142/2018/11-ART105 $15.00
https://doi.org/10.1145/3274374

problems that involve solving mysteries, in which analysts must sort through many snippets of textual information to identify a latent plot, such as a suspect or target.

The sensemaking process of intelligence analysts has been modeled as an iterative loop of interdependent steps that involves foraging for relevant information and synthesizing it into credible hypotheses [48]. However, the process ultimately relies on the experts to make ad hoc decisions on which steps to conduct and how to order the workflow. Making sense of massive amounts of complex information that comes from various sources and discerning critical patterns and anomalies is a cognitively demanding task for individual experts.

Several areas of research attempt to support this sensemaking process. Visual analytics tools have been developed to leverage technological support for some specific steps [20, 25]. Collaborative sensemaking among experts can bring together diverse expertise and perspectives, but often suffers from biases and inefficiencies [24, 53]. Machine learning can process the data to provide starting points for analysts [65], but deciphering the rich information encoded in text data is still AI-hard [67].

Alternatively, crowdsourcing is a powerful new paradigm that augments individual human intelligence at large scale, showing potential to bridge the gaps between the information overload and the limited cognitive capacity of individual experts. By modularizing a big problem into many small, manageable problems and aggregating results from small solutions into a big meaningful result, prior research has successfully used crowds for complex data analysis tasks like taxonomy generation [11], bottom-up qualitative analysis [1], and organizing online information [23].

However, sensemaking with large amounts of data introduces two major problems for modularization. First, the entire sensemaking process, as required for solving mysteries, is a highly integrated cognitive activity composed of iterative information foraging, schematizing, and synthesizing, that is difficult to formalize into a workflow of microtasks for novice crowd workers [55]. Second, sensemaking requires a holistic view of the data, making it difficult to subdivide the data into small local slices while preserving global data context for crowd workers. To overcome these problems, we need a model that adequately captures and translates the expert sensemaking process for large crowds of transient novice workers.

With this motivation, we address the following research questions in this paper:

RQ1. To support crowds, how can we formally modularize the sensemaking process into a series of steps that each defines the information needs (*Step Input*) and intermediate analysis results (*Step Output*)?

RQ2. Within each step, how do we slice the *Step Input* into contextualized microtasks for individual crowd workers, and aggregate the local analysis results into *Step Output*?

RQ3. How well do crowds perform in solving mysteries with the modularized sensemaking process, and specifically, how do crowds perform in each step?

To address these questions, we designed a pipeline to support crowdsourced sensemaking informed by preliminary studies with both individual users and crowds. We implemented the pipeline as a software platform called CrowdIA. We evaluated the pipeline by deploying CrowdIA on Amazon Mechanical Turk (MTurk) to guide crowds in solving mysteries. In these empirical studies, the crowds successfully solved easy and moderate mysteries, and were one step away from solving a difficult mystery. Our main contributions are as follows:

(1) We modularized the sensemaking loop into a pipeline with clearly defined *Step Inputs* and *Step Outputs*, such that each step can be separately investigated by crowd workers.

(2) We designed methods to (a) restructure and distribute *Step Inputs* into cohesive "context slices" as local task inputs for each individual crowd worker to contribute meaningfully, and (b) to combine crowd results into the corresponding *Step Outputs*.

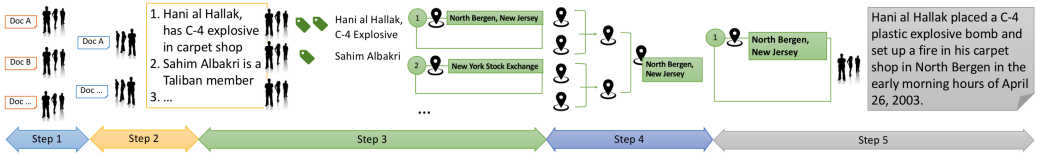

Fig. 1. Overview of our pipeline for crowdsourced analysis of the difficult *Sign of Crescent* dataset.

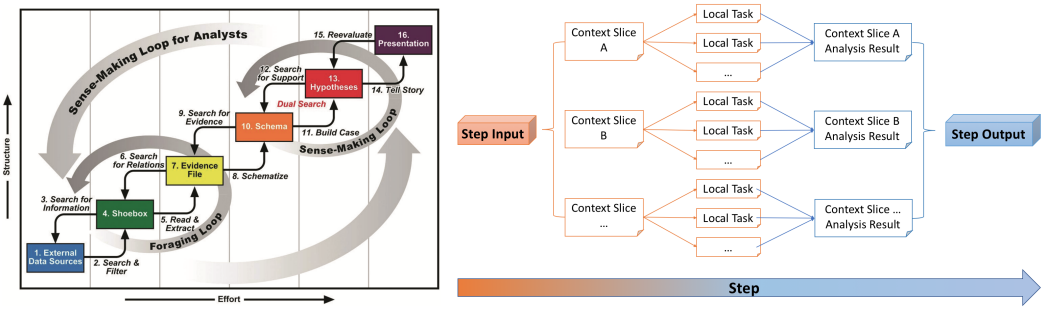

Fig. 2. Sensemaking loop [48], image source [15] (left). Example of how each component is modularized in the pipeline (right).

(3) We developed CrowdIA, a system that facilitates execution of our pipeline, and evaluated its feasibility in enabling novice, transient crowd workers to solve mysteries.

We further suggest that the modularized pipeline can open up the sensemaking process as a test-bed environment for researchers to design and evaluate novel interfaces for each step.

2 RELATED WORK

Intelligence analysts make sense of large amounts of information by iteratively foraging for relevant source data (*1. search and filter*), extracting useful information (*2. read and extract*), organizing and re-representing the information with their mental models (*3. schematize*), developing hypotheses from different perspectives (*4. build case*), and deciding on the best explanation (*5. tell story*). Pirolli et al. [48] informally modelled this process as a main loop composed of an information foraging sub-loop and a sensemaking sub-loop, each iterating on smaller intertwined steps (Fig. 2 left). HCI research on improving sensemaking in different domains and settings, as we review in later sections, can be considered as fitting in different parts of the sensemaking loop.

2.1 Collaborative Sensemaking: Communication and Hand-off

Analysts collaborate on sensemaking tasks by using visual analytics tools to annotate, link, and spatially organize documents and named entities; forage for information; identify topics; and plan more in-depth analysis [5, 13, 19, 21]. One key challenge is to collect and mentally compare the relevant information scattered across many locations. Metaphors like folders and bookmarks are used to organize fragments of information to create task-specific contexts [10].

Another challenge is the hand-off of intermediate results between analysts. Given the non-routine nature and black-box mental models of analysts, hand-off in collaborative sensemaking is seldom successful unless it happens very early (transfer) or very late (referral) in the process [49]. Analysts examine the entities in the documents from different perspectives (e.g., categories, document contents) [4, 52] and data structures (concept map [13], bicluster [54]). A visualization of data links

is more effective as an intermediate analysis artifact than a notepad of annotations [27]. Such hand-offs still rely on a shared understanding of the schema and visual layout of the information [2, 21, 50]. To help establish the shared understanding, Zhao et al. [69] developed Knowledge-Transfer Graphs that automatically capture, encode, and streamline analysts' interactions 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. To address this issue, Goyal et al. [26] proposed a social translucence interface to raise analysts' self-awareness, shedding light on when and how distributed collaborative pairs share intermediate hypotheses to enhance the analysis quality. However, these and most other collaborative sensemaking projects focus mainly on certain sensemaking components [30, 61, 63] or assume the same analysts are involved in the entire session [12, 16, 56].

We explore how sensemaking process can be modularized so that intermediate analysis results can be passed to subsequent analysts with minimal hand-off learning curve. Small group collaboration relies heavily on analysts spending enough time and attention understanding and building on previous work by others, an approach that generally does not scale well for crowdsourcing. We next consider existing crowdsourcing approaches in complex sensemaking work.

2.2 Crowdsourcing Complex Cognitive Tasks: Large-Scale Coordination

Researchers have found success in systems that leverage crowdsoucing for information synthesis guided by experts. Crowds have improved the quality of selected components of sensemaking. For example, Wang et al. [60] use crowds to verify and remove duplicated database records identified by computers, and Soylent [3] used crowds to shorten and proofread text as part of a word processor.

As tasks become more complex, processes become more interdependent, and workflow and task designs play a more important role in crowd sensemaking. Current research relies on experts to provide ideal input and generate a guideline [8, 33] or specific goals [29, 58] to address the tension between local micro tasks and the global view of the whole dataset, balancing structure, flexibility and expert guidance [36]. Crowd Synthesis [1] scaffolds expertise for novice crowds via a classification-plus-context approach, where crowds first re-represent the text data then iteratively elicit categories. We build on the re-representation stage in the design of Step 2 of our pipeline to extract key information pieces from the documents and simplify the clustering stage in Step 3 with predefined tags to facilitate later hypothesize. Crowdlines [23] found that exposing individual crowd workers to more information (high context) and less guidance (low structure) and using tournament-style workflows yields higher quality results, faster completion times, and higher completion rates in topic merging tasks. We take inspiration from Crowdlines' synthesis interfaces and parallelized, hierarchical workflows. But rather than synthesizing information into a summary of a given topic, we explore the possibility of leveraging crowds to solve mysteries, which requires discovering less obvious connections and developing hypotheses to uncover the hidden truth.

Parikh et al.'s [47] notion of human-debugging, originally applied to computer vision research, takes out each specific component in the computational system's pipeline and uses human subjects to transform the same input given to machine into the output. Drawing inspiration from this paradigm, and the CrowdForge framework for complex crowd tasks [37], **we modularize the components of the sensemaking loop [48] and decompose each step input into context slices so that distributed novice crowd workers can contribute meaningfully**.

Below we describe the process and challenges involved in designing the pipeline, particularly focusing on issues central to provenance and hand-off within and between sensemaking components.

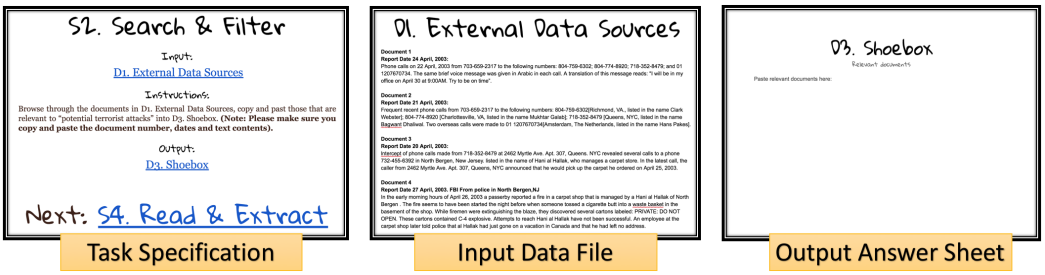

Fig. 3. Prototype Interface: Example task specification interface, input data file, and output answer sheet.

3 DESIGN PROCESS: PRELIMINARY STUDIES

The above related work suggests two hard problems in crowdsourcing the entire sensemaking process. One problem (RQ1) involves identifying information needs (*Step Input*) and intermediate analysis results (*Step Output*) of each step in the analysis, such that each step has a clear goal and progresses along the sensemaking process while preserving provenance. A second problem (RQ2) involves distributing the work required to analyze each *Step Input* among crowd workers and meaningfully aggregate local results as the *Step Output*. Below, we describe our approach to addressing these problems, leading to the final pipeline design implemented in CrowdIA.

3.1 RQ1: Identifying Step Inputs and Outputs with Individual Participants

The sensemaking loop is a "broad brush description" [48] of an expert's cognitive process of information transformation. The boundary of each step is not clear-cut and the expert might skip steps. In order to support large-scale distributed sensemaking, a robust pipeline has to explicitly represent each intermediate analysis result. Our first preliminary study sought to uncover the information needs (*Step Input*) and intermediate results (*Step Output*) of each step when individual analysts rigidly follow the sensemaking loop (RQ1).

We designed a series of prototype interfaces to sequentially guide individuals' exploratory text analysis and specified the intermediate analysis results at five different steps. These five steps were based directly on the forward arrows in Pirolli's sensemaking loop (labeled 2, 5, 8, 11, and 14 in Fig. 2): 1) *Search and Filter* relevant documents; 2) *Read and Extract* key information pieces; 3) *Schematize* information pieces into meaningful node-link graphs; 4) *Build a Case* with graphs to form hypotheses; and 5) *Tell a Story* with winning hypotheses in a narrative conclusion. The prototype served as a common artifact that participants could jointly author [21]. Each step was comprised of two information items (input data file and output answer sheet) and one task specification (Fig. 3). The input data file listed the contents of the input data. The output file was a blank sheet to fill in. The task specification gave the task name, instructions, the name of available input, the name of expected output, and the name of the next step. We designed the interfaces as lightweight prototypes [28] in Google Docs and Google Slides. Each participant completed Step 1 through 5, and then had the option to go back and refine their intermediate analysis in a second path edit. Participants attempted to solve mysteries using easy, moderate, and difficult data sets.

Provenance. Several participants found some of the sensemaking steps unnecessary when solving the easy mystery, while others appreciated the step-by-step pipeline to organize their thoughts and consolidate their analyses. All participants solving the difficult mystery needed to go back and refine their previous step output. Most participants added more documents and information pieces (Fig. 4), and modified their original node-link graphs (Fig. 5, left) after finishing the first path. They found the modularized steps helpful to keep track of their analysis process, and to trace back

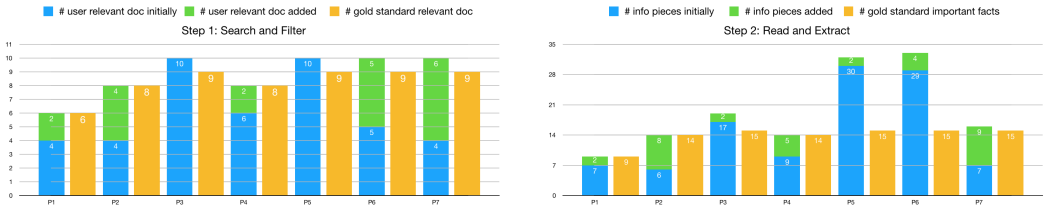

Fig. 4. Intermediate analysis results by individual participants in Steps 1 and 2. Blue bars are their initial results, green bars are their second path edits, and orange bars are correct answers contained in each participant's results.

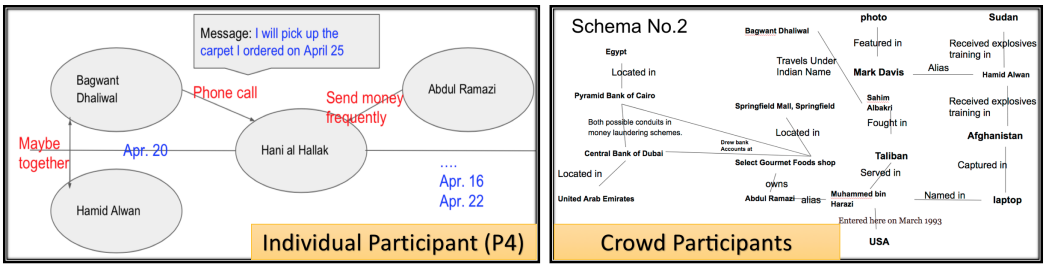

Fig. 5. Example schemas created by individual participants (left) and crowd workers (right) in the preliminary studies.

and improve certain intermediate results when refining their final conclusion. *Lesson*: Modularized steps with clearly defined intermediate analysis results help ensure analysis credibility and support backtracking when refining the previous analysis.

3.2 RQ2: Distributing Input and Aggregating Output with Crowd Workers

Given the *Step Input* and *Step Output* of each step, our second preliminary study explored how to distribute the *Step Input* among crowd workers and aggregate local analysis results into *Step Output* (RQ2). We deployed the same prototype interfaces on Amazon Mechanical Turk and added separate input and output interfaces for each crowd worker. After each step, we manually copied and pasted the crowd results to fill in the input interface for the next step.

Handoff Within Steps. We compared parallel and iterative human computation processes [40] for crowd collaboration within the steps. We found that a parallel approach allowed crowds to search and filter relevant documents (Step 1), but the same approach led to duplicated evidence extracted from the documents (Step 2). Workers rarely revised previous schema or hypotheses created by others and tended to create their own new ones (Steps 3, 4). However, they refined and improved previous narrative conclusions written by other workers (Step 5). *Lesson*: A parallel process works better for decision tasks, and an iterative process works better for information extraction and synthesis tasks. When new structures are introduced to reorganize the current information (Step 3), previous analysis results become difficult to understand and build on.

Handoff Between Steps. We observed that crowds were most challenged to understand schemas created by previous workers. The hypotheses these workers developed did not include key findings from the schema (Fig. 5, right) or had wrong or contradictory information (Step 4). *Lesson*: The node-link graph is difficult for later crowd workers to understand since it does not provide an obvious starting point and represents many implicit, personal thought processes. We suggest a more

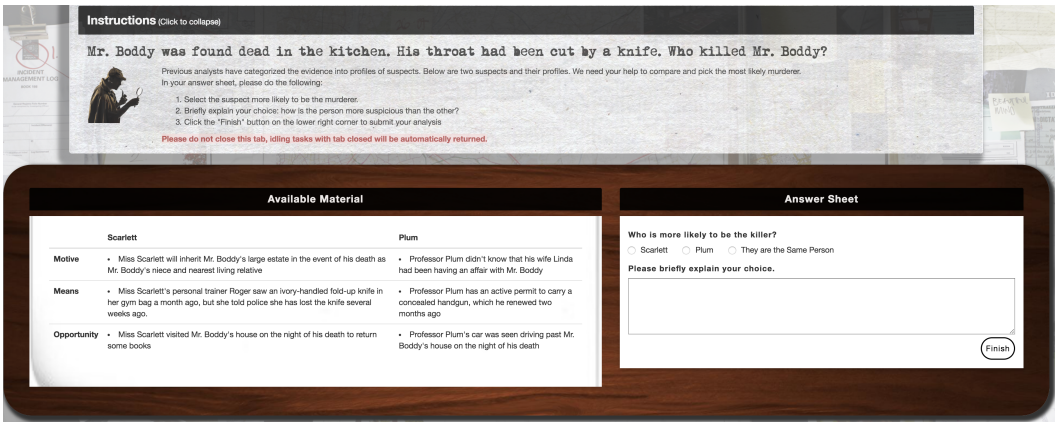

Fig. 6. Example crowd worker interface for Step 4. On the top are task instructions including the global context (first line), task overview (first paragraph), and action items (bullet points). On the bottom left is one context slice as local task input (available material). On the bottom right is the local task output where crowds fill in and submit their analysis.

effective approach to crowdsourced schematization is to use a less abstract and more well-defined structure, such as workers assigning appropriate pre-defined tags to information pieces, which can then be organized accordingly.

4 THE CROWDIA SYSTEM

Guided by the lessons learned from the preliminary studies, we refined our pipeline and implemented a web-based application, CrowdIA, to automate its facilitation. Fig. 6 shows an example interface from the system.

4.1 Implementation

CrowdIA is implemented with the Django web framework and deployed on the Heroku cloud platform.

The back end is implemented in Python with a PostgreSQL database and uses the boto3 API to communicate with MTurk. It is responsible for 1) partitioning current step input into context slices; 2) sending context slices and corresponding contents to the front end when a worker accepts a task; 3) receiving and saving local task results (encoded as JSON strings) into the database when a worker submits a task result; 4) keeping track of local task status by detecting submitted, returned, or abandoned tasks; 5) aggregating local task results into *Step Output*; 6) keeping track of step completion status; 7) transforming current completed *Step Output* into the next *Step Input*; and 8) automatically releasing corresponding tasks to MTurk.

The front end is implemented with the Bootstrap UI framework in HTML, CSS, and JavaScript / JQuery. It is responsible for 1) rendering the UI design; 2) supporting user interaction (e.g., when extracting information pieces, a crowd worker is required to put the "who, what, where, when" elements of an information piece into four separate blanks); 3) validating results to ensure work quality; and 4) sending requests to the server to fetch task content and submit analysis results (via AJAX and JSON strings).

In the following sections, we first explain the overall pipeline structure, and then focus on describing the different input, output, context slices and aggregation mechanism of each step.

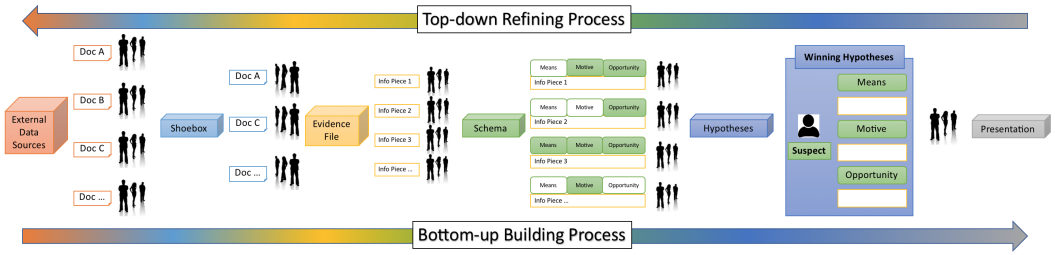

Fig. 7. Modularized sensemaking pipeline. Step 1 searches external data sources for relevant documents. Step 2 extracts important information pieces from the relevant documents. Step 3 organizes information pieces into profile schemas. Step 4 compares and merges schemas to develop hypotheses. Step 5 synthesizes the best hypotheses as the final presentation.

4.2 Pipeline Structure and Step Definition

The CrowdIA pipeline is composed of five *Step*s, corresponding to the five data transformation processes in the sensemaking loop [48]. Each *Step* is a dedicated module defined by *Step Input* and *Step Output* (Fig. 2 right). Each *Step Output* equals the *Step Input* of the next *Step*.

Each *Step Input* is restructured and partitioned into multiple *Context Slices*, each of which is a meaningful subset of *Step Input* and contains semantically cohesive data. The *Context Slice Results* are aggregated into *Step Output* without further processing.

Each *Context Slice* is rendered in one or more *Local Tasks*, each of which will be assigned to one crowd worker. The results of *Local Tasks* submitted by crowd workers contributes to *Context Slice Result* via an *Aggregation Mechanism*. In this paper, we implemented two commonly used aggregation mechanisms from the crowdsourcing community: majority vote and create-review [70]. The majority vote mechanism applies to rating, tagging, or voting tasks that are distributed in parallel among workers (Steps 1, 3, 4). A *Context Slice* uses the answer chosen by most of the workers (above a threshold) as the *Context Slice Result*. The create-review mechanism applies to free-text input (Steps 2, 5). The first crowd worker creates a free-text answer (one information piece or one narrative presentation); then, a second worker reviews and refines this result. The process continues until no new revisions are made. A *Context Slice* uses the final unrevised answer as the *Context Slice Result*. To ensure the quality of work, we ask crowds to provide brief explanations of their choices for the majority vote tasks; for create-review tasks, we provide self-assessment rubrics [18] below the free-text input boxes.

4.3 Step 1: Search and Filter

Step Input: All the raw text documents available for analysis.

Step Output: The subset of documents that are relevant to the global context.

Context Slices and Local Tasks: Each *Context Slice* contains n documents with shared entities, which could potentially "connect the dots" among documents and help workers better determine document relevance. Each *Context Slice* has no more than 600 words (2 minutes of reading for the average adult) and is rendered in $c \geq 3$ *Local Tasks*. Each crowd worker gives a relevance rating and provides a brief explanation.

Aggregation Mechanism: Majority Vote. A document is considered relevant if a majority of workers deemed it so. All relevant documents are put into the *Output* (Fig. 8).

4.4 Step 2: Read and Extract

Step Input: Relevant documents found in Step 1.

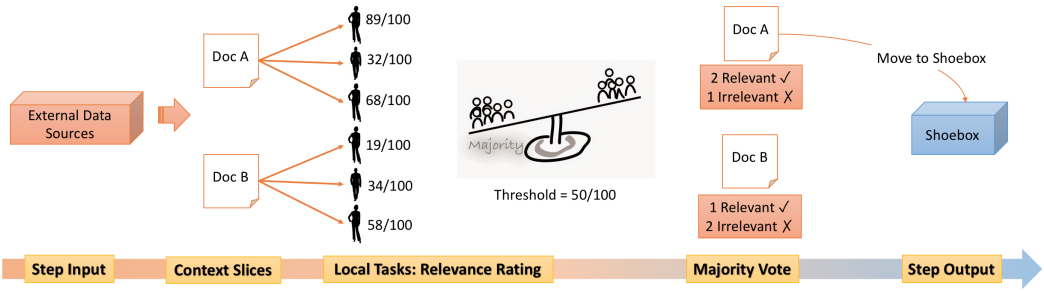

Fig. 8. Step 1: Search and Filter. Crowds independently rate document relevance from 0 (completely irrelevant) to 100 (completely relevant). Using a predefined threshold, each relevance rating is converted to a binary vote. Documents with the majority vote will be passed to Step 2.

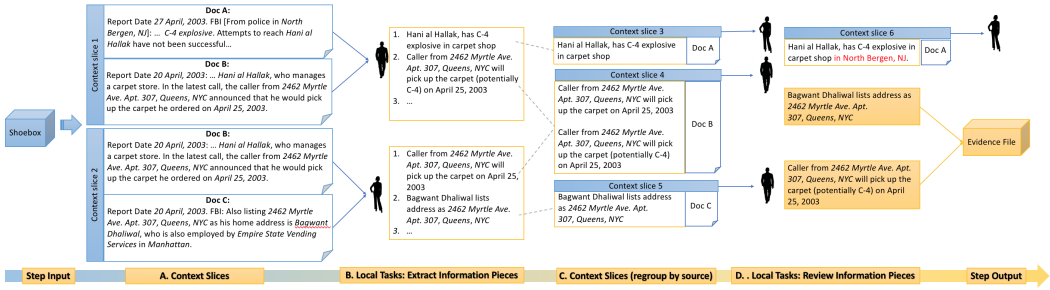

Fig. 9. Step 2: Read and Extract. CrowdIA groups documents with overlapping entities into context slices of size $n = 2$ (A). The first batch of crowd workers extracts information pieces from context slices (B). The information pieces are then regrouped by their source documents into new context slices (C). The following batches of crowds review information pieces (D). The process continues until no new revisions are made.

Step Output: A set of information pieces comprised of the key entities in the relevant documents.

Context Slices and Local Tasks: Each *Context Slice* contains n documents and (if not the first worker) information pieces extracted from the documents. For each *Context Slice*, $c \geq 2$ *Local Tasks* are rendered sequentially and the crowd workers extract or review the information pieces.

Aggregation Mechanism: Create-Review. Each *Context Slice* ends up with a list of final information pieces. Notably, when context slices have overlapping documents, crowds can extract information pieces that synthesize information from multiple documents. The disadvantage is that the same information pieces could be extracted multiple times by different workers. The reviewing process re-organizes information pieces by source documents into new context slices to help remove duplicates (Fig. 9).

4.5 Step 3: Schematize

Step Input: Information pieces extracted in Step 2.

Step Output: Tags on information pieces that identify targets and form a categorical schema.

Context Slices and Local Tasks: Each *Context Slice* contains n information pieces and is rendered in $c \geq 3$ *Local Tasks*, where each crowd worker fills in one free-text target and selects one or more predefined tags independently. The free-text target identifies candidates for the unknown element of the global task (e.g., potential target locations of a terrorist attack, or the suspect in a murder case). The predefined tags link each information piece to the known elements of the global task.

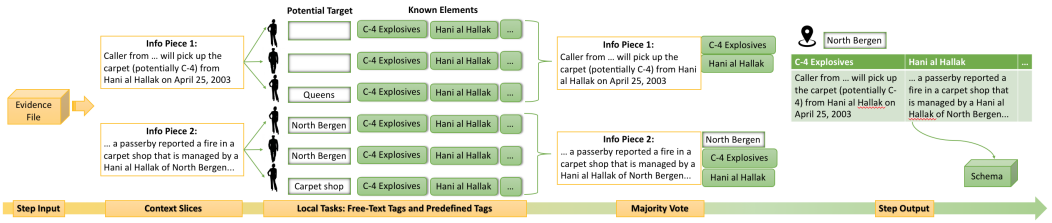

Fig. 10. Step 3: Schematize. Crowds identify potential target locations and tag the information pieces with known elements. Information pieces are tagged with tags that earned the crowd's majority vote and organized into profiles of the candidate targets.

Inducing structure from new data with distributed crowds is a challenging problem [1, 9, 35]. In CrowdIA, predefined tags depend on the types of data and known facts. For example, we used "means," "motive," and "opportunity" tags for the moderate dataset because these three categories are common practices in determining the suspect in a criminal case. In the difficult dataset, we used all known elements as predefined tags because the target location is assumed to have the highest number of abnormal activities.

Aggregation Mechanism: Majority Vote. Tags that received a majority vote are retained for the information pieces in each *Context Slice*. Each free-text target has at least one information piece which also has at least one predefined tag. This creates a profile for each target candidate, which marshals the relevant information pieces into a tabular structure according to the predefined tags (Fig. 10). For example, each murder suspect's profile organizes all of their means, motive, and opportunity evidence.

4.6 Step 4: Build Case

Step Input: Target profiles containing information pieces tagged from Step 3.

Step Output: Preliminary hypotheses developed by comparing the target profiles.

Context Slices and Local Tasks: Each *Context Slice* contains n profiles and is rendered in $c \geq 3$ *Local Tasks*. Each crowd worker selects the most likely candidate or (in the case of aliases) declares them identical and provides a brief explanation. As a proof-of-concept, we adopt the single elimination tournament among candidates [23], each competition being a *Context Slice*. Profiles are initially ranked by the number of tags and information pieces (Fig. 11).

Aggregation Mechanism: Majority Vote. For each *Context Slice*, the profile with majority vote enters the next round of comparisons until only one profile is left.

4.7 Step 5: Tell Story

Step Input: The best preliminary hypotheses from Step 4.

Step Output: A narrative conclusion backed up by supporting evidence.

Context Slices and Local Tasks: Each *Context Slice* contains n profiles and is rendered in $c \geq 2$ *Local Task*s sequentially. Crowds write and review a narrative that integrates the profile information into a complete story (Fig. 12). In our experiments with one missing element (person or location), only the winning profile is considered the *Step Input*, so there is not really a need for a *Context Slice* when $n = 1$.

Local Task Aggregation Mechanism: Create-Review. Each *Context Slice* ends up with a reviewed presentation, which is also the output of the entire pipeline.

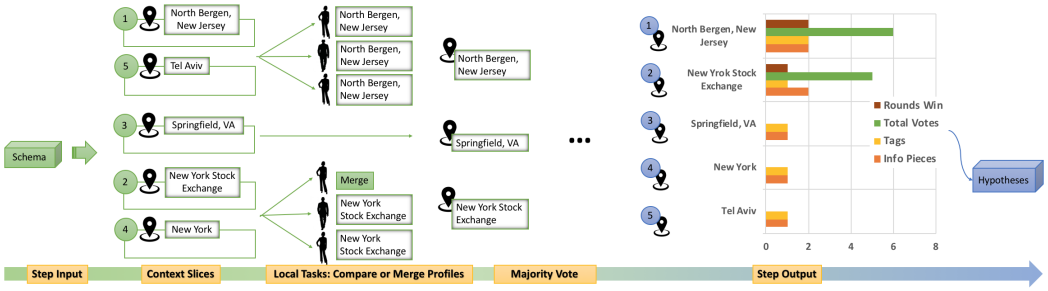

Fig. 11. Step 4: Build Case. Crowds compare candidate profiles and merge aliases. As in a single-elimination competition, workers in Step 4 rank candidates by their perceived likelihood of being the target location.

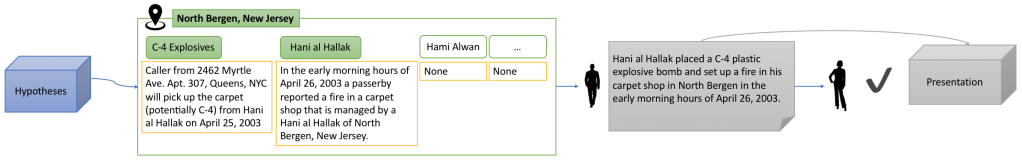

Fig. 12. Step 5: Tell Story. Crowds put together the information in the winning profile and write a complete narrative. The presentation is ready when no new revisions are made.

4.8 Refining Path: Top-Down

When the final presentation does not meet the expectation of clients, we need to refine the previous analysis. Furthermore, complex sensemaking like intelligence analysis is never fully complete, but rather becomes more valid with the available evidence [12]. As the analysis proceeds, new insights and lines of inquiry may arise. These issues motivate the needs to refine the previous intermediate analysis with the new knowledge learned and any feedback provided by clients.

There are two questions to ask before triggering the refining path: 1) What are the problems with the current analysis, and 2) From which steps did the problems originate? The first question decides if a refining path is needed, and the second question decides where and how the intermediate analysis should be refined. A reasonable start could be to ask an expert to review the current pipeline result, locate problematic intermediate results, and provide feedback for the corresponding step. It is also worth exploring how crowds can be leveraged in answering the two questions.

Once the step is identified and feedback is provided, new context slices are needed for crowd workers to address the feedback. We designed a feedback format with three elements: *context* (where is the problem), *critique* (what is the problem), and *task specification* (what is needed to fix the problem). The new *Context Slices* contain the formatted feedback and a subset of *Step Input* that produced the feedback *context*. Each *Context Slice* is then rendered in local tasks similar to the bottom-up path. Each crowd worker first evaluates whether the feedback can be addressed with the given material, then submits an explanation of why not or new results to address the feedback. The new results are aggregated into a refined *Step Output*. If the new *Step Output* is different from the previous one, the subsequent steps will be re-executed with the new analysis. Otherwise, the expert will need to find another previous step to fix.

5 EVALUATION: SOLVING MYSTERIES WITH CROWDS

To evaluate the feasibility of the pipeline, we deployed CrowdIA on Amazon Mechanical Turk (MTurk) and asked crowd workers to solve three mysteries of increasing difficulty levels. To evaluate the quality of crowd analysis with CrowdIA, we compare the conclusions by crowds to the correct answers of the mysteries. In this paper, we focus on evaluating the bottom-up forward pipeline, leaving experiments with the top-down refining path for future work.

5.1 Method

We recruited participants from MTurk US-only pool and paid at least minimum wage in our location ($7.25 per hour). Based on the time the individuals and crowds spent in the preliminary studies, as well as the effects of queuing and/or idling tasks, we determined the time needed to complete local tasks in each step as: Step 1: 1.7 min, Step 2 extract: 2.8 min, Step 2 review: 2.5 min, Step 3: 1.6 min, Step 4: 3.3 min, Step 5 create and review: 2.2 min. Consequently, to ensure at least minimum wage, we provided the following payments for the local tasks: Step 1: $0.20, Step 2: $0.34 and $0.30, Step 3: $0.19, Step 4: $0.40, Step 5: $0.24 and $0.24. CrowdIA posts each step on MTurk as a Human Intelligence Task (HIT) that dynamically renders the context slices and assigns a worker to each slice. We assign each worker to only one context slice of one step to demonstrate the capability of distributed novice crowds to solve mysteries and mitigate learning effects or collusion. Crowd workers who quit an accepted HIT without submitting it were not allowed to resume the unfinished work or take a new HIT.

We evaluate our pipeline with easy, moderate and difficult datasets (for details on the datasets, see Appendix A). We consider datasets with more documents, more elements (who, what, where, when) and more complicated relationships among elements to be more difficult.

5.2 Results of Easy Dataset

The easy dataset contains three documents about three girls who might have ruined Mr. Potter's flowerbed [17]. The crowd workers successfully found the culprit and presented their conclusion with supporting evidence.

We recruited five crowd workers, each working on one step in the pipeline. The first worker took 34 seconds to find the two relevant documents, the second took 12.5 minutes and extracted 10 important information pieces, the third took 22.7 minutes and organized the information pieces into four groups, the fourth took 10.1 minutes and generated three hypotheses for each suspect, and the fifth took 5.7 minutes to pick the most likely culprit, offering the following conclusion:

> I think it was Serina who had the muddy shoes after playing hopscotch. Her shoes were muddy so that could indicate that she went into the just watered flower bed. Maybe she only ran through it to get to her friends so they could play but she might have stomped on the flowers on her way to the play area. She was in a hurry and not paying attention to what she was doing.

Detailed crowd analysis results for the easy dataset are presented in Appendix B.

5.3 Results of Moderate Dataset

The moderate dataset, inspired by the popular Clue board game, has nine documents; three suspects with different means, motives, and opportunities to kill Mr. Boddy; and four witnesses. The crowds successfully identified the murderer and backed up the conclusion with supporting evidence.

We recruited a total of 76 crowd workers to analyze the dataset. In Step 1, 27 workers found seven relevant documents and excluded the two documents about wrong means and wrong opportunity of the two wrong suspects (time spent in minutes: mean=5.8, median=2.9, std=8.78). In Step 2,

14 workers extracted eight information pieces, of which seven were important (time spent on creation tasks in minutes: mean=10.7, median=4.6, std=18; time spent on reviewing tasks in minutes: mean=7.6, median=3.81, std=9.61). In Step 3, 24 workers tagged the seven important pieces with person names and means / motive / opportunity evidence type, whereas the useless information piece received a None tag (time spent in minutes: mean=7.7, median=3.5, std=12.79). One of the witnesses also got tagged as a potential suspect, resulting in four profiles. In Step 4, nine workers weighed the suspect profiles in the single elimination tournament and Scarlett was deemed the most likely murderer (time spent in minutes: mean=13.0, median=7.2, std=14.92). Finally, in Step 5, 2 workers narrated the conclusion (creating worker took 23.7 minutes, reviewing worker spent 2.05 minutes):

> Miss Scarlet killed him [Mr. Boddy]. She was seen at Mr Boddy's house on the night of his death. She also had the murder weapon seen by her trainer in her bag. Also, she had the motive since she would inherit his estate.

Detailed crowd analysis results for the moderate dataset are presented in Appendix C.

5.4 Results of Difficult Dataset

The difficult dataset is part of the *Sign of Crescent* dataset [32] used as training material for professional intelligence analysts. We streamlined the Crescent dataset to only cover one terrorist plot, added extra documents as noise, and specified the goal as identifying the target location of the attack. There are 13 documents, four terrorists, and 12 locations mentioned in the documents.

We recruited a total of 135 crowd workers: 18 workers in Step 1 (time spent in minutes: mean=10.9, median=4.3, std=15.6), 22 workers in Step 2 (time spent in minutes: mean=17.3, median=10.7, std=17.2), 78 workers in Step 3 (time spent in minutes: mean=12.1, median=3.1, std=16.0), 15 workers in Step 4 (time spent in minutes: mean=9.9, median=7.3, std=9.0), and two workers in Step 5 (creating worker took 47.4 minutes, reviewing worker spent 1.5 minutes).

Echoing the results from our preliminary study, the crowd workers were one step away from the actual target location New York Stock Exchange. However, they found the weapon storage location North Bergen, New Jersey and ranked New York Stock Exchange as the second possible target. Below, we examine the crowds' performance in detail.

Step 1: Crowd successfully retrieved indirectly relevant documents. In Step 1, seven documents out of 13 directly mentioned one or more key elements and are automatically considered relevant. The remaining six documents (three indirectly relevant and three irrelevant) are each rated by three crowd workers on a 0–100 scale. The crowds found four relevant documents (with one extra irrelevant document) from the six documents with a threshold of 50, resulting in 11 relevant documents. A follow-up analysis found that thresholds ranging from 30 to 60 would lead to the same result, with lower thresholds increasing false positives and higher ones increasing false negatives. Thresholds 0–10 would include all three irrelevant documents, 15–25 would include two irrelevant documents, 65–70 would include one irrelevant document and miss one relevant document, 75 would include one irrelevant document and miss two relevant documents, and above 80 would not include irrelevant documents but miss three relevant documents.

Step 2: Crowd extracted most key useful information pieces. In Step 2, a total of 26 information pieces were extracted from the documents, of which 18 were useful ones. The information pieces cover key evidence about terrorists' real names and aliases, phone calls, and the bomb and the storage location. We found that the crowds were able to synthesize information across two documents, viz.: "Hani al Hallak's carpet shop in North Bergen caught fire" and "Police found C-4 explosives in the carpet shop reported on fire in North Bergen" were extracted as one information piece: "Hani al Hallak's carpet shop has C-4 explosives." The crowd's review process solved issues

like misspellings, incomplete name references, missing elements (who, what, where, when, etc.), and duplicates.

On the other hand, not all important information pieces were extracted. One important information piece showing that one of the terrorists works in the actual target location did not get extracted. Some information pieces about the relationships and roles of terrorists also did not get extracted. The missed activities were cover-ups of terrorists and not obvious to an early-stage investigation. These issues could be improved by re-executing Step 2 with additional feedback.

Step 3: Majority vote elicits accurate tagging and potential target identification. After Step 3, 18 of the 26 information pieces were tagged, excluding the four information pieces from the irrelevant document. Following the majority vote aggregation, all information pieces were accurately tagged with the key evidence.

We closely examined the tags and found that individual crowd workers tended to give information pieces more tags than strictly needed. Some workers just selected every tag and others selected nothing when they could not identify any locations. The majority vote mechanism helped eliminate the influence of such low quality work and only kept the accurate tags.

Five location tags were created. One notable development was that two different workers both identified a location "Tel Aviv" in the information piece: "I will be in my office on April 30 at 9:00AM. Try to be on time." One of the workers even gave very specific information: "the location is Israel at Mike's Place, a restaurant in Tel Aviv." We later learned that there was a real Palestinian suicide attack perpetrated by British Muslims which killed three civilians and wounded 50 at Mike's Place in Tel Aviv on April 30, 2003, the same timeframe as the dataset. Although crowds were instructed not to add extraneous information, these two workers aligned the information in the given context slice with their external knowledge and mental model.

Step 4: Crowd logically reasoned and weighed hypotheses. The ranked location tag results are shown in Fig. 11. The final winning location was North Bergen, New Jersey, the last place the bomb was stored before transferred to the target location. The runner-up, losing by only one vote, was the correct answer, New York Stock Exchange. Even though the crowd narrowly missed the actual target, the winner is the second-most crucial location to investigate. The correct answer, New York Stock Exchange, was merged with another location, New York, and won one of the competitions with insightful explanations by workers:

> The New York Stock Exchange is a specific, high value target for terrorists because a bomb attack there would likely cause many casualties and have a negative effect on the US economy. Springfield, VA is a very broad target and besides the fact that one of the terrorists lives there there isn't much evidence than an attack will take place there.
> — *Worker 1, New York Stock Exchange vs. Springfield, VA (Round 2)*
> There are multiple pieces of evidence showing suspicious activity centered on the NYSE. There's just one pieces of evidence pointing to Springfield, and it's just that a suspect lives there, there's no real evidence he's doing anything there.
> — *Worker 2, New York Stock Exchange vs. Springfield, VA (Round 2)*

Unfortunately, New York Stock Exchange lost in the final competition with North Bergen, New Jersey, where the terrorists store the bomb in a carpet store before transferring it to New York. However, the explanations were not as insightful or convincing, e.g.:

> They found an actual C4 in New Jersey, which makes me believe that was more likely meant to be the target.
> — *Worker 3, New York Stock Exchange vs. North Bergen, New Jersey (Round 2)*

Step 5: Crowd wrote a clear narrative presentation. Using the profile of North Bergen, New Jersey, workers from the last step created a narrative that connected the evidence to current findings

and justified the likelihood of this place being a potential target. The final presentation created by the crowds was: "Hani al Hallak placed a C-4 plastic explosive bomb and set up a fire in his carpet shop in North Bergen in the early morning hours of April 26, 2003."

6 DISCUSSION

6.1 RQ1: How can we modularize the sensemaking process?

Analysis provenance enables step-wise debugging of the sensemaking process. Modularizing the sensemaking steps with explicit definitions of information needs (inputs) and intermediate analysis results (outputs) enables step-wise debugging and refinement, breaking down a big black box into smaller, more inspectable modules. For example, when the crowds analyzed the difficult dataset, they failed to extract some important information pieces (false negatives) in Step 2. This resulted in incomplete profiles of the potential target locations, which we believe led to the narrow miss of the correct target. By examining the intermediate analysis results, either experts or crowds could have potentially debugged the situation and refine the analysis in a top-down refining path (as in Figure 7). In future work, experts could manage the execution of the pipeline, similar to CrowdWeaver [34], and provide structured and situated feedback [38] for crowd workers to refine previous analyses. Alternatively, crowds themselves could critique and refine intermediate analysis results in a feedback pass. In prior work, crowds have been used to provide feedback [42, 66] on visual designs, accurately evaluate each other's credibility [62], and react to personalized expert feedback while brainstorming [8].

Modularization enhanced scalability, resusability of analysis, and efficient division of labor. CrowdIA enabled as many as 134 transient novice crowd workers to collaborate to solve a difficult mystery, producing high-quality, insightful analysis output. The same number of people working together in a collocated way would be a big challenge to coordination and communication. Getting the same number of trained analysts working at the same time would be even more difficult. CrowdIA's automated facilitation mitigates logistics burdens, enabling workers to invest their time and efforts in the core analysis tasks. Later crowd workers continued the analysis from where previous workers left off, without requiring the previous workers to explain their intermediate results or thought processes. Furthermore, CrowdIA did not require crowds to have significant sensemaking expertise. Workers were all novices and transient, without prior exposure to the dataset, and giving only a small time commitment (typically a few minutes) each. These features can open up the sensemaking process to dynamically recruit from a much bigger pool of contributors.

Alternative strategies to schematize information. Steps 3 and 4 could be modified to support many different types of schemas. Different structures could be helpful for different types of analyses [22]. A node-link graph structure is very general to capture many types of relationships in the data, but can be difficult to hand off during collaboration [69]. CrowdIA implemented a more specific tabular structure to represent suspect profiles, which resulted in accurate hypotheses.

In CrowdIA, we implemented both location-centered (difficult mystery) and person-centered (easy and moderate mysteries) profile schema strategies. Alternative methodologies, such as analysis of competing hypotheses (ACH) [31], could also be applied. We found that an effective strategy is to tag information with appropriate categories with which the information pieces can be organized from different perspectives. This strategy is simple for novice crowd workers and highly scalable. Future work can also explore a data-centric approach by inducing tags directly from the documents [1, 35].

Optimizing each step with best-suited techniques. The modular design can support future research on optimizing each step of the pipeline as well as the overall workflow. For information foraging stages (Step 1 and 2), advanced algorithmic techniques [51, 59] can be leveraged to improve efficiency. Crowds can focus on edge cases where machine learning models do not perform well [9], which in

Step	Context Slicing Goal	Context Slices	Alternative Slicing Methods
1	Provide context to define relevance	Documents that share entities	Documents of the same topic, sources
2	Provide context to complete missing parts of facts	Information pieces from the same source documents	Information pieces created by the same people, date
3	Provide context to identify evidence type	Information pieces that share entities	Information pieces with similar relationship types, connotation
4	Provide context to combine and/or compare schema	Profiles of suspects organized by evidence types	Evidence types organized by date, location
5	Provide context to back up each hypotheses statement	Most likely suspect and complete profile	Strongest suspects for each evidence

Table 1. Customized context slicing of each step depends on the level of analysis and the goal of the step.

turn helps train the machine learning models. For information synthesizing tasks, crowds are better suited than algorithms and have shown success in other applications [29, 43]. More complicated approaches like online contest webs [43] can be applied to guide the crowds to build hypotheses. Experts can also take over whenever they deem it appropriate.

6.2 RQ2: How do we distribute and aggregate the analysis in each step?

Decomposing a big problem into small manageable problems has been a major challenge for the crowdsourcing community. To make sense of large amounts of data, many solutions employ a single step for sensemaking and distributing the work by 1) showing each worker all the data, 2) showing each worker one piece of data, or 3) showing each worker an arbitrary subset of data. All such microtasks are linearly defined, similar to how we divide the documents in the first step. Instead of naively passing uniform local tasks from Steps 1 to 5, we create *Context Slice*s that divide each *Step Input* into cohesive subsets, and aggregate the *Context Slice Results* into *Step Output* (which is also the next *Step Input*), before creating new context slices.

Context slices enable meaningful and scalable division of work. Although the steps in the sensemaking pipeline were already modularized for experts, our concept of *Context Slice*s partitions the each *Step Input* so that novice crowd workers can contribute meaningfully. *Context Slices* enable workers to generate meaningful results that synthesize information beyond what can be extracted from a single piece of information or an arbitrary subset of information. For example, in the first step of solving the difficult dataset, when given a context slice containing one directly relevant document and one unrated document with shared entities, the crowds were able to identify other indirectly relevant documents. Without context slices, workers in the preliminary study were not able to identify these indirectly relevant documents.

Context slice design depends on data and differs among steps. An open challenge is that context slicing methods must be carefully designed for each step. CrowdIA implemented customized context slicing methods for each step in the pipeline (Table 1). We specify the context unit (how the slices are defined) according to the level of analysis in each step: documents, information pieces, and profiles, and the context slicing goal (how the slices are determined) according to the step goal. Exploring the trade-off between the size of context slices and the quality of local task output, and exploring alternative context slicing methods, are both promising directions for future work.

6.3 RQ3: How do crowds perform in solving mysteries with the modularized pipeline?

Handling false positives and false negatives. Crowds handled false positives within the pipeline. For example, in the difficult dataset, crowds included one irrelevant document (a false positive) in Step

1, which propagated the useless information to later steps. However, Step 3 guaranteed that only useful information pieces are tagged with evidence types and put into profiles of candidates. Thus, the useless information was filtered out in Step 3. The pipeline was able to recover from the false positive and save worker labor in later steps.

We did not encounter false negatives in Step 1 with a rating threshold of 50. In general, the trade-off between false positives and false negatives could potentially be controlled by the rating thresholds [39]. We encountered false negatives in Step 2, where the crowds failed to extract all the relevant information pieces. The bottom-up pipeline alone did not recover from the false negatives. Future work on the top-down refining path could help resolve false negatives.

Crowds used external knowledge in information foraging. We found that sometimes crowds connected their own knowledge to the dataset. In Step 3, the workers created a location tag "Tel Aviv" which is not mentioned anywhere in the documents. This connects the information from the documents under investigation to external knowledge from the crowds. Despite our assumption (Appendix D) that no external knowledge is needed to solve these mysteries, the wisdom of crowds potentially broadens the coverage of the investigation.

Crowd explanations provided diverse perspectives in information synthesis. In the synthesizing stages (Steps 3–5), we found that the crowds provided diverse perspectives. When comparing suspects in the moderate dataset, one crowd worker chose the wrong suspect (Professor Plum instead of Miss Scarlett), and provided the explanation: "To cut a man's throat you would need to be at least as strong as him, I don't think women in general have the same sort of physical power as men, therefore I don't think Miss Scarlet had the physical strength to overpower Mr. Boddy and cut his throat..." Although this hypothesis does not align with the correct answer, real-world investigation can benefit from such insights for further data collection and analysis. Further exploration on collecting, structuring and making use of crowd explanations would be valuable future work.

6.4 Generalizability

We chose to deploy our pipeline to solve intelligence analysis mysteries because they exemplify the challenge of exploratory analysis: building robust and logical hypotheses from known facts to achieve a final conclusion as close to the hidden truth as possible. However, we envision the pipeline as adaptable to broader applications with different sensemaking challenges, as well as opening up more in-depth research within each step.

Broader applications beyond intelligence analysis. The general class of "mysteries" CrowdIA may help solve is potentially broad, including investigations in law enforcement, journalism, and human rights advocacy. However, we also expect that our pipeline can be adapted for other sensemaking tasks and domains. While future work is needed to understand the trade-offs, we anticipate that our approach will translate most directly to sensemaking tasks that involve uncovering hidden patterns or relationships among many text-based documents, such as coding qualitative data [1] or synthesizing creative ideas [7]. Our approach may also be suitable for sensemaking tasks that incorporate personal preferences, such as trip planning [68], online shopping [35], or researching home improvement solutions [29], because our pipeline already assumes iterative cycles of client feedback and revision. Finally, our approach may support crowdsourced sensemaking to generate hypotheses of biological and environmental phenomena [41, 46]. For these latter problems, it may be necessary to modify CrowdIA's steps to better align with the scientific method rather than the sensemaking loop of intelligence analysts.

Flexible crowd compositions and collaboration settings. A major constraint of using crowds on MTurk is that workers are transient and novices. This serves well for our purpose as a proof-of-concept, but in real-world intelligence analysis, the crowd's analysis may serve as an assistance

to experts. When solving the difficult dataset, crowds were able to prune the noisy information from the documents without much loss of important information. Expert analysts could focus on the pruned information for more advanced analysis. Along with being a more efficient division of labor, this approach also allows for professional oversight, preventing novice crowds from jumping to wrong conclusions that can result in harmful consequences.

When confidentiality is a concern, one possibility is to incorporate task assignment techniques for sensitive documents [6], but these may limit workers' access to global context and degrade quality. Another possibility is to use a trusted internal group who can access the confidential documents when collocated, synchronous, devoted experts are not available. Many data management businesses already employ or have access to such internal worker pools [44].

6.5 Limitations and Future Work

In this paper, we focused on first establishing a proof-of-concept pipeline that orchestrates crowdsourced sensemaking, and then investigating how well the pipeline can facilitate novice asynchronous distributed crowds in solving mysteries. However, we did not empirically compare our approach to other sensemaking techniques or systems. Future evaluation studies could compare CrowdIA's highly structured process to more free-form approaches (e.g., [26, 55]) to articulate the trade-offs of exploiting Pirolli and Card's sensemaking structure. Comparing to alternative data slicing techniques, such as 1) extreme slicing, in which each worker gets only one document and votes for the likely target based entirely on local information; or 2) no slicing, in which each worker sees all the documents and attempts to solve the mystery, could suggest pipeline modifications that enable greater flexibility in worker time commitments and microtask granularities. CrowdIA's modular approach also suggests opportunities to experiment with alternative task designs for specific steps within the proposed pipeline, using this paper's configuration as the baseline, similar to the experimental framework organized by Parikh et al. [47] for computer vision research.

A challenge in conducting controlled studies of CrowdIA is cost. Running the pipeline to solve the moderate and difficult mysteries required approximately 100 crowd workers and cost $50 per execution. Although other crowd-based systems, especially those requiring workers with specialized expertise, can be much costlier [29, 43, 57], CrowdIA executions will become more expensive as the mystery gets more complicated, to compensate the increasing numbers of workers. Alternatively, researchers could leverage public enthusiasm for solving mysteries [45] to recruit volunteer crowds. These self-selected participants may bring more dedication and expertise to the problem, but may be less motivated by the artificial data sets common to controlled experiments.

To manage the scope of the problem, we enforced some key assumptions (see Appendix D) on the initial data input and the final result output associated with the target mysteries. Further field research is needed to understand how to relax these assumptions when applying the pipeline to more complex, real-world mysteries.

7 CONCLUSION

In this paper, we proposed a modularized pipeline that guided crowds to collaborate on solving mysteries. With clearly defined inputs, outputs, and context slicing methods for each step, crowd workers on MTurk successfully undertook the entire sensemaking process and solved mysteries of increasing difficulty levels. We implemented the pipeline as CrowdIA, a web-based crowdsourcing system that provides automated facilitation of the sensemaking process for novice transient crowd workers. Our pipeline enables research on different sensemaking steps to be dynamically plugged in and tested, thereby coordinating large-scale efforts from the sensemaking research community. Our hope is that this pipeline will serve to accelerate research on sensemaking, and contribute to helping people conduct in-depth investigations of large collections of information.

Document 1	Document 2	Document 3
One hot, dry day Neva saw Mr. Potter shaking his head as he stood by his flowerbed. "Somebody ruined all my flowers," he said. "I had the hose out watering them. When I went to put it away, somebody tromped through the rows and stomped on my flowers." "Who'd do a mean thing like that?" Neva asked. Mr. Potter sighed. "Somebody who likes mischief, I guess." "I'm walking to the mall now. Maybe later I can find out who did it," Neva told him.	On Neva's way to the mall, she saw three girls playing hopscotch. She decided to stop and watch how expertly they moved over the chalk marks. Lucy had to hop very carefully because one of the straps was broken on her left sandal. Cathy hopped slowly. She wore purple sneakers that looked worn-out. Cathy seemed worn-out, too. Serina hopped the fastest. The muddy soles of her white jogging shoes hardly seemed to touch the sidewalk as she moved.	Ada Peterson is a graduate student in ABC Tech. She went on a vacation to Yellowstone National Park this August with her family. She spent 3 days at her cousin Elaine's house in Los Angeles before that. She spent a week there, and before she came back to school on Aug 28th, she went to Utah for 2 days to visit her old friend Cindy.

Table 2. Easy dataset adapted from a brain teaser. The correct answer is that Serina is the culprit.

A DATASETS

Given the complexity of the entire sensemaking process, we evaluate our pipeline with simplified datasets in three levels of difficulty. Dataset with more documents, more elements (who, what, where, when) and more complicated relationships among elements are considered more difficult.

The easy dataset is adapted from one of the brain teasers from Braingle [17], the answer being *Serina is the culprit*. There are two relevant documents: Document 1 introduces the background setting and Document 2 lays out the suspects and information about them. We added a third irrelevant document as noise, and masked the original names to prevent crowds from finding the solution online. There are 227 words in all three documents, with 162 words in the two key documents (Table 2).

The moderate dataset is adapted from the popular board game Clue, where there is a limited number of suspects (who), weapons (what), locations (where), and one known murder time (when). We picked three suspects, two weapons and three locations in the whole dataset (Table 3), and the correct answer being *Miss Scarlett killed Mr. Boddy (victim) in his kitchen with a knife, because she will be bequeathed with the large estate after his death.*

The difficult dataset is part of the *Sign of Crescent* dataset [32]. It is used as a training material for intelligence analysts. There are 41 fictional text intelligence reports about three coordinated terrorist attack plots in three US cities. Each plot involves a group of at least four suspicious people. And each report document contains a single prose paragraph of 33 to 210 words (Fig. 4). We took nine of the documents that contain evidence of one of the attack plots: *A C-4 plastic explosive bomb, will be detonated at 0900hrs on 30 April,2003, by Hamid Alwan [alias Mark Davis] in New York Stock Exchange. Support Hamid with money and bomb storage and transportation is a group of terrorists: Muhammed bin Harazi [alias Abdul Ramazi], Hani al Halak, Sahim Albakri [alias Bagwant Dhaliwal].* We added two irrelevant documents as noise.

Evidence	Miss Scarlett*	Prof. Plum	Reverend Green
Means	has knife	handgun (wrong weapon)	has knife
Motive	inherit victim's property on his death	wife has affair with victim	victim stole money from their business
Opportunity	visited victim's house	car driving past (wrong location)	phone call (wrong location)

Table 3. Moderate dataset skeleton adapted from the card game Clue.

Report Date 14 April, 2003. CIA: From an interrogation of a cooperative detainee in Guantanamo. Detainee says he trained daily with a man named Ziad al Shibh at an Al Qaeda explosives training facility in the Sudan in 1994. From a captured laptop computer in Afghanistan it is learned that Ziad al Shibh holds a United Arab Emirates passport in the name Faysal Goba. INS check reveals that a Faysal Goba, from the United Arab Emirates, entered the USA on a travel visa in January of 2003 stating that he would be visiting a person named Clark Webster in Richmond, Va. The contact address given by Goba was: 1631 Capitol Ave., Richmond VA; phone number: 804-759-6302.
Report Date 27 April, 2003. Intercept of cell phone 804-774-8920. In a very brief call from this number to phone number 703-659-2317 on 26 April, 2003, the caller speaks in Arabic. A translation reads: "We are now prepared to take the crescent to victory".

Table 4. Example documents from the difficult dataset adapted from *The Sign of the Crescent*.

B CROWD ANALYSIS OF EASY DATASET

Step 1: The crowds successfully identified that Document 1 and Document 2 are relevant.
Step 2: The information pieces extracted by the crowds are:

(1) Neva witness Mr. Potter shaking his head by the flowerpots.
(2) Mr. Potter claimed someone ruined his flowers by stomping on them after he watered them.
(3) Neva wondered who it was.
(4) Mr. Potter didn't know but guessed it was for mischief's sake.
(5) Neva said he might help figure it out later.
(6) On that day, Neva saw three girls playing hopscotch and watched them.
(7) Lucy had a broken strap on her left sandal.
(8) Cathy looked worn out and her shoes were worn too.
(9) Serina was fast but had muddy shoes.
(10) Any of these three girls could be the culprit.

Step 3: The schemas organized by crowds are:

Group 1: Possible suspects
6. On that day, Neva saw three girls playing hopscotch and watched them.
7. Lucy had a broken strap on her left sandal.
8. Cathy looked worn out and her shoes were worn too.
9. Serina was fast but had muddy shoes.

10. Any of these three girls could be the culprit.

Group 2: Evidence Mr. Potter's flowerpots were destroyed
1. Neva witness Mr. Potter shaking his head by the flowerpots.
2. Mr. Potter claimed someone ruined his flowers by stomping on them after he watered them.

Group 3: Reflections
3. Neva wondered who it was.
4. Mr. Potter didn't know but guessed it was for mischief's sake.
5. Neva said he might help figure it out later.

Group 4: Could be Neva
4. Neva said he might help figure it out later.

Step 4: The hypotheses developed in Step 4 are:

Hypothesis 1, Group 1, 2, & 3: Mr. Potter's flowerpots were destroyed and claimed someone ruined his flowers by stomping them. Neva saw three girls playing hopscotch and noticed that one girl, Lucy had a broken strap on her sandal. It is likely that Lucy stomped on the flowers and in doing so her strap on her left sandal caught on something and broke, or from the force or her action that is how the flowers were ruined.

Hypothesis 2, Group 1, 2, & 3: Mr. Potter's flowerpots were destroyed and claimed someone ruined his flowers by stomping them. Neva saw three girls playing hopscotch and noticed that one girl, Cathy looked exhausted and had worn out shoes. Cathy looked so worn out due to the fact that she ruined Mr. Potter's flowers and ran off in a rush as to not get caught

Hypothesis 3, Group 1, 2, & 3: Mr. Potter's flowerpots were destroyed and claimed someone ruined his flowers by stomping them. Neva saw three girls playing hopscotch and noticed that one girl, Serina was very agile in playing hopscotch and had muddy shoes. Since Mr. Potter had just watered his flowers the soil would have been moist/damp and it makes sense that Serina ruined his flowers and ran back to her friends to play with them which is why her shoes are muddy

Step 5: The final presentation by crowds is:

I think it was Serina who had the muddy shoes after playing hopscotch. Her shoes were muddy so that could indicate that she went into the just watered flower bed. Maybe she only ran through it to get to her friends so they could play but she might have stomped on the flowers on her way to the play area. She was in a hurry and not paying attention to what she was doing.

C CROWD ANALYSIS OF MODERATE DATASET

Step 1: The crowds identified all documents about Miss Scarlett, documents about Prof. Plum's motive and wrong location, and Reverend Green's means and motive as relevant, i.e. all relevant documents were retrieved, one irrelevant document was retrieved (precision=85.7%, recall=100%).

Step 2: The information pieces extracted by crowds are:

(1) Miss Scarlett visited Mr. Boddy's house on the night of his death to return some books
(2) Miss Scarlett will inherit Mr. Boddy's large estate in the event of his death as Mr. Boddy's niece and nearest living relative
(3) Miss Scarlett's personal trainer Roger saw an ivory-handled fold-up knife in her gym bag a month ago, but she told police she has lost the knife several weeks ago.
(4) Professor Plum didn't know that his wife Linda had been having an affair with Mr. Boddy
(5) Professor Plum's car was seen driving past Mr. Boddy's house on the night of his death

	Miss Scarlett	Prof. Plum	Reverend Green	Roger
Means	Miss Scarlett's personal trainer Roger saw an ivory-handled fold-up knife in her gym bag a month ago, but she told police she has lost the knife several weeks ago.		Reverend Green kept a black utility knife in his pocket, but he told authorities that the knife had broken months ago and he had discarded it.	Miss Scarlett's personal trainer Roger saw an ivory-handled fold-up knife in her gym bag amonth ago, but she told police she has lost the knife several weeks ago.
Motive	Miss Scarlett will inherit Mr. Boddy's large estate in the event of his death as Mr. Boddy's niece and nearest living relative	Professor Plum didn't know that his wife Linda had been having an affair with Mr. Boddy	Reverend Green suspected Mr. Boddy had stolen money from the failed business they had together several years ago	
Opportunity	Miss Scarlett visited Mr. Boddy's house on the night of his death to return some books	Professor Plum's car was seen driving past Mr. Boddy's house on the night of his death		

Table 5. Profiles generated from information pieces tagged by crowds in Step 3.

(6) Reverend Green called Mr. Boddy's house twice on the night of his death
(7) Reverend Green kept a black utility knife in his pocket, but he told authorities that the knife had broken months ago and he had discarded it.
(8) Reverend Green suspected Mr. Boddy had stolen money from the failed business they had together several years ago

Step 3: The information tagged by crowds can be used to generate the following profiles: *Step 4:* The single elimination competition results by crowds are shown in Fig. 13

Step 5: The final presentation created by the crowds is:

Miss Scarlet killed him. She was seen at Mr Boddy's house on the night of his death. She also had the murder weapon seen by her trainer in her bag. Also, she had the motive since she would inherit his estate

C.1 Additional Experiment Results of Moderate Dataset

Step 1: Same as the above described experiment.

Step 2: The information pieces extracted by crowds are:

(1) Professor Plum has an active permit to carry a concealed handgun
(2) Professor Plum most recently renewed the concealed carry permit two months ago.

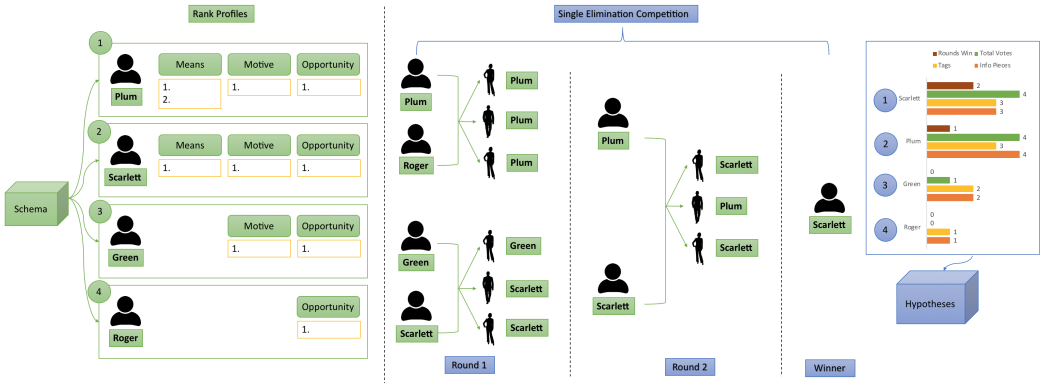

Fig. 13. Single elimination competition of profiles in Step 4.

(3) Phone records show Reverend Green called Mr. Boddy's house twice on the night of his death
(4) Sharon Miller possibly saw a car matching the description of Professor Plum's car driving past Mr. Boddy's house on the night of his death.
(5) Sharon Miller is Mr. Boddy's neighbor.
(6) Green kept a black utility knife in his pocket, for whittling wood carvings in his spare time.
(7) Green told authorities that the knife had broken months ago, and he had discarded it.
(8) Miss Scarlett was at Mr. Boddy's house the night he died.
(9) Miss Scarlett said she was only returning books.
(10) Miss Scarlett's personal trainer, Roger, told authorities that a month ago, he had seen an ivory-handled fold-up knife in Scarlett's gym bag
(11) Ms. Scarlett lost the knife weeks ago.
(12) Mr. Boddy has a large estate that will go to his nearest relative.
(13) Mr. Boddy's nearest living relative was Miss Scarlett.
(14) Professor Plum's wife, Linda was having an affair with Mr. Boddy.
(15) Linda said she didn't believe that Professor Plum knew about the affair.
(16) Reverend Green and Mr. Body were partners in a failed business
(17) Reverend Green suspected Mr. Boddy of stealing money

Step 3: The information tagged by crowds generated five profiles. Following the order of amount of evidence: Scarlett, Green, Plum, Miller, and Linda (Table 6).

Step 4: There were two rounds of competition. The first round knocked out Miller and Linda, and the second round knocked out Green and Plum.

Step 5: The final presentation created by the crowds is:

Scarlett was known to have a knife similar to the weapon used in the murder, she was at Mr Boddy's house and was also the last person who saw the victim alive.

D ASSUMPTIONS

To manage the scope of the problem, we enforce some assumptions on the initial data input and the final result output, focusing on the specific problem of text analysis we aim to tackle. The identification of data elements in each step of the pipeline will be defined based on these assumptions in following sections. We assume a crime solving or intelligence scenario.

	Scarlett	Plum	Green	Miller	Linda
Means	Miss Scarlett's personal trainer, Roger, told authorities that a month ago, he had seen an ivory-handled fold-up knife in Scarlett's gym bag. Ms. Scarlett lost the knife weeks ago.	Professor Plum most recently renewed the concealed carry permit two months ago.	Green kept a black utility knife in his pocket, for whittling wood carvings in his spare time. Green told authorities that the knife had broken months ago, and he had discarded it.		
Motive	Mr. Boddy has a large estate that will go to his nearest relative. Mr. Boddy's nearest living relative was Miss Scarlett.	Professor Plum's wife, Linda was having an affair with Mr. Boddy. Linda said she didn't believe that Professor Plum knew about the affair.	Reverend Green and Mr. Body were partners in a failed business. Reverend Green suspected Mr. Boddy of stealing money		Professor Plum's wife, Linda was having an affair with Mr. Boddy.
Opportunity	Miss Scarlett was at Mr. Boddy's house the night he died. Miss Scarlett said she was only returning books.	Sharon Miller possibly saw a car matching the description of Professor Plum's car driving past Mr. Boddy's house on the night of his death.	Phone records show Reverend Green called Mr. Boddy's house twice on the night of his death	Sharon Miller possibly saw a car matching the description of Professor Plum's car driving past Mr. Boddy's house on the night of his death.	

Table 6. Additional experiments: Profiles generated by information pieces tagged by crowds in Step 3.

D.1 Assumptions about External Data Resources

A1. There is a general investigation goal (global context). We assume there is a general investigation goal to guide the whole sensemaking process. For example, we know there is a murder case (known

victim, time, and location), or we suspect there is a potential terrorist attack (undecided who, what, where, when).
A2. Source materials are narrative texts. Crime plots are diffused or obfuscated in the text with noise, in some latent structure. The documents are modularized in some uniform way, and can be disassembled into sentences, paragraphs or whole documents.
A3. Entities and their relationships are from the source texts. The key entities constructing the crime plots are all in the text, but relationships between them may be more or less explicit, which is why algorithms are not enough to uncover the hidden plots.
A4. No external information is required to solve the case. Common sense knowledge is enough to understand and analyse the source material. All necessary information is covered in the narrative texts and agents do not need to consult external information sources.
A5. Privacy and confidentiality is out of scope. Although we are using fictional crime-related evidence data as the example dataset, our main focus is the analysis of text data. For confidential datasets, the strategy could be applied within a private crowd (e.g. employees).

D.2 Assumptions about Reportable Results

A6. Assume an investigation report. Since the initial data input assumes crime plots are hidden in narrative texts, the final outputs should be an investigation result reportable to a potential client.
A7. Event description fulfills formula for complete stories. The final results should conform to a simple template of a complete story comprised of: who, what (method of crime), where, and when with necessary supporting evidence.
A8. Each answer component has a finite number of options. There is a finite number of options for each of the four W's mentioned above, based on the content of the dataset.
A9. Missing links mean the solution is not correct. In terms of evaluation, if any of the W's are missing or any of the connections between entities are missing, the results are considered incomplete.
A10. Simpler explanations are preferred. In the final stage of analysis, among several candidate hypotheses with the same level of correctness, simpler candidates are preferred.
A11. Constraints are enforced by resources. There are limits like elapsed time, number of guesses allowed with given resources that constrains the analysis procedure. The pipeline cannot keep running forever and must stop when the limits are met.

ACKNOWLEDGMENTS

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

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] L. Bavoil, S. P. Callahan, P. J. Crossno, J. Freire, C. E. Scheidegger, C. T. Silva, and H. T. Vo. 2005. VisTrails: enabling interactive multiple-view visualizations. In *VIS 05. IEEE Visualization, 2005.* 135–142. https://doi.org/10.1109/VISUAL.2005.1532788

[3] Michael S Bernstein, Greg Little, Robert C Miller, Björn Hartmann, Mark S Ackerman, David R Karger, David Crowell, and Katrina Panovich. 2010. Soylent: a word processor with a crowd inside. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM Press, New York, New York, USA, 313–322. https://doi.org/10.1145/1866029.1866078

[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. https://doi.org/10.1109/VAST.2008.4677362

[5] 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, 1078–1088. https://doi.org/10.1016/j.ijhcs.2013.07.004

[6] L. Elisa Celis, Sai Praneeth Reddy, Ishaan Preet Singh, and Shailesh Vaya. 2016. Assignment Techniques for Crowdsourcing Sensitive Tasks. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. ACM, New York, NY, USA, 836–847. https://doi.org/10.1145/2818048.2835202

[7] Joel Chan, Steven Dang, and Steven P. Dow. 2016. Comparing Different Sensemaking Approaches for Large-Scale Ideation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 2717–2728. https://doi.org/10.1145/2858036.2858178

[8] Joel Chan, Steven Dang, and Steven P. Dow. 2016. IdeaGens: Enabling Expert Facilitation of Crowd Brainstorming. *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion - CSCW '16 Companion*, 13–16. https://doi.org/10.1145/2818052.2874313

[9] 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 Press, New York, New York, USA, 3180–3191. https://doi.org/10.1145/2858036.2858411

[10] Wen-Huang Cheng and David Gotz. 2008. Context-based page unit recommendation for web-based sensemaking tasks. In *Proceedingsc of the 13th international conference on Intelligent user interfaces - IUI '09*. ACM Press, New York, New York, USA, 107. https://doi.org/10.1145/1502650.1502668

[11] 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 Press, New York, New York, USA, 1999. https://doi.org/10.1145/2470654.2466265

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

[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. *VAST 10 - IEEE Conference on Visual Analytics Science and Technology 2010, Proceedings*, 107–114. https://doi.org/10.1109/VAST.2010.5652932

[14] Robert M Clark. 2013. *Intelligence analysis: a target-centric approach* (4th ed.). CQ Press, Thousand Oaks, Calif.

[15] Kristin A Cook and James J Thomas. 2005. *Illuminating the path: The research and development agenda for visual analytics*. Technical Report. Pacific Northwest National Lab.(PNNL), Richland, WA (United States).

[16] R. Jordon Crouser and Remco Chang. 2012. An Affordance-Based Framework for Human Computation and Human-Computer Collaboration. *IEEE Transactions on Visualization and Computer Graphics* 18, 12, 2859–2868. https://doi.org/10.1109/TVCG.2012.195

[17] doggyxp. 2018. Braingle: 'The Case of the Mischief Maker' Brain Teaser. (2018). http://www.braingle.com/brainteasers/17325/the-case-of-the-mischief-maker.html

[18] 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

[19] Philipp Drieger. 2013. Semantic Network Analysis as a Method for Visual Text Analytics. *Procedia - Social and Behavioral Sciences* 79, 4–17. https://doi.org/10.1016/j.sbspro.2013.05.053

[20] Alex Endert, Patrick Fiaux, and Chris North. 2012. Semantic interaction for visual text analytics. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, 473–482. https://doi.org/10.1145/2207676.2207741

[21] 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

[22] Brooke Foucault Welles and Weiai Xu. 2018. Network visualization and problem-solving support: A cognitive fit study. *Social Networks* 54, 162–167. https://doi.org/10.1016/J.SOCNET.2018.01.005

[23] D Gary. 2011. Patty Wagstaff's Second Act. *Air & Space Smithsonian* 26, 3, 20–25.

[24] Daniel Gigone and Reid Hastie. 1993. The common knowledge effect: Information sharing and group judgment. *Journal of Personality and social Psychology* 65, 5 (1993), 959.

[25] D. Gotz, M. X. Zhou, and V. Aggarwal. 2006. Interactive Visual Synthesis of Analytic Knowledge. In *2006 IEEE Symposium On Visual Analytics Science And Technology*. 51–58. https://doi.org/10.1109/VAST.2006.261430

[26] 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 - CSCW '16*. ACM Press, New York, New York, USA, 287–301. https://doi.org/10.1145/2818048.2820071

[27] 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 Press, New York, New York, USA, 2721. https://doi.org/10.1145/2470654.2481376

[28] Catherine Grevet and Eric Gilbert. 2015. Piggyback Prototyping: Using Existing, Large-Scale Social Computing Systems to Prototype New Ones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 4047–4056. https://doi.org/10.1145/2702123.2702395

[29] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big Picture Thinking in Small Pieces. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, 2258–2270. https://doi.org/10.1145/2858036.2858364

[30] Jeffrey Heer, Fernanda B. Viégas, and Martin Wattenberg. 2007. Voyagers and voyeurs. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '07*. ACM Press, New York, New York, USA, 1029. https://doi.org/10.1145/1240624.1240781

[31] Richards J. Heuer and Center for the Study of Intelligence (U.S.). 1999. *Psychology of intelligence analysis*. Lulu. com. 184 pages.

[32] F Hughes and D Schum. 2003. Discovery-proof-choice, the art and science of the process of intelligence analysis-preparing for the future of intelligence analysis. *Washington, DC: Joint Military Intelligence College*.

[33] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S Bernstein. 2016. Mechanical Novel: Crowdsourcing Complex Work through Revision. In *Proceedings of the 20th ACM Conference on Computer Supported Cooperative Work & Social Computing*. 233–245. https://doi.org/10.1145/2998181.2998196

[34] Aniket Kittur, Susheel Khamkar, Paul André, and Robert Kraut. 2012. CrowdWeaver. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work - CSCW '12*. ACM Press, New York, New York, USA, 1033. https://doi.org/10.1145/2145204.2145357

[35] Aniket Kittur, Andrew M. Peters, Abdigani Diriye, and Michael Bove. 2014. Standing on the Schemas of Giants: Socially Augmented Information Foraging. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14 (CSCW '14)*. ACM Press, New York, New York, USA, 999–1010. https://doi.org/10.1145/2531602.2531644

[36] Aniket Kittur, Andrew M. Peters, Abdigani Diriye, Trupti Telang, and Michael R. Bove. 2013. Costs and benefits of structured information foraging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. ACM Press, New York, New York, USA, 2989. https://doi.org/10.1145/2470654.2481415

[37] Aniket Kittur, Boris Smus, and Robert Kraut. 2011. CrowdForge Crowdsourcing Complex Work. *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems - CHI EA '11*, 1801. https://doi.org/10.1145/1979742.1979902

[38] Jean Lave and Etienne Wenger. 1991. *Situated learning: Legitimate peripheral participation*. Cambridge university press.

[39] Tianyi Li, Asmita Shah, Kurt Luther, and Chris North. 2018. Crowdsourcing Intelligence Analysis with Context Slices. *Chi '18 Sensemaking Workshop*.

[40] Greg Little, Lydia B Chilton, Max Goldman, and Robert C Miller. 2010. Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 68–76. https://doi.org/10.1145/1837885.1837907

[41] Kurt Luther, Scott Counts, Kristin B. Stecher, Aaron Hoff, and Paul Johns. 2009. Pathfinder: an online collaboration environment for citizen scientists. In *Proceedings of the 27th international conference on Human factors in computing systems*. ACM, Boston, MA, USA, 239–248. https://doi.org/10.1145/1518701.1518741

[42] 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*. ACM, 473–485. https://doi.org/10.1145/2675133.2675283

[43] Thomas W. Malone, Jeffrey V. Nickerson, Robert J. Laubacher, Laur Hesse Fisher, Patrick de Boer, Yue Han, and W. Ben Towne. 2017. Putting the Pieces Back Together Again. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing - CSCW '17*. ACM Press, New York, New York, USA, 1661–1674. https://doi.org/10.1145/2998181.2998343

[44] Adam Marcus and Aditya Parameswaran. 2015. Crowdsourced Data Management: Industry and Academic Perspectives. *Foundations and Trends in Databases* 6, 1-2 (Dec. 2015), 1–161. https://doi.org/10.1561/1900000044

[45] Johnny Nhan, Laura Huey, and Ryan Broll. 2017. Digilantism: An Analysis of Crowdsourcing and the Boston Marathon Bombings. *The British Journal of Criminology* 57, 2 (2017), 341–361. https://doi.org/10.1093/bjc/azv118

[46] Vineet Pandey, Amnon Amir, Justine Debelius, Embriette R. Hyde, Tomasz Kosciolek, Rob Knight, and Scott Klemmer. 2017. Gut Instinct: Creating Scientific Theories with Online Learners. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 6825–6836. https://doi.org/10.1145/3025453.3025769

[47] Devi Parikh and C. Lawrence Zitnick. 2011. Human-Debugging of Machines. In *Neural Information Processing Systems*. 1–5. https://filebox.ece.vt.edu/~parikh/human_debugging/

[48] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of International Conference on Intelligence Analysis* 2005, 2–4. https://doi.org/10.1007/s13398-014-0173-7.2 arXiv:gr-qc/9809069v1

[49] Nikhil Sharma. 2009. Sensemaking handoff: When and how? *Proceedings of the American Society for Information Science and Technology* 45, 1, 1–12. https://doi.org/10.1002/meet.2008.1450450234

[50] Nikhil Sharma and George Furnas. 2009. Artifact usefulness and usage in sensemaking handoffs. *Proceedings of the American Society for Information Science and Technology* 46, 1, 1–19. https://doi.org/10.1002/meet.2009.1450460219

[51] Stephen Soderland. 1999. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning* 34, 1/3, 233–272. https://doi.org/10.1023/A:1007562322031

[52] John Stasko, Carsten Görg, and Zhicheng Liu. 2008. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. *Information Visualization* 7, 2, 118–132. https://doi.org/10.1057/palgrave.ivs.9500180

[53] Garold Stasser and William Titus. 1985. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology* 48, 6 (1985), 1467.

[54] Maoyuan Sun, Peng Mi, Chris North, and Naren Ramakrishnan. 2016. BiSet: Semantic Edge Bundling with Biclusters for Sensemaking. *IEEE Transactions on Visualization and Computer Graphics* 22, 1, 310–319. https://doi.org/10.1109/TVCG.2015.2467813

[55] Yla Tausczik and Mark Boons. 2018. Distributed Knowledge in Crowds: Crowd Performance on Hidden Profile Tasks. In *Twelfth International AAAI Conference on Web and Social Media*. https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17817

[56] Alice Toniolo, Timothy J. Norman, Anthony Etuk, Robin Wentao Ouyang, Nir Oren, Timothy Dropps, John A. Allen, Federico Cerutti, Robin Wentao Ouyang, Mani Srivastava, Nir Oren, Timothy Dropps, John A. Allen, and Paul Sullivan. 2015. Supporting Reasoning with Different Types of Evidence in Intelligence Analysis. In *AAMAS '15 Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '15)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 781–789. http://dl.acm.org/citation.cfm?id=2772879.2773254http://dl.acm.org/ft_gateway.cfm?id=2773254&type=pdf

[57] 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 Press, New York, New York, USA, 3523–3537. https://doi.org/10.1145/3025453.3025811

[58] Vasilis Verroios and Michael S Bernstein. 2014. Context Trees: Crowdsourcing Global Understanding from Local Views. *Hcomp 2014* 1351131, 210–219. http://ilpubs.stanford.edu:8090/1105/

[59] Hanna M. Wallach and Hanna M. 2006. Topic modeling. In *Proceedings of the 23rd international conference on Machine learning - ICML '06*. ACM Press, New York, New York, USA, 977–984. https://doi.org/10.1145/1143844.1143967

[60] Jiannan Wang, Tim Kraska, Michael J. Franklin, and Jianhua Feng. 2012. CrowdER: Crowdsourcing Entity Resolution. 1483–1494. https://doi.org/10.14778/2350229.2350263

[61] Ashley Wheat, Simon Attfield, and Bob Fields. 2016. Developing a Model of Distributed Sensemaking: A Case Study of Military Analysis. *Informatics* 3, 1, 1. https://doi.org/10.3390/informatics3010001

[62] Mark E Whiting, Dilrukshi Gamage, Snehalkumar S Gaikwad, Aaron Gilbee, Shirish Goyal, Alipta Ballav, Dinesh Majeti, Nalin Chhibber, Angela Richmond-Fuller, Freddie Vargus, et al. 2016. Crowd guilds: Worker-led reputation and feedback on crowdsourcing platforms. *arXiv preprint arXiv:1611.01572*. https://doi.org/10.1145/2998181.2998234

[63] Wesley Willett, Jeffrey Heer, Joseph Hellerstein, and Maneesh Agrawala. 2011. CommentSpace: structured support for collaborative visual analysis. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 3131–3140. https://doi.org/10.1145/1978942.1979407

[64] William Wright, David Schroh, Pascale Proulx, Alex Skaburskis, and Brian Cort. 2006. The sandbox for analysis. In *Proceedings of the SIGCHI conference on Human Factors in computing systems - CHI '06*. ACM Press, New York, New York, USA, 801. https://doi.org/10.1145/1124772.1124890

[65] Hao Wu, Michael Mampaey, Nikolaj Tatti, Jilles Vreeken, M Shahriar Hossain, and Naren Ramakrishnan. 2012. Where do I start?: algorithmic strategies to guide intelligence analysts. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*. ACM, 3. https://doi.org/10.1145/2331791.2331794

[66] Anbang Xu, Shih-Wen Huang, and Brian Bailey. 2014. Voyant: Generating Structured Feedback on Visual Designs Using a Crowd of Non-Experts. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*, 1433–1444. https://doi.org/10.1145/2531602.2531604

[67] Roman V Yampolskiy. 2013. Turing test as a defining feature of AI-completeness. In *Artificial intelligence, evolutionary computing and metaheuristics*. Springer, 3–17. https://doi.org/10.1007/978-3-642-29694-9_1

[68] Haoqi Zhang, Edith Law, Rob Miller, Krzysztof Gajos, David Parkes, and Eric Horvitz. 2012. Human computation tasks with global constraints. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 217–226. https://doi.org/10.1145/2207676.2207708

[69] Jian Zhao, Michael Glueck, Petra Isenberg, Fanny Chevalier, and Azam Khan. 2017. Supporting Handoff in Asynchronous Collaborative Sensemaking Using Knowledge-Transfer Graphs. In *IEEE Transactions on Visualization and Computer Graphics*, Vol. 24. 1–1. https://doi.org/10.1109/TVCG.2017.2745279

[70] Haiyi Zhu, Steven P. Dow, Robert E. Kraut, and Aniket Kittur. 2014. Reviewing versus doing. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14*. ACM Press, New York, New York, USA, 1445–1455. https://doi.org/10.1145/2531602.2531718

Received April 2018; revised July 2018; accepted September 2018