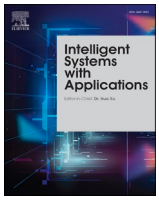
Intelligent Systems with Applications 17 (2023) 200151

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications



Human-machine collaboration in intelligence analysis: An expert evaluation☆

Alice Toniolo [a,*], Federico Cerutti [b,c], Timothy J. Norman [d], Nir Oren [e], John A. Allen [f], Mani Srivastava [g], Paul Sullivan [h]

[a] University of St Andrews, School of Computer Science, UK
[b] Università degli Studi di Brescia, Dipartimento di Ingegneria dell'Informazione, Italy
[c] Cardiff University, Crime and Security Research Institute, UK
[d] University of Southampton, School of Electronics and Computer Science, UK
[e] University of Aberdeen, Department of Computing Science, UK
[f] Honeywell, USA
[g] University of California, Los Angeles, Electrical and Computer Engineering, USA
[h] INTELPOINT Incorporated, USA

A B S T R A C T

In this paper we illustrate how novel AI methods can improve the performance of intelligence analysts. These analysts aim to make sense of — often conflicting or incomplete — information, weighing up competing hypotheses which serve to explain an observed situation. Analysts have access to numerous visual analytic tools which support the temporal and/or conceptual structuring of information and collection, and support the evaluation of alternative hypotheses. We believe, however, that there are currently no tools or methods which allow analysts to combine the recording and interpretation of information, and that there is little understanding about how software tools can facilitate the hypothesis formation process. Following the identification of these requirements, we developed the CISpaces (Collaborative Intelligence Spaces) decision support tool in collaboration with professional intelligence analysts. CISpaces combines multiple AI-based methods including argumentation theory, crowdsourced Bayesian analysis, and provenance recording. We show that CISpaces is able to provide support to analysts by facilitating the interpretation of different types of evidence through argumentation-based reasoning, provenance analysis and crowdsourcing. We undertook an experimental analysis with intelligence analysts which highlights three key points. (1) The novel, principled AI methods implemented in CISpaces advance performance in intelligence analysis. (2) While designed as a research prototype, analysts benchmarked it against their existing software tools, and we provide results suggesting intention to adopt CISpaces in analysts' daily activities. (3) Finally, the evaluation highlights some drawbacks in CISpaces. However, these are not due to the technologies underpinning the tool, but rather in its lack of integration with existing organisational standards regarding input and output formats. Our evaluation with intelligence analysts therefore demonstrates the potential impact that an integrated tool building on state-of-the-art AI techniques can have on the process of understanding complex situations, and on how such a tool can help focus human effort on identifying more credible interpretations of evidence.

1. Introduction

An intelligence analyst's job is to construct coherent hypotheses despite significant gaps and inconsistencies in gathered evidence, and present them clearly to decision-makers to inform their interventions (Heuer, 1999).

Current automated systems to support the day-to-day practice of analysts are (almost) exclusively focussed on two aspects of the problem. The first is data collection, aggregation and visualisation (IBM, 2017; Wright et al., 2006). Such tools help analysts collate, inspect and interact with a large dataset, and support the identification of relationships, for example through link analysis (Prunckun, 2010). Recently, crowdsourcing tools that enable the public to contribute information have also been introduced to integrate more traditional intelligence collection approaches (Stottlemyre, 2015). The second problem on which tools focus involves listing and weighing up alternative hypotheses (Heuer, 1999), through automated analysis of competing hypotheses (Burton and Knowles, 2010; Schrag et al., 2016; Stefik, 2014; Tecuci et al., 2010). This analysis requires that all alternative hypotheses be identified from available evidence, and, if aided by automated inferential reasoners such as Bayesian networks (Schrag et al., 2016), the tools also require that each (aggregated) piece of evidence is given a weight or a

☆ Dedicated to the memory of Paul Sullivan.

* Corresponding author.
E-mail address: a.toniolo@st-andrews.ac.uk (A. Toniolo).

https://doi.org/10.1016/j.iswa.2022.200151
Received 21 April 2022; Received in revised form 27 August 2022; Accepted 8 November 2022
Available online 15 November 2022

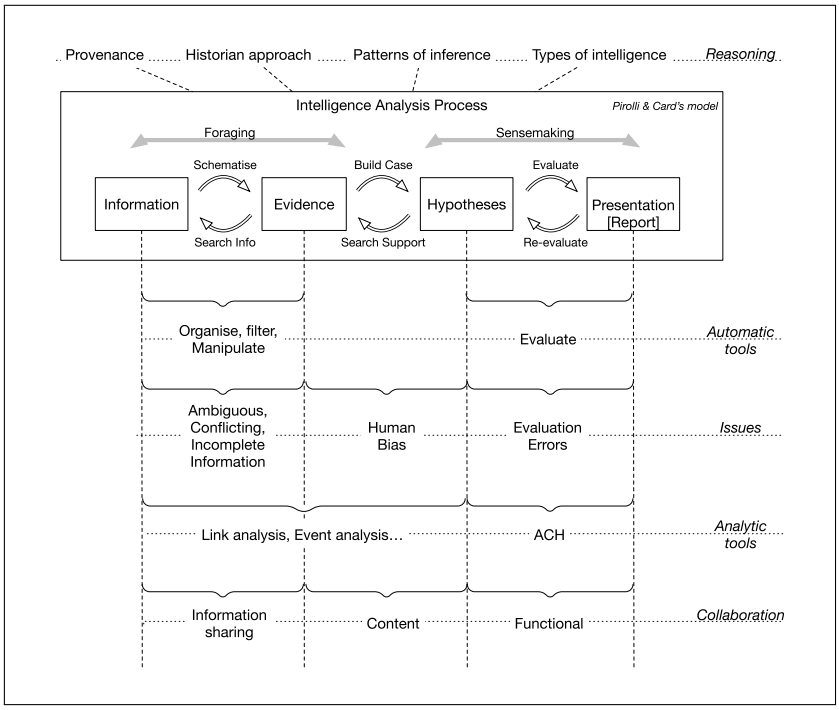


Fig. 1. A visualisation of Intelligence Analysis Issues and Requirements extending Pirolli and Card (2005).

degree of certainty.

We observe, however, that there is a gap in technology that supports the process that analysts perform after inspecting the data, and before the identification of hypotheses. The task of the analyst here involves the structuring of evidence in a consistent manner to select plausible hypotheses. This is currently done manually, supported only by generic spreadsheet and text-processing tools. The challenge – which we seek to address in this research – is to understand how automated reasoning can best complement human expertise in this evidential reasoning process.

Experienced analysts currently identify plausible hypotheses using a combination of manual approaches to assess available evidence, establish what information is credible, and understand what additional evidence may be required or what questions to ask to determine plausibility. This activity may be time critical so as to enable effective situational understanding, which poses significant challenges for individual analysts. The volume and variety of information that analysts must consider is significant, and, evidence may be unreliable or conflicting, with important information missing. Collaboration may be used to provide peer-review, sharing the burden of analysis and helping in the validation of conclusions (Heuer, 1999). Such collaboration, however, requires analysts to work with a common model and a consistent world-view, which is hard to achieve in the real world.

When data is diverse and comes from different sources, analysts must reason about the reliability of the evidence leading to claims from information such as how, where, when and by whom the evidence was gathered and analysed. Cognitive biases may inadvertently be introduced in the process, preventing an analyst from drawing accurate conclusions. This process of interpreting evidence relies heavily on the expertise and training of analysts, and there is a distinct lack of methods to ease the high cognitive burden involved in forming hypotheses. Furthermore, there is a general lack of understanding of how the hypothesis formation process works, as it is not normally recorded, making it difficult for senior analysts to pass on their analytical skills to trainees. The analytical process is also resistant to automation due to the significant knowledge engineering effort required to process data and express reasoning patterns (Llinas, 2013).

In this paper, we illustrate how novel AI methods, based on a combination of argumentation theory, crowdsourcing and provenance reasoning, can contribute to improved performance in intelligence analysis. While existing systems care mostly about information presentation and collection (Billman et al., 2006; Burton and Knowles, 2010; IBM, 2017; Wright et al., 2006), we co-designed our software tool — Collaborative Intelligence Spaces (CISpaces) — with intelligence analysts to focus on the sensemaking activities around forming hypotheses from available evidence using patterns of defeasible inferences, or *argumentation schemes* (Walton et al., 2008). Our formal evaluation of CISpaces using the Technology Acceptance Model (TAM) (Davis, 1989; Venkatesh and Bala, 2008; Venkatesh and Davis, 2000) provides evidence that intelligence analysts benefit from the support they receive from the tool in their sensemaking activities.

The contributions of this paper are thus many fold:

- In Section 3, we provide a blueprint for further co-design of artificial intelligence driven tools, by showing how to govern the process for a successful outcome.
- In Section 4, we expand on our preliminary conference paper (Toniolo et al., 2015) to illustrate the delicate interconnection between the various artificial intelligence techniques utilised and extended to achieve the co-designed objectives. In particular:
  - we advance the engineering of *argumentation-based reasoning* (Prakken, 2010) to identify plausible hypotheses as sets of acceptable arguments;
  - we show how to argue with, and about, *crowd-sourced* information (Brabham, 2008; Kamar et al., 2012; Whitehill et al., 2009) pre-analysed using *Bayesian analysis*;
  - we embed *provenance* analysis (Hartig and Zhao, 2009) in the argumentative process to establish the credibility of hypotheses.
- In Section 5,we provide empirical evidence that intelligence analysts benefit from the unique mixture of formal argumentation theory, crowdsourcing support, and provenance recording provided in CISpaces, for the first time, using the Technology Acceptance Model (TAM) (Davis, 1989; Venkatesh and Bala, 2008; Venkatesh and Davis, 2000) in an argumentation-based system.

Our results suggest that the novel, principled AI methods implemented in CISpaces may advance performance in intelligence analysis. Despite having designed CISpaces as a basic research prototype (Technology Readiness Level 3 — TRL 3), during their evaluation, analysts benchmarked the quality of its features against commercial systems they use everyday: we compare against them in Section 6.

We collected evidence suggesting that the AI methods implemented in CISpaces can have a behavioural effect on the intention to adopt CISpaces by end users. The analysts' evaluation highlights drawbacks in CISpaces that predominantly result from the interface between the tool and data sources and from aspects of the user interface (rather than the underlying AI methods). We, therefore, conclude that for successful adoption by the intelligence analysis community, CISpaces will need data integration with existing organisational standards both for the input and the output of information. These and other engineering and usability aspects, while being essential for commercialisation, are beyond the scope of this paper.

2. Challenges of intelligence analysis

Intelligence analysis is the application of individual and collective cognitive methods to evaluate, integrate, and interpret information about situations and events, aiming to provide warning regarding potential threats or to identify opportunities (Heuer, 1999). Various types of intelligence can be distinguished based on source. HUMINT (human intelligence), for example, is intelligence gathered from human sources. IMINT (imagery intelligence) is derived from image or video sources. OSINT (open source intelligence) is acquired from sources such as social media (Prunckun, 2010) and more recent types of intelligence include for example crowdsourced intelligence, made up of structured information acquired from or volunteered by the general public (Stottlemyre, 2015). Analysts often specialise in a specific type of intelligence, and may be focused on particular objectives (e.g., tracking activities of a criminal organisation). In the military context, strategic analysts focus on studying long term objectives and intentions of foreign actors, while operational and tactical analysts are focused on supporting specific actions or providing timely responses to emerging situations. The examples used in this paper focus primarily on the operational and tactical analysis of HUMINT, although related work has also considered field intelligence (Toniolo et al., 2016) and OSINT (Cerutti et al., 2018b).

To *conceptualise* the process, Pirolli and Card's (2005) model of intelligence analysis is one of the most influential in training and practice. It consists of two high-level iterative loops: *foraging* for *information* which is collected, filtered and collated into *evidence files*; and *sense-making*, where the evidence files are interpreted through logical reasoning by drawing inferences and identifying *hypotheses*, which are then brought together to form a coherent explanation of the situation. A sketch of this process is shown within the box at the top of Fig. 1. The top row represents general features that characterise or influence the reasoning process during analysis. The other rows in this figure represent concepts related to different dimensions of intelligence analysis corresponding to a specific phase of analysis represented by the curly bracket in the column. More generally, this figure provides a reference for the components and challenges which inform the remainder of our discussion, and are ordered according to the topics covered in this Section.

Hypotheses formation and automation. While automation has been deployed in previous research for information collection (IBM, 2017; Wright et al., 2006) and evaluation of hypotheses (Billman et al., 2006; Burton and Knowles, 2010; Stefik, 2014), the analysis of information to form hypotheses remains a mostly human-driven task. This is difficult to fully automate due to the significant effort required to engineer both the explicitly available data for a situation, and the implicit knowledge that analysts use to form a hypothesis (Waltz, 2003). Heuer (1999) argues that these hypotheses are often created by analysts by adopting a *historian approach* to reconstruct a narrative looking at the available information to explain events. Recent experiments confirmed the use of narratives for collaborative analysis (Saletta et al., 2020). This approach is also advocated by Bex and Verheij (2012) in formalising reasoning with respect to a criminal investigation, where the evidence is used to establish the plausibility of an existing story.

The output of an analysis normally consists of an *intelligence report* that presents the most plausible hypothesis. Many hypotheses, however, are considered during the process of making sense of the situation. For example, Klein et al. (2006), while modelling the cognitive behaviour of analysts argue that "*people don't engage in simple mental operations of confirming or disconfirming a hypothesis.*" They propose a model in which data are interpreted according to a *frame* (such as a story, a diagram or a map) that is questioned and reframed as new information and links are formed. Their observations indicate that experts consider multiple, competing frames while making sense of events to establish those most accurate. More recently, Baber et al. (2016) used this framework to better understand the use of frames in the context of intelligence analysis. Two groups of participants, professional analysts and students, were observed and compared during an exercise to identify suspects of criminal activities. They concluded that "*tools that support the collation of information…might help with* down-collection *of data but do not provide support for* conflict and corroboration *or for* hypothesis-exploration" where down-collection means sampling the available data for material deemed to be relevant to the analysis (Baber et al., 2016). This is attributed to differences in how frames or representations are constructed, used and shared depending on participants' expertise. We note that there is little prior research, however, on how tools can facilitate the *externalisation* of an analyst's *reasoning* process while forming frames or hypotheses.

Analysis Issues. One of the main challenges to such facilitation is the need to identify factors that contribute or hamper the externalisation of the reasoning process. Of primary importance among these factors is the role of evidence. Ambiguous, conflicting, unreliable or incomplete evidence might lead to multiple alternative hypothetical explanations of a situation, and such evidence in turn may be used to evaluate the strength of the different hypotheses due to the cyclic nature of hypotheses formation and evaluation.

Evidence may be ambiguous or conflicting for a variety of reasons. Heuer (1999) claims that most human-sourced information is second-hand at best, and furthermore this might be reported by sources that have varying degrees of trustworthiness. Information might have been purposely manipulated or simply reported by several sources from alternative points of view. Evaluating the provenance of this information is fundamental to forming a more objective assessment (Toniolo et al., 2014). When information is sought specifically to answer questions and requirements, such as for example through crowdsourcing (Stottlemyre, 2015), its quality varies significantly and analysts must employ methods to aggregate results and establish the truthfulness of this evidence. In addition, the analysis might be hampered by biases, such as confirmation bias, whereby an analyst (often unconsciously) prioritises information that confirms current beliefs, which may affect the accuracy of conclusions. Other biases are also associated with human working memory and its difficulties in remembering all the underlying reasons for an explanation, as well as the difficulties in revising links that have already been made on the basis of new information and its credibility (Heuer, 1999).

Analytic tools. Intelligence analysts are specifically trained in developing critical thinking through a variety of analytic approaches and techniques (Heuer, 1999; Prunckun, 2010; United Nations, 2011; US Army, 2006), to address the challenging process of identification of evidence and formation of hypotheses. These approaches are derived from logical and statistical models of reasoning, and are concerned with both analysing and interpreting data, and understanding and avoiding biases of different sorts. Examples include: *link analysis* that aims at identifying relationships between entities, resources, and events;

red-team versus blue-team exercises where teams play the roles of attacker (red) and defender (blue); and SWOT (Strength, Weaknesses, Opportunities and Threats) analyses. The resulting observations from these approaches are patterns and relationships that link information; such patterns vary significantly, however, they share similarities in that logical inferences are made among events (in line with the historian view) and other elements of analysis related to these events such as entities, resources, indicators, etc. These patterns constitute inferences that are fundamental to hypotheses formation, however, the reasoning step that leads an analyst to make the observations and then construct a hypothesis is primarily manual and often an internal process, and as such unrecorded.

Hypotheses Assessment. Once hypotheses are identified, and in order to evaluate these against evidence, the Analysis of Competing Hypotheses (ACH) (Heuer, 1999) approach is considered fundamental among the analytic methods employed in both training and practice. This approach aims to provide a reliable evaluation of hypotheses and support the mitigation of reasoning biases. The application of ACH to a problem promotes a systematic and objective approach, which aids in the management of complexities inherent in real-world scenarios. A table is used to weigh alternative explanations by asserting whether there is some evidence in support of a hypothesis. Typically, ACH is performed manually, and so demands significant cognitive effort: the analyst is required to retain multiple hypotheses in memory together with the evidence acquired to support these hypotheses. The complexity of this process leads to a high risk of erroneous assessment of the plausibility of hypotheses, which can have substantial effects on the quality of an analysis. For example, the report of the Iraq inquiry (Chilcot, 2016), when referring to whether Iraq possessed weapons of mass destruction (WMD) in 2003 states that "*Intelligence and assessments were used to prepare material to be used to support Government statements in a way which conveyed certainty without acknowledging the limitations of the intelligence*" and that "*The question is whether, in doing so, they conveyed more certainty and knowledge than was justified*". The effect of the errors made in the assessment of plausibility of this hypothesis has been so significant that it is now a textbook exercise when training intelligence analysts (Lahneman and Arcos, 2014).

Collaboration. In addition to the challenges listed above, analysis of complex, real-world situations is typically team-based within or across agencies (Kang and Stasko, 2011). Collaboration brings advantages in providing diverse and complimentary perspectives and expertise, and mitigating biases. The differences between groups of analysts in expertise, resources, and capabilities may, however, lead to conflicts of opinions with regards to what hypotheses are plausible. While out of scope of the current work, we note that issues including policies restricting information sharing may also come into play (Verma, 2010). Externalisation of the reasoning process is key in collaborative analysis, and one of the most critical problems in supporting collaboration is to identify what part of the process should be *externalised* or shared with other collaborators (Mahyar and Tory, 2014). Existing models focus on *information sharing* and *functional* collaboration (Kang and Stasko, 2011). Information sharing in this context is the process of identifying information that may be of interest to others and sharing that which is relevant. This is very much an activity which occurs early on in the foraging loop, and is concerned with collecting and filtering information. Functional collaboration in this context is the process of editing reports to complete an analysis, and hence telling a combined story. These collaborative tasks are at either end of the Pirolli and Card conceptual model (Pirolli and Card, 2005). An additional type of collaboration, referred to as the *content* level collaboration (Kang and Stasko, 2011), sees analysts work together to structure and link information and evidence. For analytic problems to be addressed in time-stressed contexts, analysts from different organisations with different expertise, perspectives and resources need to work together to form these hypotheses within a common framework. Despite limited prior research on how content-level collaboration can be supported, this type of collaboration would be most beneficial as it would permit the elaboration and sharing of alternative hypotheses across a team, enabling more effective criticism and robust reasoning.

From the challenges highlighted within intelligence analysis, we must distil the most important and unaddressed issues. This is the focus of the next section, where we consider and prioritise the most important analyst requirements which have not yet been adequately addressed by existing tools.

3. Co-design with analysts

In this research, we study how novel AI methods advance performance in intelligence analysis by providing analysts with computational support in *externalising their reasoning* while building and comparing alternative hypotheses from interpretation of reports, observations, inferences, or crowdsourced data (potentially with various degrees of reliability) and conflicting or corroborating evidence. Addressing this challenge by providing meaningful computational support will bring advantages both for individual analysts and analysis teams. At the individual level, there is potential to provide automation in checking the validity of the reasoning process, support the understanding of where the information has come from, and to focus reasoning on important areas of uncertainty. At the team or collaborative level, advantages may manifest in improved mutual understanding of how other analysts reason, and may help elucidate others' opinions of how evidence links together to form hypotheses.

Our objectives are, therefore:

- OJ1 — support the identification of what is believed to have happened from the evidence gathered (the *sensemaking process*);
- OJ2 — support the integration of different forms of explicitly requested information (the *crowdsourcing process*);
- OJ3 — support the assessment of information credibility according to the history of its collection and manipulation (the *provenance reasoning process*).

For the support provided to these underpinning processes to be effective, we aim to develop a principled approach that aligns with analysts' methods for analysis.

Identifying these research objectives is not sufficient. If we are to develop interventions that have the potential to be acceptable to and adopted by practitioners, we require a deeper understanding of the process of intelligence analysis as experienced by experts as well as their requirements and priorities. To achieve this, we closely collaborated with various groups of expert analysts in the US and UK to validate our objectives, co-develop a system to enact these objectives as well as a scenario to help understand the kind of tasks that analysts need support for, and finally to evaluate our system.

In the next subsections, we present the results of our work in exploring requirements in collaboration with experts to prepare the development of CISpaces. In particular, Section 3.1 discusses results from a focus group, Section 3.2 presents an intelligence analysis scenario which will be used throughout this paper, and Section 3.3 summaries the requirements presented in this section and emerging from our review in Section 2.

3.1. Requirement validation: Focus group

We now introduce the results of a focus group conducted to better understand the analysis process and refine and contextualise our objectives in line with analysts' experience. This focus group involved five experienced analysts from an international agency who consented to participate in this academic study and have their opinion analysed and reported in publications. Although this group was relatively small, participants brought extensive expertise to the study, and the discussion lasted for around three hours. The questions were exploratory in nature,

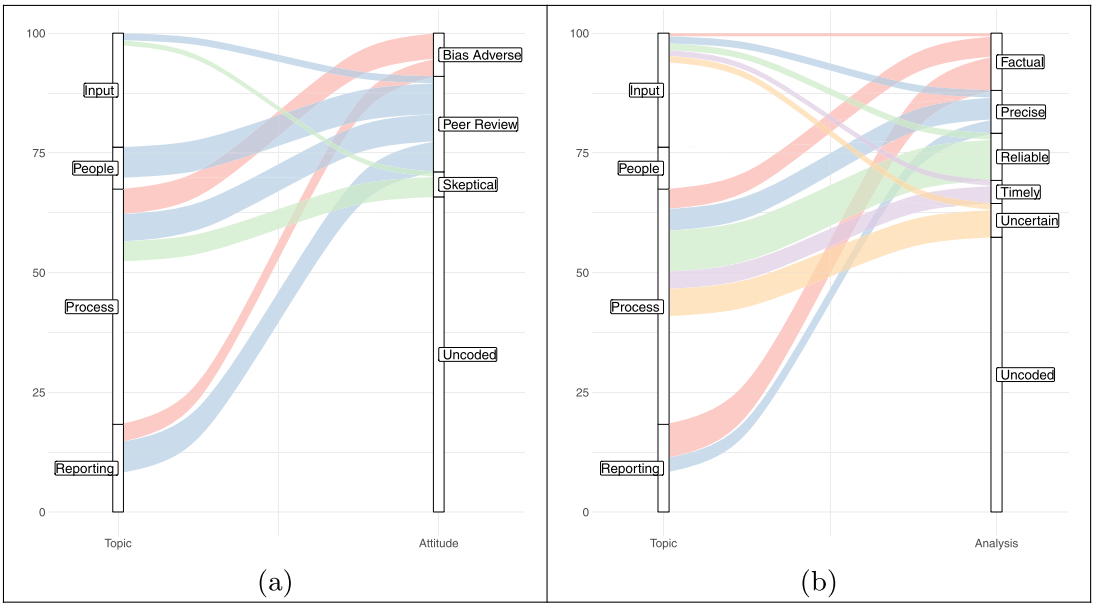


Fig. 2. Theme analysis grouped by topics on the left-hand axis: (a) Analysts' attitudes towards analysis; and (b) Quality features of analysis.

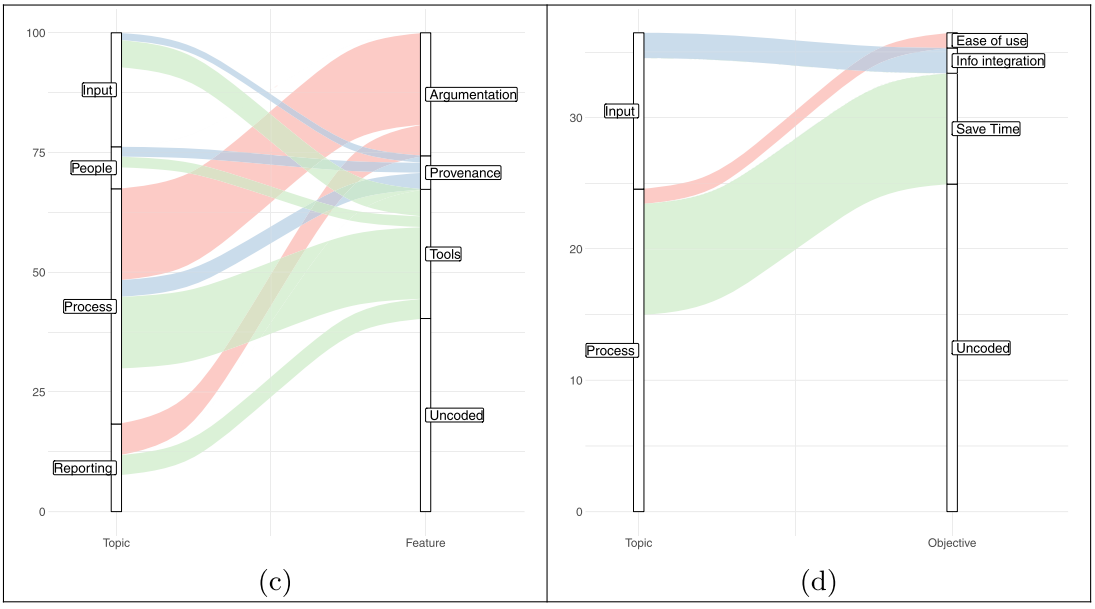


Fig. 3. Theme analysis grouped by topics on the left-hand axis: (c) Features of analytical process; and (d) Tool objectives and requirements.

but focused on four main topics:

- Inputs — the types of information collected, and issues of quality and credibility assessment.
- Reporting — outputs and conclusions delivered to decision-makers.
- Process — how analysts work in their daily activities.
- People — modes of collaboration among analysts.

The list of guiding questions is provided in Appendix D.1. The focus group discussion was recorded, transcribed and coded and we report here on insights gained that relate to our objectives (OJ1–OJ3).[1] Three main themes emerged: analysts' attitudes towards the analysis, quality of the analysis, and features and tools characterising the analytic process. In Figs. 2 and 3 we summarise the results of the coding using an adapted Sankey diagram. For each theme, the graphs highlight the most relevant concepts (right-hand axis) according to the four topics (left-hand axis). Note that in a Sankey diagram the width of links are proportional to the strengths of the associations according to the experimental data.

3.1.1. *Analysts' attitude towards the analysis*

Analysts, by training, are skeptical, and aware of potential biases. As indicated in Fig. 2(a), skepticism is maintained both for information received and during the process. Analysts invest significant effort in strategies to avoid biases, and are open to challenge, in particular through peer review. This highlights one important aspect of collaboration, which is aimed at the review of others' assessments during the analysis, as well as at the reporting phase. Peer review activities include, for example, the red-team versus blue-team technique (Section 2) and

[1] NVivo was used for the analysis (QSR International, 1999).

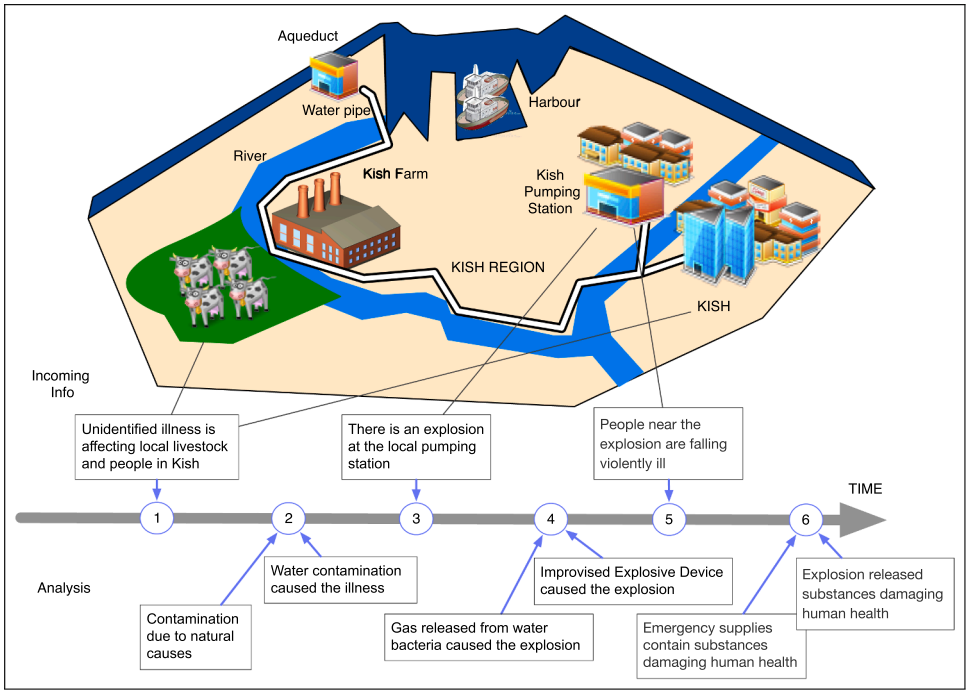


Fig. 4. An illustrative scenario with timeline.

seem to align to a *devil's advocate* type of dialogue, where a proponent proposes a specific position, and others seek to contradict it, in a process of evaluation and progressive elimination of alternatives. This is then reflected in the reporting, where the main hypothesis is elaborated, but then strengthened by the alternative hypotheses that have been discounted.

3.1.2. Quality of the analysis

Analyses have specific characteristics that ensure their quality (see Fig. 2(b)):

- Factual and Precise — given the aim is to increase situational awareness, explanations of what is happening must be grounded on facts. This is important with respect to inputs, the process and particularly in reporting. Access to primary source information was considered important in making precise assessments, but this is often difficult as analysts often have to rely on second-hand information.
- Reliable or Uncertain — the likelihood of a situation guides the prioritisation of assessment. Information is assessed on the basis of its likelihood and source reliability. In reports, information, hypotheses and their assessment are often characterised by specific wording expressing their likelihood.
- Timely — timeliness of information plays a crucial role in understanding a situation. Time of generation and time of receipt determine the cut off date for information when conducting derivative analyses. Different types of analysis at strategic, operational or tactical level have different requirements.

3.1.3. Features and tools characterising the analytical process

The focus group also explored attitudes around features and tools characterising the analytical process (Fig. 3). This exploration focused on understanding the features of, and requirements for, support tools. Fig. 3(c) indicates strong reference to both informal argumentation and provenance.

The analysis of the transcript showed strongly positive attitudes towards the argumentative nature of the analysis process. We note that analysts are, by training, exposed to concepts such as informal argumentation and mind mapping as methods to organise, structure and assess the quality of the intelligence reports prepared. This assessment is based on supporting evidence and facts, where no hypothesis is considered *wrong* but can be rebutted by alternative hypotheses or facts through personal evaluation or peer review. Argumentation also appears important in reporting, where arguments both for the main hypothesis and against alternatives considered play a role. In this work, we leverage the analysts' familiarity with informal argumentation to apply computational argumentation, as discussed in Section 4.

The provenance of information and the analysis are both crucial features for reliable assessment, particularly when this is derivative. Recording how an analysis was formed and how source reliability is considered was viewed as having potential for training.

The discussion around the potential for, and use of tools to support the analytical process drew heavily on participants' prior experience, but as indicated in Fig. 3(d) tools are viewed as important across all our four main topics (inputs, reporting, process and people). Areas of specific interest for future tools included support for organising and improving integration of input information; collaborating with other analysts for peer review and training; the process of assessment and report preparation. Fig. 3(b) summarises a small number of key concerns for novel tools; by far the most important being to save analysts' time.

The findings from our focus group exercise validates our objectives: analysts appreciate support in their sensemaking activities; value timely data; and treasure provenance of information.

3.2. An illustrative scenario

A realistic scenario for use in better understanding the demands and requirements of the analysis process, for demonstrating our approach and for evaluation was co-created with the help of two US expert analysts.

This scenario centres on an investigation into possible water

Table 1
Summary of Challenges and Requirements for Intelligence Analysis, in relation to the focus group (FG) and our objectives (OJ1,OJ2,OJ3). A symbol • in column FG indicates that the focus group raises the challenge in a specific row. In the OJ columns, • indicates that the objective is designed to address the respective challenge in the row.

Topic	Challenge	FG	OJ1	OJ2	OJ3
Input	Identify Information requirements			•	
Input	Deal with second hand information and their provenance	•			•
Input	Identify and collect factual information	•		•	
Input	Establish source reliability	•		•	•
Input/Process	Analyse results of information requirements			•	
Input/Process	Integrate and analyse conflicting, incomplete and uncertain information	•	•		
Process	Support the reconstruction of events (historian approach)		•		
Process	Support the iteration through information foraging and sensemaking		•	•	
Process	Construct multiple alternative hypotheses and keep them in working memory	•	•		
Process	Support hypotheses exploration and externalisation of the reasoning process		•		
Process	Support an informal argumentative approach to analysis	•	•		
Process	Mitigate bias in the analysis and encourage skepticism	•	•	•	•
People/Process	Support analysis through peer review to mitigate bias	•	•		•
People	Support collaboration in collection and integration of information			•	
People	Support collaboration in content of analysis		•		
People	Support collaboration for training junior analysts	•	•		•
People/Reporting	Support peer review of intelligence reports	•			
Reporting	Evaluation of hypotheses	•	•		
Reporting	Identification of the most plausible hypothesis	•			
Reporting	Provide timely analysis	•	•	•	•

contamination in and around the fictional city of Kish. Reports from rural areas indicate an unidentified illness affecting livestock, and, from the city, an increase in patients reporting common symptoms. Analysts identify contamination of drinking water as a possible cause. An intelligence requirement is issued to determine whether this is accidental, or related to other suspicious activities such as a local pumping station explosion. The event of the explosion requires immediate response to both causes and the potential threats to the population. The analysis team requires support both to understand the evolving situation and to liaise with local authorities to gather further information about the spread of the illness in the region. Fig. 4 provides an overview of the scenario and a timeline of events, reporting on the upper part of the timeline information received, and on the lower part key parts of the analysis.

3.3. Intelligence analysis requirements and priorities

In this work, we aim to introduce AI techniques to facilitate analysis in a principled way that aligns with analysts' methods employed in everyday activities. In Section 2, we explored key issues raised in the literature which motivate our objectives, while the focus group complements these objectives by providing further insight in the analysis process and by eliciting priorities. Table 1 provides a summary of requirements and challenges raised in the literature and remarked by the focus group, alongside an indication of how the objectives – namely the support provided for the sensemaking (OJ1), the crowdsourcing (OJ2), and the provenance (OJ3) reasoning process – align with the challenges identified. The table further organises the challenges following the coding scheme of the focus group. In this research, we mostly focus on the hypotheses formation leaving some of the requirements specifically related to general collection of information and reporting for future research. In the next section, we will discuss how our AI approaches have been designed to address these challenges.

4. Automated support for intelligence analysis

Grounded in the requirements and priorities elicited from the focus group sessions discussed in Section 3, we developed a tool, Collaborative Intelligence Spaces (CISpaces) (Toniolo et al., 2015), which uses multiple AI techniques to support intelligence analysts in:

- *evidence-based sensemaking* by employing adapted argument schemes (Walton et al., 2008) to guide critical review of evidence, and a tailored model of argumentation-based reasoning (Prakken, 2010) to identify plausible hypotheses (OJ1);
- *gathering crowd-sourced evidence* by interpreting responses to structured requests for information from groups of collectors (Kamar et al., 2012) and feeding the results back into the analysis as arguments (OJ2); and
- *assessing the provenance of information* by inspecting the provenance of information to identify critical meta-data that may inform the credibility of hypotheses (Toniolo et al., 2014) once again by interpreting them as arguments (OJ3).

Support is provided to analysts via an interface with two core components: the *InfoBox*, where collected information relevant to a task are streamed from external sources, typically from intelligence reports; and an individual *WorkBox*, the analytical space used in the construction of hypotheses. Each component in the analysis has a provenance chain attached. Different forms of collaboration are supported in CISpaces. Portions of the *WorkBox* can be shared between analysts via drag-and-drop, enhancing collaboration. An analyst may also canvas groups of contributors via the *ReqBox*, by creating forms for collecting structured information via crowdsourcing.

Fig. 5 provides a screenshot – edited to enhance its readability – of the system during use within our scenario. On the left, the *InfoBox* collects relevant information: for the purpose of this paper we assume there is an information stream connected to it providing a stream of information to the analysts. Then analysts can move selected pieces of information from the InfoBox into the *WorkBox* and use these to build their analysis. Recalling our running scenario from Section 3.2, among many theories, the contamination in Kish could potentially be explained by bacteria in the water supply system, and a local Non-Governmental Organisation (NGO) ran some tests for these. Depending of the type of

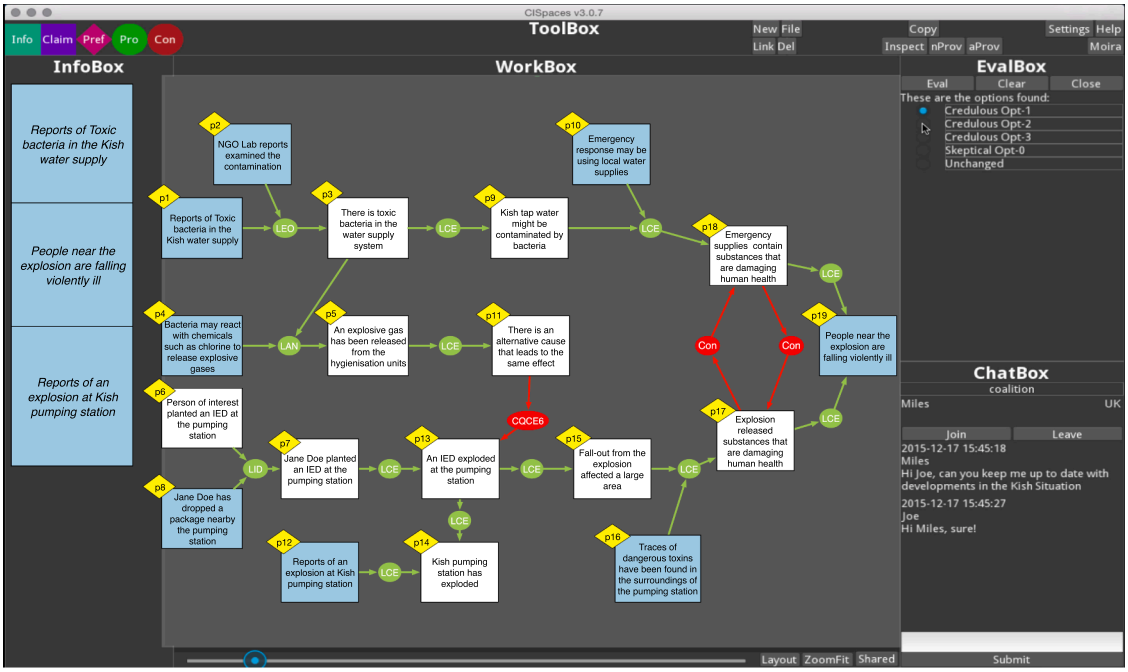


Fig. 5. A CISpaces view of evidence-based sensemaking task of events happening in Kish.

bacteria, it is known that there might be a reaction with chlorine, used in the water system, which could release explosive gases. This, in turn, could potentially cause an explosion. However, the opposite could also be true: an Improvised Explosive Device (IED) that uses highly-toxic chemicals could have been planted next to the pumping station, leading to poisoning of the local water supplies. In the next subsections, we describe the AI-based support provided to analysts for this ongoing evidence-based sensemaking activity. In particular, in Section 4.1 we discuss how evidence is structured and analysed using argumentation approaches following OJ1. Section 4.2 focuses on the crowdsourcing data collection and analysis following OJ2. Provenance assessment (OJ3) is demonstrated in Section 4.3. In the last part of this section, we provide an overview of how CISpaces was developed (Section 4.4).

4.1. Evidence-based sensemaking

In Sections 2 and 3, we highlighted several characteristics of sensemaking: hypotheses are formed by drawing inferences over information, guided by patterns such as from link analysis, where each hypothesis can be considered as a story, underpinned by a sequence of events (the *historian* approach). There are multiple alternative hypotheses considered during analysis, arising from conflicting information or from alternate events that explain this information. Furthermore, analysts are familiar with informal argumentation, as the principal method to evaluate conclusions on the basis of arguments supported by evidence. CISpaces leverages these characteristics to provides automated reasoning support for analysts in structuring and elaborating individual and collaborative analysis. CISpaces makes extensive use of computational argumentation, in particular by introducing patterns to help draw inferences, and by providing a conceptual and computational framework to guide the identification of justified hypotheses.

A graphical representation of inferences enables analysts to form conclusions on the evidence acquired through the *WorkBox* in the CISpaces interface, and is based upon other argument mapping tools (Reed and Rowe, 2004; van Gelder, 2007). An inference rule is a set of propositions (p_i) divided into one or more *premises* that are linked with a *conclusion*. Premises and Conclusions are represented as either white *information* or light-blue *claim* nodes in Fig. 5 (respectively in green and purple in the system, modified here for readability purposes). For instance, to explain the *information* of illness within the population near the explosion in Kish (p_{19} on the right-hand side of Fig. 5), our analyst might identify a premise (*claim*) that this is caused by the supplies used by the emergency services (p_{18}) forming an inference rule between p_{18} and p_{19} linked by a green round node called a *Pro-link*. We will write Pro-links as $p_{18} \xrightarrow{p} p_{19}$ within this paper. Note that Pro-links can be annotated within CISpaces to provide additional meta-information about the type of inference between nodes.

Propositions can also be in conflict with other components on the basis of an asymmetric *contrariness* relation (Prakken, 2010). Conflicting propositions are linked through red round nodes referred to as *Con-links*. In the example in Fig. 5, the analyst might question whether the explosion released gas that is causing the illness (p_{17}). Specifically the Con-links capture contradicting relationships between pieces of information: p_{17} is the contrary of p_{18}, and p_{18} is the contrary of p_{17}, hence p_{17} and p_{18} are said to be *contradictory* as they represent two alternative, mutually exclusive, interpretations of reality (conclusion p_{19}). We will write Con-links as, for example, $p_{17} \xrightarrow{c} p_{18}$.

4.1.1. Patterns of inference for intelligence analysis

The underlying structure used to support analysts in drawing inferences is based on argumentation schemes — reasoning patterns that commonly occur in human reasoning and dialogue (Walton et al., 2008). They represent templates for making presumptive inferences formed by premises supporting a conclusion, and by critical questions (CQs) that can be put forward against the applicability of the inference. A commonly used example is the *argument from expert opinion*, used to describe an assertion warranted by expertise:

- Source E is an expert in domain S containing proposition A
- E asserts that proposition A is true

⇒ Therefore, A may plausibly be true.

For instance, p_3="There is toxic bacteria in the water supply system" might be the result of analysts' speculation (hence being a claim) and subject of further investigation. Such an investigation can be summarised by an argument from expert opinion reporting a test that was conducted by an NGO (Non-Governmental Organisation) laboratory

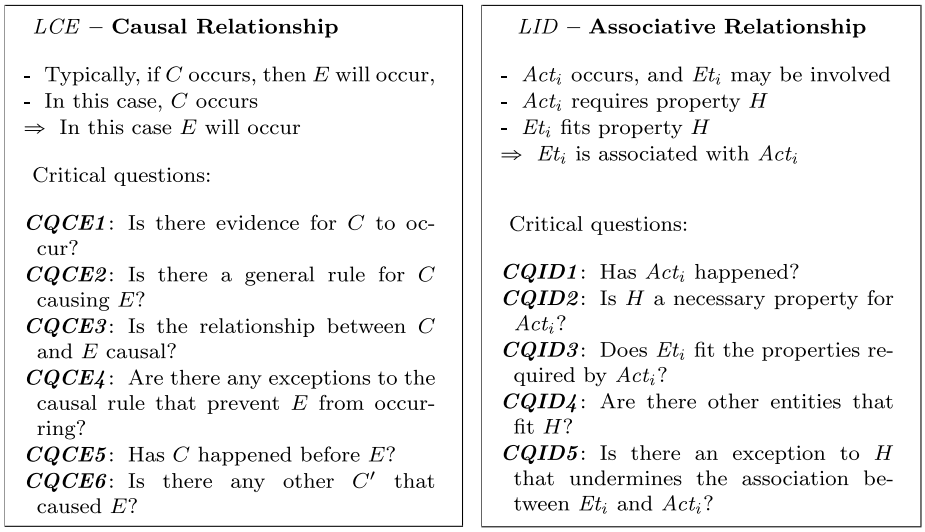


Fig. 6. Argumentation schemes and critical questions for intelligence analysis.

(p_2), which points towards expertise, with a second premise relating to the assertion reporting of Toxic bacteria in the Kish water supply (p_1). In Fig. 5, the link $p_1, p_2 \underset{\rightarrow}{p} p_3$ can be labelled with an argumentation scheme, in this case "LEO," standing for "Link from Expert Opinion."

There are, however, no explicit statements about this NGO laboratory having expertise for testing bacteria in water samples, or regarding other issues such as their reliability. To this end, *critical questions* can then be asked to strengthen (or weaken) arguments instantiated from argumentation schemes. Some of the critical questions relevant to an argument from expert opinion are (Walton et al., 2008):

CQEO1: How credible is E as an expert source?
CQEO2: Is E an expert in the field that A is in?
CQEO3: Is A consistent with the testimony of other experts?

Uniquely in CISpaces, to our knowledge, is how the critical questions are used to drive further analysis. When the analyst selects a question that must be answered for the conclusion to be acceptable, the system generates a negative answer in a new node connected via a Con-link; e.g., for CQEO1 "Source E is not an expert source." This asymmetric conflict prompts the analyst to challenge assumptions that may lead to bias, and requires them to find a reason for source E to be considered an expert, as otherwise the conclusion that there is bacteria in the water supply p_3 would remain unsupported in the evaluation of hypotheses, as discussed later in this section.

To provide analysts with a coherent model, we worked closely with experts to identify the most common argumentation schemes used in intelligence analysis. Analysts are mostly concerned about: **Activities** (*Act*) including *actions* performed by actors, and *events* happening in the world; **Entities** (*Et*) including individuals or groups, and objects such as resources; and **Facts** (*Ft*) including statements about the state of the world regarding entities and activities. There are several critical relations among these elements: *causal* relations representing the distribution of activities, their correlation and (possible) causality; and relations that connect entities and activities through temporal, geographic or thematic *associations*. Intelligence elements then act as premises for inferences, and conclusions are tentatively drawn by discovering relations among them. In line with the historian approach (see Fig. 1), analysts then use these relations to reconstruct a narrative that explains events forming alternative hypotheses. According to the type of relation (causal or associative) we can now instantiate two main types of schemes for the sensemaking process (cf. Fig. 6).

An **argument scheme from cause to effect** may be used to provide an explanation for some set of observations on the basis of activities and events that shows how the situation has evolved. This is referred to as an inference link *LCE*, and considers a cause *C* (referring to some fact Ft_i or activity Act_i), its effect *E* (also referring to some fact or activity), and a causal rule that links *C* to *E*. In Fig. 6 we present *LCE*, which has been adapted from Walton et al. (2008). In our previous example, as illustrated within Fig. 5, the two explanations of usage of defective emergency supplies (p_{18}) or gas releases (p_{17}) may be events that cause the spreading of the illness among the population (p_{19}). The links $p_{18} \xrightarrow{p} p_{19}$, $p_{17} \xrightarrow{p} p_{19}$ can then be considered instances of a causal argument scheme.

Instances of the causal argumentation scheme form a chain of events that constitute the backbone of the hypothesis. These can further be reviewed through critical questions, challenging the instantiation of the causal argumentation scheme (CQCE4); the order of events (CQCE5); and evidence for the premises (CQCE1, CQCE2, CQCE3). Critical question CQCE6 has a different purpose. In CISpaces, analysts are required to represent a cause as a Pro-link to an effect, but analysts may have evidence for the effect and infer a plausible cause using abductive reasoning. In this case, alternative causes must be considered. CQCE6 is used to consider these alternatives by interpreting this question as a rebuttal for *C*. CQCE6 then results in alternative incoming nodes to the Pro-link representing a contradictory relation between causes.

An **argument for identifying an agent from past actions** (Walton et al. 2008) (*LID*, Fig. 6) encodes the sensemaking process that shifts from understanding what happened to understanding what entities were involved and their association with the activity. In this scheme, properties, *H*, are facts *Ft* of type "Et_i is affected by Act_i/Et_j" or "Et_i is in the location Et_j of Act_i". An instantiation of this scheme can, for example, be used to assert that the observed activity, Act_i, "An unidentified person (aka Jane Doe) of interest planted an improvised explosive device (IED) at the pumping station (p_6)" and certain properties (e.g., p_8) can be used to draw the conclusion "Jane Doe planted the IED"(p_7) through p_6, $p_8 \xrightarrow{p} p_7$.

Associated critical questions CQID1, CQID2 and CQID3 can be used to review the inference by challenging respectively: that the act (planting the IED) has occurred; that entity (Jane Doe) has some property (seen at the location); and that the act requires that property. CQID4 identifies alternative conclusions, while CQID5 challenges the instantiation of the scheme.

The links between these two major schemes for causal and associative relations are primarily forged through questions CQCE1 and CQID1. A response to question CQCE1, for example, may claim that some entity, Et_i, was associated with the cause. In this way, *LID* may be used to

answer a challenge to an instance of *LCE*. Similarly, an instance of *LCE* may answer question CQID1, which is concerned with whether some activity happened, linking association back to causality. In addition, different schemes may be used by the analyst to respond to the various critical questions (Walton et al., 2008). Examples include: *arguments from the group* where properties of a member are applied to an organisation for providing evidence to *LID*; an *argument from analogy* reporting a case with similarities to *LCE*; or an *argument from sign* to explain that an event is likely to happen if its indicator is verified, following common indicators such as those presented in training manuals (US Army, 2020).

4.1.2. Hypotheses identification

Following our running example, suppose that analysts have prepared the analysis shown in Fig. 5 to identify what the coherent explanations for this situation are, in order to evaluate their hypotheses.

CISpaces provides automated support to identify what claims and pieces of evidence can together form a plausible hypothesis, and what other alternatives exists that are also plausible, by employing computational models of argumentation. In such models, a fundamental concept is that of an inference rule, where a statement (antecedent) becomes a (prima facie) reason to believe another statement (consequent). For instance, "Reports of Toxic bacteria in the Kish water supply" (antecedent) can be seen as a *prima facie* reason to believe that "There are toxic bacteria in the water supply system" (consequent). In this research, we only make use of a small set of concepts derived from formal argumentation, specifically borrowing from the ASPIC literature (Modgil and Prakken, 2014) (see Appendix A for further details). For example, scholars in this area distinguish between strict and defeasible rules in their approach to formal argumentation and preferences to establish defeats between arguments (Modgil and Prakken, 2014). Analysts by training are familiar with informal argumentation concepts (see Section 3), including premises and conclusions of an argument, supporting and conflicting arguments. In this research, in order to limit the training burden for analysts, we chose concepts which we could align with these informal concepts but limited to a small set which we deemed necessary for representing and evaluating an analysis. We, therefore, will not make use of strict rules or preferences in this work, and we will not discuss those further.

Rules provide the building blocks for the notion of argument, that is iterative in the chaining of rules. Statements that are tentatively assumed to hold provide the base case for such an iteration, and thus they are defined as arguments having the statement itself both as a singular premise and as a conclusion, where such premises and conclusion are two attributes of an argument. The premises of arguments constructed using this base case also take the name of ordinary premises in our approach. As an iterative step, an argument requires the existence of a rule whose antecedents are the conclusions of other arguments (sub-arguments), and as a consequent a statement that forms the conclusion of this new argument. The premises of such a (compound) argument are the union of all the premises of its sub-arguments. A statement is the contrary of another one when they cannot be both accepted, albeit they can both be rejected. Borrowing from the literature, a flexible way for using such a notion of contrariness is by allowing for a statement to be the contrary of another one, while not explicitly requiring the opposite. Two statements which are contrary to each other are said to be *contradictory* as mentioned above.

The notion of contrariness between statements leads to the concept of defeat between arguments: an argument defeats another argument if the former rebuts or undermines the latter. When the conclusion of an argument contradicts the conclusion of another argument, it is the case that the first rebuts the second, as well as any other compound argument that has the second argument as sub-argument. If, instead, the conclusion of an argument contradicts one of the premises of another one, then the former is said to undermine the latter. Exceptions to the application of a rule of an argument scheme are also considered contrary undermining arguments to implicit premises in this work.

The graphical map of inferences constructed by the analyst, cf. Fig. 5, is transformed into the corresponding premises, contrariness relationships and inference rules as follows:

- Premises are considered those propositions that are not conclusions of inferences (incoming Pro-link edges) and constitute part of the knowledge base.
- A contrary relationship is added if a Con-link is drawn between two propositions. In addition, critical questions that point towards alternative conclusions are mapped as contradictory relationships.
- Pro-links map to inference rules.

In our previous example, the LCE Pro-link $p_1, p_2 \xrightarrow{p} p_3$ represents an inference rule $r : p1, p2 \Rightarrow p3$ and gives rise to three arguments: Arg_1 with p_1 both as premise and as conclusion; Arg_2 with p_2 both as premise and as conclusion; and Arg_3 with $\{p_1, p_2\}$ as premises and p_3 as conclusion following the rule r.

On the basis of the asymmetric *contrariness* relation, an argument can attack another one: if the conclusion of an argument Arg_X is the contrary of one of the premises of argument Arg_Y, then Arg_X *undermines* Arg_Y; if the conclusion of Arg_X and the conclusion of Arg_Y are contradictory then the two arguments *rebut* each other.

Arguments and attacks form a Dung argumentation framework (Dung, 1995) from which sets of acceptable arguments surviving the attack together (extensions) can be computed according to a *semantics*. In this work we consider the preferred semantics. This semantics selects a maximal set of arguments: maximal with respect to set inclusion extensions that are conflict free (i.e., no arguments in any extension attack each other), and admissible (i.e., each argument in the extension is defended against the attacks it receives).

For the first time in the intelligence analysis and argumentation literature – to our knowledge – we associate each extension to an intelligence analysis hypothesis. The process of intelligence analysis is driven by the identification of hypotheses and a discussion of any potential alternative that may explain the information received about a situation. This is a key concept in the intelligence literature (see Section 2) and has strongly emerged during our focus group (Section 3). Analytic methods such as the red-team versus blue-team or the Analysis of Competing Hypotheses show that this process is embedded in analysts' work (e.g. Heuer, 1999; Klein et al., 2006). The preferred semantics is a multiple-status semantics which provides a set of alternative labellings and, therefore, is well suited to represent alternative explanations for a situation. A set of acceptable arguments identified in the extension permits the extraction of acceptable propositions about events, entities and activities that together are plausible. This is formed by the conclusions of the arguments. We refer to this set as a single, coherent hypothesis or a plausible hypothesis in short. This set is then presented to the analyst using colour coded symbols in the interface. For each hypothesis, green (V) is used to indicate a supported conclusion of an argument belonging to the extension (IN), and red (X) is used to indicate an unsupported conclusion of an argument that is attacked by some argument of the extension (OUT).[2] The formal correspondence between argumentation semantics extensions and identification of alternative hypotheses is discussed in Appendix A.

To recall our previous example, we can extract two hypotheses as shown in Fig. 7, corresponding to the conclusions of arguments accepted by one of three preferred extensions. In addition, unsupported statements for a hypothesis are also highlighted, which in turn might show analysts the consequences of not responding to a critical question (as per our previous example in Section 4.1.1). Assuming that there is no evidence that the NGO is an expert in water contamination (p_2) as

[2] There is also a third case of undecided conclusions. Further details can be found in Appendix A.

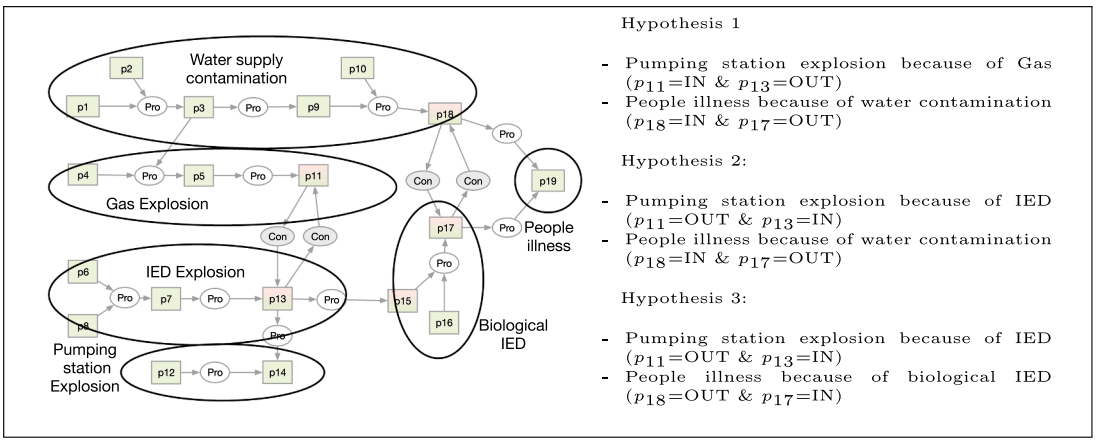


Fig. 7. CISpaces analysis and evaluation: Three alternative hypotheses.

suggested by CQEO1, this invalidates the first two hypotheses.

To conclude our section on sensemaking, our novel approach includes a tailored set of argumentation schemes for intelligence analysis and an automatic use of critical questions to effectively support analysts in reducing their cognitive biases, leveraging in full the computational argumentation paradigm by creating directed attacks from unanswered critical questions. In addition, with our approach analysts can automatically identify alternative hypotheses, thanks to the correspondence between an extension and a plausible hypothesis. An analyst can more readily observe the effects of adding further information or advancing critiques to parts of the analysis on the set of plausible hypotheses, demonstrating visually the availability of a new hypothesis or the rejection of an unsubstantiated one.

4.2. Crowd-sourced evidence

Continuing with our running example, the analyst might require additional location-sensitive information to ascertain whether there is evidence that Kish tap water is contaminated, p_9. As highlighted by analysts in our interviews (see Section 3), timely information is crucial in this context in particular when the contamination may explain why people are falling ill, as this would allow for more rapid intervention. To do so, analysts could initiate a request for information distributed to the local population to collect evidence about the status of the tap water in Kish. Crowdsourcing uses human computation to sense information and discover truth in a timely, large-scale and cost-efficient manner (Brabham, 2008; Kamar et al., 2012; Whitehill et al., 2009), and it is particularly effective in event detection (Ouyang et al., 2016b). How to interpret and integrate such crowdsourced evidence into an analysis is, however, an open issue.

In CISpaces, we proposed an online method to analyse results of the reports and instantiate them within a novel argumentation scheme which integrates these results into the analysis. More information about the formalism can be found in Appendix B.

4.2.1. Task initialisation

In CISpaces, a crowdsourced query task is initiated by asking specific CQs; e.g. a claim p_t may be challenged by the analyst via the question "Is there evidence for p_t?" In our example, this is initiated as "Is there evidence that the tap water in Kish is contaminated"?. We assume that after some time, people in Kish respond to this request by reporting the colour and temperature of their tap water.

In relation to a specific task, we can automatically introduce novel data and inference links in the graph of arguments from a number of questions Q for the crowd, together with associated information enabling data collection and aggregation of results. An example could include the two questions $Q = \{q_0, q_1\}$:

- q_0: "What is the temperature of your cold water?", of numerical type. If the temperature reported is $< 20°C$ the results will provide evidence against the claim p_t that the tap water in Kish is contaminated, otherwise the response provides evidence for the claim.
- q_1: "What colour is your tap water?", of categorical type with m possible categories. If the water is *Clear* or *White* this would be evidence against the claim, and *Brown* and *Yellow* are considered as evidence for the claim.

The task terminates when it reaches a deadline or some pre-specified number of reports are acquired.

4.2.2. Analysis of results

The results are aggregated in different ways depending on the type of data. For categorical data we are interested in knowing the probability of the categories of a multi-valued answer to question q_k. Using a Dirichlet prior for this multinomial distribution, the posterior is thus a Dirichlet distribution (Jøsang and Haller, 2007) that combines prior beliefs and collected reports for question q_k from which we obtain a vector of expected values $\overline{c}_k$ for the m categories of question q_k. The prior used in the simplest case is a uniform distribution over the answers, but a more sophisticated approach would consider crowd features such as reliability and location by manipulating the prior (e.g., Etuk et al., 2013; Ouyang et al., 2016b). For numerical data, we consider a weighted mean μ_k of the collected reports for q_k where in the simplest case weights are all assumed to be 1, although these may vary according to features of the reports as for the prior probability.

A novel aspect of our approach occurs after aggregating the results for each question q_k, when CISpaces uses the task definition to *automatically* build a partial argument map that is integrated within the overall analysis. The *argument from generally accepted opinion* (Walton et al., 2008), *LCS*, represents the defeasible inference that a statement is plausible if a significant majority in a group accepts it.

- Given that the crowd was asked q_k and
- Answer A is generally accepted as true
⇒ Therefore, A may plausibly be true

Critical questions focus on whether the crowd is believable, or corroborating evidence is needed to accept the conclusions:

CQCS1: Is the claim A supported by evidence?
CQCS2: Is the group in a position to know about q_k?
CQCS3: Is the claim consistent with others' claims?
CQCS4: Does the group present characteristics appropriate for answering q_k?

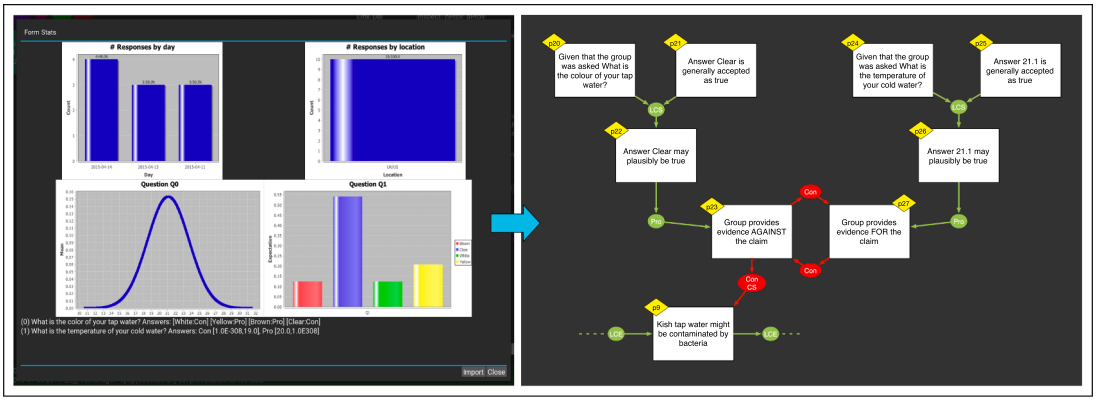


Fig. 8. CISpaces Crowdsourcing and Analysis Example for p_9: "Kish tap water might be contaminated by bacteria," cf. Fig. 5.

The system constructs a *LCS* argument for each question q_k where the answer A corresponds to the mean μ_k for numerical questions, or for categorical ones the category with maximal expected value $\epsilon_j \in \bar{\epsilon}_k$. Each conclusion either provides evidence for or against the main claim p_t.

In our running example, assume we have collected 10 reports for q_0, q_1 such that:

- q_0: $\{21, 22, 25, 24, 18, 17, 22, 20, 23, 19\}$ with $\mu_0 = 21.1$
- q_1: $\{Clear : 6, Brown : 1, Yellow : 2, White : 1\}$ with corresponding expectation $\boldsymbol{\epsilon}_1 = (0.542, 0.125, 0.208, 0.125)$

Fig. 8 illustrates the CISpaces interface for data collection, consisting of the data itself and information such as the location and time of responses. CISpaces allows for inspection of the data before importing, and the imported section of the results (from which arguments are derived) is shown on the right-hand side of the diagram. The *LCS* argument p_{24}, $p_{25} \xrightarrow{p} p_{26}$ on the right-hand side states that the temperature of the tap water is 21.1°, reporting the result of q_0. The *LCS* argument for q_1, p_{20}, $p_{21} \xrightarrow{p} p_{22}$ reports that the colour of the tap water is Clear.

CISpaces also uniquely aggregates the results of the various questions once again in a purely argumentative manner stating that the group provides evidence either *for* or *against* the claim, respectively p_{27} and p_{23} in Fig. 8. The *against* claim is then linked to the claim that originated our crowdsourcing request, (p_9 in the figure), via a Con-link, $p_{23} \xrightarrow{c} p_9$, in line with our definition of critical questions. If one or more conclusions of *LCS* exist providing evidence for the claim (e.g. p_{27}), individual Pro-links are used to connect these conclusions to the aggregated for statement *for*, e.g. $p_{26} \xrightarrow{p} p_{27}$; a single Con-link attacks the opposite *against* claim, $p_{27} \xrightarrow{c} p_{23}$. Similar links are added if one or more conclusions exist providing evidence against the claim. Hence, if all evidence is for a claim, the claim will be accepted (assuming no other arguments exist against the claim), otherwise the claim will not be accepted. In our example, we show however that given that there is not decisive evidence, we currently obtain an inconclusive result, and the two hypotheses are still valid.

To conclude, gathering additional information is necessary to avoid the rejection of hypotheses on the basis of insufficient evidence (Heuer, 1999). Our novel approach to crowdsourcing, evidence interpretation and *automated* integration of the outcome(s) into an analysis using specifically designed argumentation schemes and procedures, provides an effective method to integrate this form of human intelligence into the sensemaking process.

4.3. Provenance

As previously described, each component in the analysis, whether input information or the analysis itself, has a provenance chain attached: data representing the phases of manipulation of that component from its primary sources. In our focus group (Section 3) analysts have highlighted that the origins of information (including information from the crowd), and how and by whom this information is interpreted during analysis are important factors to establish the credibility of hypotheses. Provenance can be used to annotate how, where, when and by whom some information was produced (Moreau and Missier, 2013). Understanding the provenance of information more broadly, however, is fundamental to assessing its credibility. Information may of course come from sources of varying veracity, but it may also have been manipulated or combined with other information before reaching the analyst, and the relative timeliness of information is important for many problems. The interpretation of information and understanding how information and hypotheses are linked must take into consideration all aspects of information provenance. Further, when we consider that analysis of more complex, real-world situations is typically team-based and may involve hand-over between teams or involve multiple agencies, it is important to understand an individual's contributions and what data was used to reach conclusions (Wu et al., 2013). Inspecting long provenance chains to identify relevant provenance information to assess credibility, however, remains cognitive demanding. Through the use of argumentation schemes, here we extract relevant provenance data to be introduced in the analysis following our previous work (Toniolo et al., 2014).

4.3.1. Recording provenance

Provenance is recorded in CISpaces using the W3C standard PROV Data Model (Moreau and Missier, 2013). PROV-DM expresses provenance in terms of p-entities (A_{pv}), p-activities (P_{pv}), and p-agents (Ag_{pv}) that have caused an entity to be, and defines different relationships between these elements. Note that in PROV-DM these elements are referred to as entities, activities and agents; we use the *p-* prefix to refer to provenance elements explicitly.

The left part of Fig. 9 illustrates a provenance graph used and manipulated by CISpaces for the information node p_{10}: "Emergency response may be using local water supplies," cf. Fig. 5. Reading the graph from right to left, we can see orange round nodes that are directly associated to the information nodes in CISpaces and hence that Joe, an analyst (p-agent), has imported a piece of information within CISpaces. We can also walk back in time, and thus see that this information has been delivered to the "InfoBox" by a "NGO_Officer" who has communicated key data extracted from a "Crisis_Report", all the way back to the primary sources (i.e., those that first reported or created the information) "Field_Observations" of the area of interest and local "Water_Samples".

Therefore, the provenance chain of a node p_j is represented as a directed acyclic graph $G_P(p_j)$ of relationships between A_{pv}, P_{pv}, and Ag_{pv}. $G_P(p_j)$ is a joint path from the node containing p_j to its primary sources; i.

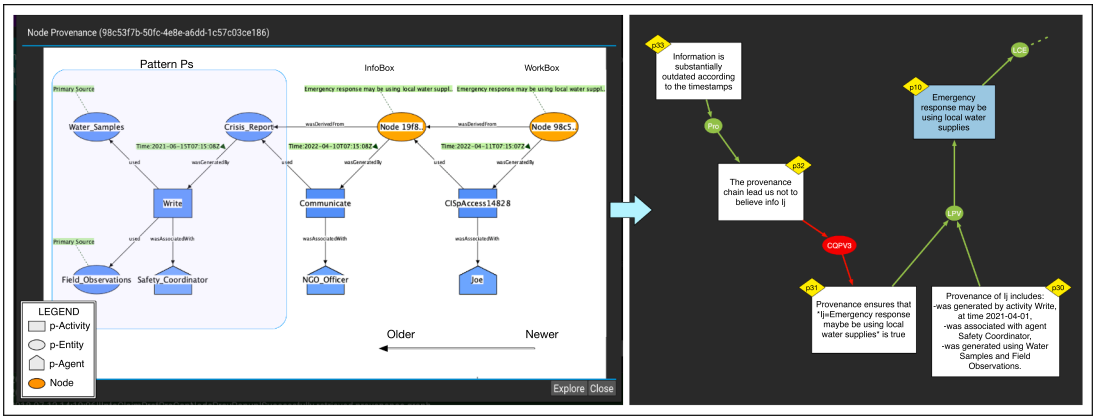


Fig. 9. Provenance chain associated to p_{10}: "Emergency response may be using local water supplies," cf. Fig. 5.

e., sources that first produced the information. More details on the formal treatment of provenance graphs is provided in Appendix C.

4.3.2. *Reasoning about provenance*

A provenance chain $G_P(p_j)$ can be queried as a graph pattern P_m which is a structured graph with nodes being variables on the p-elements. Following our previous work (Toniolo et al., 2014), we consider three commonly used patterns for intelligence analysis: P_g indicating how a p-entity was generated; P_s used to identify the primary sources used in the generation of a p-entity; and P_t which connects a piece of information with its intelligence requirement. For example, consider in Fig. 9 the provenance chain involved in the acquisition of p_{10}="Emergency response may be using local water supplies," cf. Fig. 5, referred to as p-entity "Node 98c5...". A pattern P_g shows that the piece of information p_{10} contained in the InfoBox node ("Node 19f8...") has been extracted from the Crisis Report and communicated to analyst Joe by the NGO Officer. P_s highlighted in blue Fig. 9 shows that at the source of p_{10} the Crisis Report was created on the basis of field observations and water samples by the safety coordinator.

These patterns allow us to check the presence of relevant provenance information that may warrant the credibility of p_j and the information about activities (Act_i), entities (Et_i), or facts (Ft_i) p_j is concerned with. The patterns can be integrated into the analysis by applying the *argument scheme for provenance* (*LPV*) we first introduced in a previous paper (Toniolo et al., 2014):

- Given information p_j
- The provenance chain $G_P(p_j)$ of p_j includes pattern P_m of p-entities A_{pv}, p-activities P_{pv}, p-agents Ag_{pv} involved in producing p_j
- P_m is a reason to believe that information p_j is true

⇒ Therefore, p_j may plausibly be true

Critical questions for this scheme are:

CQPV1: Is p_j consistent with other information?
CQPV2:Is p_j supported by evidence?
CQPV3: Does $G_P(p_j)$ contain p-elements that lead us not to believe p_j?
CQPV4: Is there any other p-element that should have been included in $G_P(p_j)$ to infer that p_j is true?

A question "Can it be shown that the information is verifiable?" (e.g. CQID1, CQCE1, cf. Fig. 6) shifts the reasoning process to provenance analysis. Questions CQPV1 and CQPV2 shift back to sensemaking by requiring further evidence for Act_i, Et_i, or Ft_i to be supported.

To integrate the provenance elements into the analysis, CISpaces extracts and shows available patterns P_m to the analyst. The analyst can choose a pattern deemed important for a specific part of the analysis in the Workbox. As per crowdsourced evidence, CISpaces provides an argumentative method to import this in the analysis via a *LPV* argument scheme. The conclusion already exists in the WorkBox since p_j concerns an Info or a Claim node, and the premises of *LPV* form a Pro-Link to provide additional evidence for p_j. This is the case in Fig. 9, where the pattern P_s allows us to instantiate an argument from provenance whose premises are p_{30} and p_{31}, representing respectively P_s and the warrant which justifies the credibility of p_j on the basis of P_s. Claims p_{30} and p_{31} then provide additional information on why we should believe that "Emergency response may be using local water supplies " through a link $p_{30}, p_{31} \xrightarrow{p} p_{10}$.

Provenance data supporting a claim might be helpful in further stages of the analysis, and might demonstrate to other analysts that this information was considered important. On the other hand, a pattern P_m may be a reason for believing that p_j is not credible, based upon reasons expressed by CQPV3 or CQPV4. As discussed in the more general argumentative process (Section 4.1.1), a negative answer to one of the critical questions triggers a new Con-link being formed, representing an attack on the premises of *LPV*, and therefore indicating that p_j would not be supported. In our example, looking closely at the provenance graph of Fig. 9, we notice that from the timestamps attributes of wasGeneratedBy for the Crisis_Report, it appears that this report is one year older, and therefore likely to be not relevant to the current crisis, raising critical question CQPV3. Claim p_{33} provides support to the critical questions CQPV3 (p_{32}), which, in turn, undermines p_{31} ($p_{32} \xrightarrow{c} p_{31}$) and consequently p_{10}. Extended patterns looking at the timeliness of information could also be considered to assess the credibility of a given piece of information automatically as discussed in our previous work (Toniolo et al., 2014).

With the process suggested above, CISpaces supports analysts in extracting relevant provenance information to be consumed in the process of reviewing the credibility of evidence and hypotheses. Indeed, looking at the provenance of the information p_{10}, namely that "Emergency response may be using local water supplies" and at the arguments that we can extract from it (shown in the right side of Fig. 9), we can conclude that p_{10} should not be acceptable as it is based on a substantially flawed process. This has far-reaching effects: looking at Fig. 7, accepting p_{10} is instrumental to accepting p_{18}, which in turn is necessary for one of the three hypotheses explaining the situation. By knowing that there is a reason to believe that p_{10} (and thus p_{18}) is not the case as the piece of information is not timely, the hypotheses explaining the situation now become:

- Hypothesis 1:
  - Pumping station explosion because of Gas (p_{11}=IN & p_{13}=OUT)

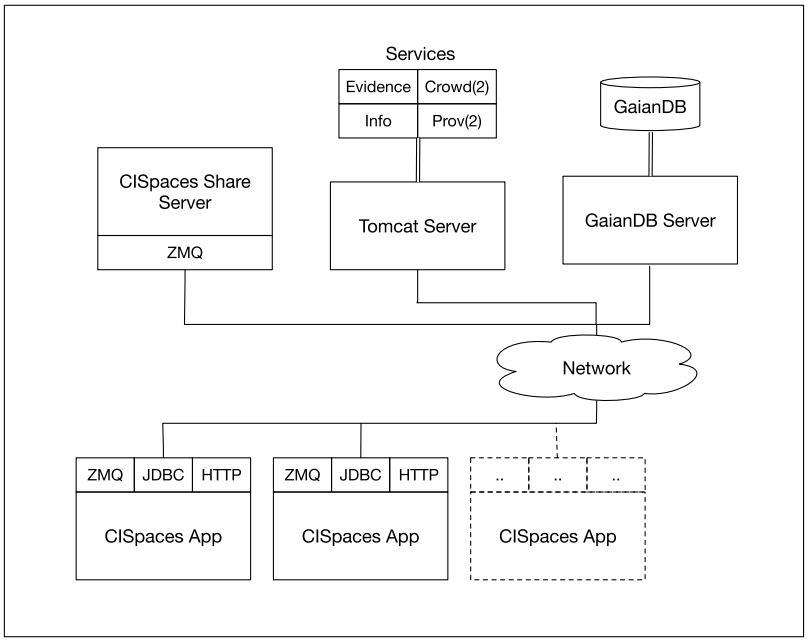


Fig. 10. CISpaces system architecture.

- Hypothesis 2:
  - Pumping station explosion because of IED (p_{11}=OUT & p_{13}=IN)
  - People illness because of biological IED (p_{18}=OUT & p_{17}=IN)

Note that in Hypothesis 1 we no longer have an explanation for people illness (p_{18}=OUT & p_{17}=OUT) which would then require further investigation if Hyp. 1 is to be taken forward.

To conclude, we created a new argumentation schemes for automatically incorporating relevant patterns linked to provenance of information. In this way, we effectively support analysts in establishing the credibility of hypotheses, as demonstrated using our running example.

4.4. System implementation

CISpaces is a web-based system. CISpaces interface is developed in Python 2.7 and deployed through the Kivy framework (The Kivy Community, 2011). CISpaces clients communicate with each other and may share analyses using a pub-sub architecture provided by ZeroMQ (The ZeroMQ Community, 2007) which is backed by the CISpaces ZMQ share messaging server. A database server is used for persistence: in our case the Gaian Database (Vyvyan et al., 2015) a dynamically distributed federated database, which allowed us to connect with other services developed within the broader scope of this project (e.g., Toniolo et al., 2016, see Section 6.2). The AI support is provided by a series of RESTful services developed in Java (Oracle, 1996))and deployed via an Apache Tomcat server (The Apache Software Foundation, 2002). Exchanges are supported via structured json data (Ecma International, 2017). These services include:

1. the *evidence reasoning service* responsible for the core evidence-based sensemaking tasks (see Section 4.1). This includes, for example, the identification of hypotheses where a call to the service is made every time the analyst intends to evaluate the current analysis and the results are displayed in the interface. The current view of the argumentation framework is posted to the service structured according to the Argument Interchange Format (AIF, Cerutti et al., 2018c).
2. two *crowdsourcing services* to handle the Crowd-sourced evidence (see Section 4.2): one for the collection of crowdsourced data, one for the analysis of the results.
3. two *provenance services*: one for recording, storing and retrieving provenance data, one for the analysis and visualisation of provenance records as discussed in Section 4.3. All provenance data is stored in the Gaian Database and is handled and queried via the Apache Jena framework (The Apache Software Foundation, 2010). The provenance is RDF-compliant following the PROV-O ontology (PROV Working Group, 2013).
4. a simulated *information retrieval service* demonstrating how a stream of information may flow into the system.

In Fig. 10 we depict the CISpaces architecture. Note that the evidence reasoning service is currently openly available as part of the newer open source CISpaces.org (Cerutti et al., 2018b), see Section 6.2.

5. Expert evaluation of CISpaces

In this section, we discuss how the AI techniques we developed and implemented in CISpaces may advance performance in intelligence analysis thanks to an evaluation of CISpaces with subject-matter experts.

Our key question in this evaluation is "Would CISpaces be adopted by professional analysts?". We intend to study: a) whether analysts consider CISpaces useful in supporting the analysis process, and b) what characteristics of CISpaces would influence the adoption of CISpaces. In Section 3 among our objectives, we discussed the aim of developing a system that supports the processes underpinning analysis by integrating and aligning with analysts' methods in order for the system to be acceptable to and adopted by practitioners. In this evaluation, we demonstrate that indeed analysts believe that CISpaces is valuable in this respect.

In the following subsections, we provide information on the methodology, hypotheses and experimental settings (Section 5.1). The quantitative results are reported in Section 5.2. We follow with a discussion of these results complemented by a qualitative analysis in Section 5.3.

5.1. Questionnaire and methodology

We run our empirical study using a questionnaire tool to investigate the analysts' response to a potential introduction of CISpaces for routine activities and its effects on the intention to adopt CISpaces in future. The

questionnaire follows an adaptation of the Technology Acceptance Model (TAM) as proposed by Davis (1989) and its subsequent versions (TAM2, TAM3) (Venkatesh and Bala, 2008; Venkatesh and Davis, 2000). This model has been developed within the Human-Computer-Interaction literature to assess the acceptability of an information system according to various *factors*[3] measured by *indicators*,[4] and use such factors to predict the potential adoption of such a system and its use (Legris et al., 2003; Park, 2009; Wu and Wang, 2005). Factors are often hard to measure directly and, therefore, in TAM indicators are used as an indirect measure of the effects of these factors.[5] Indicators represent observable characteristics of a system and their analysis alone is useful to identify whether analysts consider these as positive characteristics of CISpaces with respect to their current activities. In addition, this model provides us with a systematic method, through the analysis of relationships between factors (using PLS-PM as described below), to determine strengths and weaknesses of CISpaces which might influence the adoption and use of CISpaces.

In TAM, one of the key factor is Behavioural intention (BI) of adopting CISpaces in this case, and it is influenced by the Perceived Usefulness (PU) and by the Perceived Ease of Use (PEOU) of the system. Factors BI, PU, PEOU are the core predictors of Use Behaviour, which represents how likely it is that analysts will use the system in the future. TAM3 extends the list of factors, by introducing additional external factors influencing PU and PEOU. While maintaining the core TAM components (BI, PU, PEOU), in this research, we adapt the list of external factors introducing some indicators more relevant to our study. We reduce existing lists to those focussed to a potential adoption of CISpaces at an early stage of development and exclude those directly focussed on actual usability since analysts' direct experience with CISpaces is limited. In addition, due to the limited number of participants and the reduced number of indicators, some indicators are regrouped into more general factors.

We adapted our model considering the following three themes: Analysts' *Experience* with similar tools, perceived *Utility* of CISpaces and its potential for *Adoption* in daily activities. From these themes we selected and introduced factors and relationships forming an adapted model, referred to as TAM-A.

Experience. In Section 3, the focus group highlighted that analysts use tools to support their activities, for organising input information, for collaboration, sensemaking and reporting. Here we are interested in understanding whether the analysts' *experience* (GEX) with similar tools has a positive influence in how they perceive CISpaces' ease of use.

Utility. TAM3 external factors are pertinent to our evaluation to establish whether CISpaces features are useful in improving daily activities. Key to establish the usefulness of CISpaces in our evaluation is the perceived improvement over the *output quality* (OQL), where output is the analysis in our work, and *result demonstrability* and *relevance* of CISpaces to the analysts' tasks (GRE). *Perception of external control* and *computer self-efficacy* (GPS) may positively contribute to ease of use.

Adoption. The TAM core factors BI, PU, PEOU will reveal whether there are grounds for CISpaces to be adopted in analysts' daily activities. Relationships between BI, PU, PEOU with other external factors might indicate strengths that can be exploited or weaknesses that need addressing in further developments of our system.

In Fig. 12 we show the resulting TAM-A graphical model with a description of each factor. Links to previous relevant TAM3 factors are shown in the figure.

[3] Other authors use the terms *constructs* and *latent variables*: we will consistently use factors in this paper.

[4] Other authors use the term *determinants*: we will consistently use indicators in this paper.

[5] All factors are measured in a reflective way in this research.

5.1.1. Experiment settings

The participants were six expert analysts from UK and US, who consented to participate in this academic study and have their opinion analysed and reported in publications. These participants where different from those involved in the study presented in Section 3. While recognising that the participant sample is relatively small, this is a highly expert group of participants in a field where recruitment is challenging.

Our experiment proceeded as follows. Participants were asked to watch a 10 minutes video demonstration of the CISpaces tool using a motivating scenario similar to our running example. The video showed step by step how to create an analysis, the use of argumentation schemes, crowdsourcing, provenance and the automatic evaluation of hypotheses. After watching the video, participants were asked to respond in writing to a set of closed and open questions using the questionnaire tool (TAM-A). The questions were provided to the analysts in a semi-randomised order with respect to the indicators of TAM-A. In total, participants were asked to respond to thirty-five multiple choice questions, evaluated using a 5-points Likert scale, and seventeen related open questions aimed at gathering further information on specific system features. Questions are provided in Appendix D.2. We believe this methodology was suitable for our research questions, the participant sample and the participants' limited engagement time available for our experiments.

The system used for evaluation is as described in Section 4, which developed from our previous version (Toniolo et al., 2015) in interactive components — including more robust and reliable integration of crowdsourcing and provenance analyses, and collaborative features — and additional AI functionalities — including preference handling and additional information retrieval (see Toniolo et al., 2016; Toniolo et al., 2014). Note that these latter additional functionalities, however, are out of the scope of this research and have not been used for evaluation.

5.1.2. Hypotheses

In formulating our TAM-A questionnaire, we considered what insight could be derived from indicators of characteristics of CISpaces alone, and from the relationships between factors constructed from indicators. We, therefore, identified two hypotheses for the study:

Hypothesis 1. Analysts respond positively to indicators demonstrating that CISpaces is considered useful in supporting the analysis process.

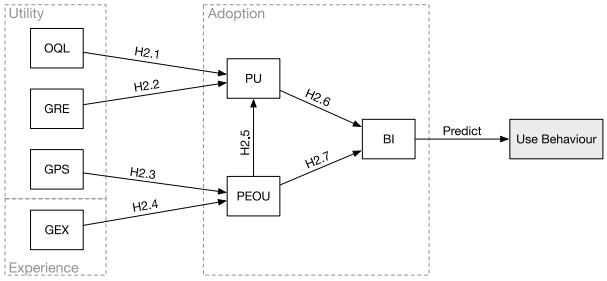
Hypothesis 2. All factors (OQL, GRE, GPS, GEX, PU, PEOU) have a positive effect of the degree of intention (BI) to use CISpaces.

Positive evidence for our hypotheses would support our general hypothesis that analysts are likely to intend to use CISpaces in their daily activities. This is based on the assumption (Venkatesh and Bala, 2008; Wu and Wang, 2005) that behavioural intention of adopting CISpaces would have a generally positive effect on actual use if this was to be deployed in future intelligence systems.

The analysis of results proceeded as follows. For H1, the individual factors extracted from the TAM-A model are analysed individually to establish analysts' response to the introduction of CISpaces in routine activities.

For H2, following common research on TAM models (Venkatesh and Bala, 2008), the data collected was analysed using the Partial Least Square Path Modelling approach (PLS-PM) (Lohmöller, 1989). This method combines factor analysis and regression by attempting to build correlations between the nodes of the graphical model in Fig. 12 along their edges.[6] Our assumption is that all edges in Fig. 12 represent a positive influence and we number each edge a sub-hypothesis. For

[6] The analysis was run using the R (The R Foundation, 2004) package *plspm* (Sanchez, 2013; Sanchez et al., 2015).



BI – *Behavioural intention:* the extent to which the analyst intends to use CISpaces.

PU – *Perceived usefulness:* the extent to which the analyst believes that CISpaces will enhance their job performance

PEOU – *Perceived ease of use:* the degree to which the analyst believes that using CISpaces will be free of effort

OQL – *Output Quality:* Perceived improvement on the quality of analysis

GRE – *Relevance and Results Demonstrability:*

- Relevance of CISpaces with respect to the analyst's daily job (TAM3-Job Relevance)
- Analysts' beliefs about whether CISpaces has a positive or negative impact on their daily job activities (TAM3-Result Demonstrability)

GPS – *Perception of system control:* Analysts' control beliefs regarding:

- Analysts' control beliefs regarding their personal ability to use CISpaces (TAM3-Computer self-efficacy)
- Analysts' control beliefs regarding the availability of organisational resources and support structure to facilitate the use of CISpaces (TAM3-Perception of external control)

GEX – *Experience with similar tools:* Analysts' experience with analytical/software tools similar to CISpaces acquired in previous work and in current activities

Fig. 11. Technology Acceptance Model TAM-A adapted for the study.

example, H2.1 indicates the hypothesis that the perceived improvement of Output Quality has a positive influence on Perceived Usefulness. We note that the PLS-PM procedure was run with a very small sample, imposing important limitations in the estimation power and in validating statistical significance. Therefore, this analysis only provides limited suggestions on the strength of the contribution of the factors to the degree of intention to use CISpaces.

To complement our hypotheses, we analysed answers to open questions and we provide quotes to support our claims. Answers were simply coded in positive comments, negative comments, or explanations given the short nature of the text provided.

Our results from the quantitative analysis are reported briefly below (see further details in Appendix E), and we then follow by contextualising our qualitative and quantitative results according to the themes that have guided the development of TAM-A : Experience, Utility and Adoption.

5.2. Results

Indicators. Fig. 11 summarises the results gathered from analysing the answers to the questionnaire using median and related interquartile ranges for a coding 1–5, where 1 indicates strong disagreement, and 5 indicates strong agreement; red and blue are used to show the two respective polarities. Detailed results are provided in Appendix E.1.

In order to understand whether the questionnaire given to analysts measured the same factor (agreement with the statement), a Cronbach's alpha (Cronbach, 1951) was run on the full results. We obtained a value of 0.89 indicating a high level of internal consistency for the scale used.

Questions related to experience have varied medians indicating more or less experience with a specific method similar to that used in CISpaces. For the remaining questions, we note that the median of most answers is above neutral (coded with value 3), which indicates generally an agreement with the statements. Our results provide positive evidence for Hypothesis H1: CISpaces is a useful tool to support analysts' tasks.

Factors and Relationships. PLS-PM runs in two phases, the first to establish the viability of the measurement model and evaluate the correlations between the indicators and their represented factors, and the second to evaluate the structural model, the hypothesised relationships between factors. Strength and direction of relationships obtained are shown in Fig. 13 and for convenience in the explanation we use arrows ↑↓ to represent positive or negative influences respectively. The figure also reports the regression weights, and the coefficients of determination of the factors, R^2. The effects between all factors but *PEOU→BI* are statistically significant at $p < 0.05$ with the limitations indicated above and high values of R^2 indicate that most of the variance in PU, PEOU, BI can be explained by their independent factors. We obtain positive influences between the relationships H2.2:GRE→PU$^{\uparrow}$, H2.4:GEX→PEOU$^{\uparrow}$, H2.6:PU→BI$^{\uparrow}$, and negative influences between the relationships H2.1: OQL→PU$^{\downarrow}$, H2.3:GPS→PEOU$^{\downarrow}$, H2.5:PEOU→PU$^{\downarrow}$. H2.7:PEOU→PU$^{\downarrow}$ is negative but not significant. While revealing information on the strength of these relationships, this analysis showed that the model created is limited in representing the data collected and in predicting power, with limitations in the measurement model and in the structural model, where indicators are only partially representing their factors and to an extent the relationships contradict common TAM results. We believe this is due to the limited sample size. Interpreting these values need caution due to these limitations, hence the conclusions we can draw are tentative observations used to complement the analysis. Further information on this analysis is provided in Appendix E.2.

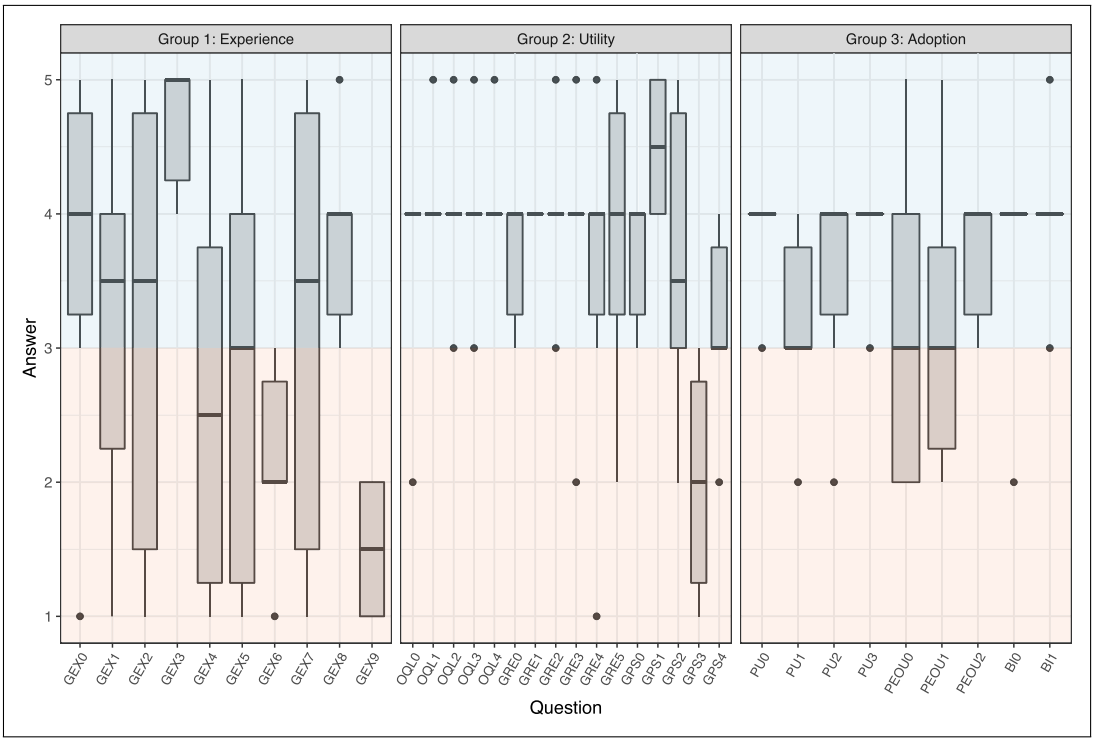


Fig. 12. Results Summary for all groups.

5.3. Evaluation discussion

We now discuss the results of the sub-hypotheses associated to the three groups of factors (Experience, Utility, Adoption) by analysing quantitative results and answers to the open questions.

5.3.1. Experience: Does analysts' previous experience align with CISpaces features?

The first set of results, on the left-hand side of Fig. 11, shows the indicators of analysts' experience (GEX): the first five indicate experience with analytical or software tools similar to those employed in CISpaces in previous work, while the subsequent five indicate current similar experience. With the exception of crowd-sourcing (GEX4), analysts had previous experience with computer-mediated analytical tools (GEX0), particularly for argument mapping (GEX1), provenance recording (GEX2), and collaborative analytical tools (GEX3). This shows that our target expert participant sample is familiar with similar tools. With respect to the question of whether those tools are currently used in participants' daily jobs, the results are more scattered. Analysts often use tools for collaboration in particular (GEX8) and provenance analysis (GEX7) as also highlighted in the focus group (Section 3), while they use argument mapping tools (GEX6) and crowdsourcing (GEX9) much less frequently.

The analysts' answers to our open questions informed us of similarities and differences of CISpaces compared to other analytical tools. Some examples are reported below:

- Analyst E: *"There are small similarities, there are other tools that seem more robust but they are not exactly like CISpaces."*
- Analyst F: *"CISpaces incorporates some features of other analytical platforms, but clearly goes much further. The provenance support is unique in my experience."*

These quotes show agreement in similarities with other tools particularly for example with link analysis tools (see Section 2). This comparison also shows drawbacks, some due to CISpaces being a research-level prototype which would require a more reliable and robust infrastructure for deployment. In answering these questions, it is also important to note that the system was directly compared by analysts with fully deployed commercial tools commonly used, highlighting intention and potential for adoption.

5.3.2. Utility: Do analysts believe that the features of CISpaces are useful in improving daily activities?

The second group of results in Fig. 11 report factors about features of CISpaces: improvement on output quality (OQL); relevance and result demonstrability (GRE); and ability to control and use CISpaces (GPS). Overall, analysts believe that CISpaces provides satisfactory features to fulfil these requirements as shown by medians mostly placed in the agreement part of the graph.

More specifically, there is evidence for the following factors:

- OQL: Analysts agree that CISpaces has the potential to improve and facilitate the analysis process
- GRE: Analysts agree that CISpaces is relevant, important, and pertinent to their daily activities. There is also a general agreement among analysts in being able to identify and explain the useful characteristics of CISpaces.
- GPS: There seems to be agreement in the perceived control of CISpaces, which is considered easy to use (GPS2). Analysts highlight that the system may not be compatible with other systems (GPS3), as expected being a research-grade prototype. There is disagreement on whether the system would change the way analysts work in daily activities (GPS4). Analysts' daily job can be completed using CISpaces (GPS0), although training would be important (GPS1). The PLS-PM analysis confirms that GPS is the least well represented factor by its indicators.

Output Quality. The perceived improvement on output quality was investigated further as this is an important reason for adopting the CISpaces solutions as novel approaches to intelligence analysis. To formulate the specific questions regarding this factor (OQL), we have

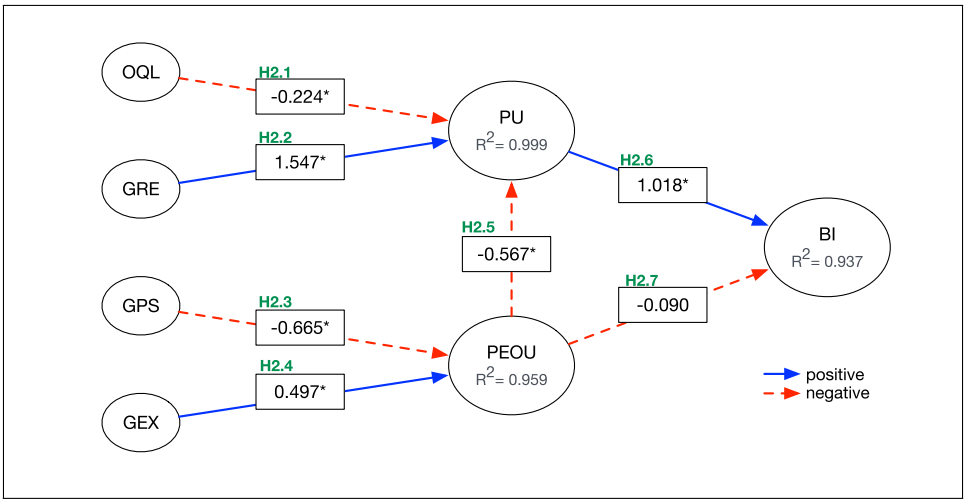


Fig. 13. PLS-PM Structural Equation Modelling Results, $^{*}p < 0.05$.

identified, through our previous focus group (Section 3), and analysis of the literature (Section 2) five criteria as key to assess analysis: time, robustness, confidence in the analysis, expression of intent, and decision-making over plausible hypotheses (OQL0–OQL4). Fig. 11 shows that analysts agreed with the proposition that CISpaces provides improvements across all these dimensions. In response to whether there are any further criteria to consider, analysts highlighted in particular the *confidence in the source documentation* and the *recognition of limitations of the available information* to avoid analysis only based on existing data.

We further asked analysts to define a robust analysis to better understand what contributes to this process:

- Analyst A: *"Amount of data used and experience of the analyst."*
- Analyst B: *"Robustness means details for me."*
- Analyst C: *"Analysis proven by a large amount of information."*
- Analyst D: *"Conclusions arrived at via logical analysis based on solid data. The uncertainty is noted in the report, along with possible alternatives."*
- Analyst E: *"The assurance that analysis has been subjected to critical review and been found to rest upon good data and good assumptions; undertaken according to valid methodology."*
- Analyst F: *"Audit trail of the conclusion and how it was developed."*

These answers show the importance placed on the rigour of the sensemaking process of analysis. Complementing the results on the output quality, these answers give some positive indication that the support that is offered by CISpaces can positively contribute to the analysts objectives and priorities during analysis in their daily activities.

Time. We note that there are conflicting views on whether the use of the system may – to a certain extent – limit the speed of the work (which is a critical characteristic for analysis as suggested in Section 3), particularly as it requires recording all analytical processes and may require additional training to use the argument mapping system. This is visible in the disagreement of GPS indicators but also in the scores of Output Quality (OQL0), where time is the only dimension with lower scores, albeit positive, highlighting a similar point that creating a visualisation of the analysis comes with a cost. When discussing disadvantages analysts expand on this issue:

- Analyst A: *"A weakness is in the time it would take to create the visualisation. It might be a better tool for training intelligence analysts."*
- Analyst B: *"Having to write them all nodes out and build the diagrams will increase the time required tenfold, and time is the one thing that analysts don't have."*

Tradeoffs between advantages of graphical representations of analysis and time and effort required to create those is an active research problem in the area of visual representations, highlighting that different representation types determine what information can be perceived (Zhang and Norman, 1994) and where different media may have advantages and disadvantages in providing contributions (Robinson and Pardoe, 2021). Beyond the objectives of this research, these tradeoffs would need consideration in interface design and further studies in future deployments of CISpaces.

Training. We additionally asked analysts about the perceived burden for training to use CISpaces. All analysts reported that a training module and manual are fundamental to be able to use the system. Furthermore, analysts highlighted other important training requirements:

- Analyst A: *"Understanding argumentation theory and analysis of competing hypotheses."*
- Analyst B: *"Determining how the information would flow into CISpaces is complicated as it depends on the type of analysis. In general, it is an easy to understand software and should have a short training requirement for analysts which are computer literate."*

Processes. Having discussed the general perspective of analysts views on CISpaces, we have also asked specific questions regarding the advantages in relation to the three core processes where automated support is provided: sensemaking, provenance and crowdsourcing through open questions. In highlighting strengths of the support to the analysis process, analysts mention:

- Analyst B: *"The strength is the visualisation of two (or more) hypotheses. Seeing where each hypothesis is supported can help determine which is the better choice."*
- Analyst D: *"An individual's reasoning is documented for others to follow/collaborate on."*
- Analyst E: *"It serves as a forcing function. [to structure the analysis]"*

For *sensemaking* there is consensus that the system would help share world views between teams, and while reservations remain with respect to time taken for creating the visualisation, mid to long term analyses would particularly benefit from this approach. Analyst C suggested that CISpaces allows others to see the reasoning behind hypothesis so that everyone can view what an analyst was thinking when they were trying to understand the situation and could better identify questions. Analyst A added *"It lays out visually for everyone to see where there are issues."*

With respect to *crowdsourcing*, analysts agree that this is a useful capability, bringing new information to the analysis. Opinions are more

divergent in that care should be taken to control reliability and expertise of the crowd to avoid potential misinformation. Active research in the area of crowdsourcing focuses on establishing and ensuring reliability of reports and can be easily integrated with CISpaces (e.g., Ouyang et al., 2016b).

Finally, analyst A informed us of issues with provenance of the analysis: *"Often a single piece of information can be reported in different ways causing an analyst to believe there are several separate pieces of information rather than a single one reported multiple times"*. All analysts agreed that tracking and analysing *provenance* is a very important feature provided that it remains non-editable and unobtrusive unless requested as currently designed in CISpaces.

5.3.3. Adoption: Do analysts believe that there are grounds for CISpaces to be adopted in their daily activities?

In the right-hand side of Fig. 11, we report the results for the last group. As for the previous questions, we notice a general agreement of analysts in perceived usefulness of the tool, and intention to adopt the system. There are, however, some drawbacks in perceived ease of use (PEOU0, PEOU1).

To better understand the causes of these drawbacks and how they impact the general behavioural intention, we turn to the results obtained with the analysis of PLS-PM, in other words considering whether there are relationships between the factors that could inform positive or negative contributions to the intention of adopting CISpaces.

In relation to perceived ease of use, we found a moderate positive effect of experience (GEX→PEOU$^{\uparrow}$) with similar tools. This presumably means that the more experience analysts have with similar tools, the easier it would be for them to use CISpaces. Disagreement in perception of system control can hinder these results (GPS→PEOU$^{\downarrow}$), however.

For perceived usefulness (PU), we note that the most influential contribution is provided by the relevance and result demonstrability factor (GRE→PU$^{\uparrow}$), but ease of use negatively influences the results (PEOU→PU$^{\downarrow}$). A relative minor negative influence on PU is given by perceived improvement in output quality (OQL) which is contradicting our assumption. OQL ratings are higher than PU in particular with respect to robustness, accuracy, expressiveness, and decision-making, showing that there is confidence in improvement perceived, but this is likely to indicate that to obtain more positive results in perceived utility, other features influencing ease of use need to be improved for adoption.

The most positive contributions to behaviour intention (BI) are provided by the perceived utility PU→BI$^{\uparrow}$ and indirectly this is provided by the relevance and result demonstrability factor (GRE). The ease of use indirectly and directly negatively influences BI even though the influence is relatively minor.

Corroborating our previous results this shows that while the principled solutions provided by CISpaces are deemed important and relevant for the analysts' daily activities, it is as important for adoption to ensure that the system is usable and integrated with other systems that analysts use. Further developments of CISpaces together with follow up user studies are needed to draw conclusions on the interface usability aspects of the system, which is, however, beyond the scope of this evaluation.

Our last open questions to analysts were in relation to additional capabilities and applications of CISpaces. The possibility of automatically creating intelligence reports from the analysis following some examples from previous research (Hossain et al., 2011) was envisaged as potential additional feature of CISpaces, and when asked all analysts agree that this would be very useful. In terms of applications, analysts highlighted opportunities specifically for analysis with medium to long term timeframe, such as for strategic analysis (Analysts B,D and E), and for complex problems with many moving parts and numerous analysts collaborating (Analysts A,C,D,and F). The use of CISpaces for training has been highlighted as an important opportunity and further to record the analytical process and ensuring robustness (Analyst F).

When looking at the ability to share information, besides some specific interface and usability issues which could be addressed with further development to higher technology readiness levels, a concern emerged in relation to the ability of sharing information, with issues raised about security restrictions and limited bandwidth. Analysts also recommended attention when deploying CISpaces in working environments to ensure compatibility with systems the organisation already adopted, to mitigate additional effort in training and usage.

5.4. Evaluation remarks

To conclude the discussion, our analysis suggests that the principled AI methods implemented in CISpaces have potential to advance performance in intelligence analysis. CISpaces has been designed to be a basic research prototype (TRL 3), nevertheless, during the evaluation analysts compared it with commercial systems they use everyday highlighting intention to adopt the system. Analysts also agree that the AI methods implemented in CISpaces are useful in improving their daily activities, in particular thanks to the perceived utility of the outputs CISpaces generates. Analysts recommended that appropriate training and integration with other systems is provided. Tradeoffs have been also highlighted in the time required to build a visualisation which inevitably has a cost. The highlighted drawbacks in CISpaces, however, do not lie on the AI methods underpinning the system: for successful adoption, CISpaces will need data integration with existing organisational standards both for the input and the output of information as well as more advancements in the user interface.

6. Discussion

6.1. Related work

Formal models of argumentation are used to capture different types of conflicts arising between information (Bex et al., 2003; Walton et al., 2008), to resolve these conflicts (Čyras and Toni, 2016; García and Simari, 2004; Modgil and Prakken, 2014), and to evaluate the reliability of conclusions (Parsons et al., 2011; Toniolo et al., 2014). Argumentation techniques, however, focus on decision-making, and such methods may require training to be used by analysts due to the extensive formalisation required. Argument mapping provides intuitive and effective support for critical thinking (Reed and Rowe, 2004; van Gelder, 2007), and shows advantages particularly in enriching and understanding of a problem over for example text representations (Carneiro et al., 2021), but does not offer support for reasoning. Argument mapping and formal argumentation can be combined to visualise and analyse arguments or conclusions (Leiva et al., 2019; Reed et al., 2017). In this work, we also combine these approaches to enable analysts to directly interact and benefit from a computational model of argumentation in the construction and evaluation of hypotheses. We chose a subset of argumentation concepts and established a formal correspondence with intelligence analysis concepts, which were identified through a co-design process and verified via a focus group (see Section 3). This subset was intentionally small to limit the training burden. Recent research has focussed on establishing connections between formal argumentation and human intuition (e.g. Cerutti et al., 2014; Cramer and Guillaume, 2019; Toniolo et al., 2018) and can guide future work on studying formally how analysts' training approaches and methods align with formal argumentation models.

To support analysts in better selecting hypotheses, we employ crowdsourcing to facilitate the acquisition of additional evidence and provenance to explore the credibility of information. In recent research, agent-based approaches have been applied to crowdsourcing to automate decision-making on behalf of the requestors such as who to hire (Kamar et al. 2012), which is more akin to a trust decision making problem. More traditional approaches focus on result aggregation to mitigate biases from unreliable sources (Brabham, 2008; Ouyang et al., 2016b; Whitehill et al., 2009). Similarly, work on provenance is

primarily concerned with data quality and interoperability (Hartig and Zhao, 2009). In this research, we study how to automatically *interpret* provenance and crowdsourced data to assist analysts and *integrate* this information in generating coherent explanations of observed evidence.

Provenance is a novel application for argumentation-based frameworks. The approach that first discussed the use of arguments underpinned by provenance is by Chorley et al. (2008), where provenance is recorded for justifications provided by users during the assessment of policy options. Using provenance for assessment of information quality has also been explored. Hartig and Zhao (2009) proposed a measure of timeliness using a specific model of provenance, according to creation and access time. In our research, we provide a method to extract information from the provenance elements according to simple intuitive patterns. More complex quality measures could also be extracted providing further automation to the analysis of provenance (Pipino et al., 2002; Toniolo et al., 2014).

6.2. Comparison with other tools

The AI techniques we implemented in CISpaces advance intelligence analysis across several dimensions: (1) visual exploration of relationships between pieces of information; (2) sensemaking and hypotheses generation; (3) evidence gathering via crowdsourcing; (4) provenance reasoning; (5) collaboration with other analysts. In related research efforts, we also experimented with social sensing (Toniolo et al., 2016) and with automatic information extraction from OSINT and report creation via natural language generation (Cerutti et al., 2019) via the spin-off CISpaces.org (Cerutti et al., 2018b).

To our knowledge, no other tool allows for such capabilities while ensuring a coherent and consistent analyst experience. There are, however, several tools that can be exploited for each of the previously mentioned capabilities.

6.2.1. Information collection and hypotheses generation

Existing visual analytics tools are primarily concerned with supporting the development of situational understanding by identifying links among – and structures present in – existing information. For instance, i2 Analyst's Notebook (IBM, 2017) offers a suite of views for analysts to organise and link information and perform sophisticated network analyses. Jigsaw (Stasko et al., 2008) enables analysts to explore different views of information available for decision-making, including viewing relationships among entities, documents, topics and visualising event/observation timelines. INVISQUE (Rooney et al., 2014), together with a number of other tools that regularly participate in the visual analytics challenge (Visual Analytics Community, 2006), offers a "suite" of perspectives over data that can be used by analysts to query and support sensemaking by facilitating access to evidence. Generally these tools are primarily focussed on organising and collating information for the analysis to take place, but on the other end of the conceptualised model of intelligence analysis proposed by Pirolli and Card (2005), once analysts have formulated available hypotheses, a variety of tools are designed to support hypotheses evaluation and selection. For example, the Xerox PARC ACH tool (Stefik, 2014), the Open ACH (Burton and Knowles, 2010) and others (e.g, Tecuci et al., 2010) provide automated means to perform a weighted ACH propagating uncertainties from evidence to hypotheses to weigh alternatives. As discussed in Section 2, a recent study from Baber et al. (2016) shows that systems available to analysts are limited in providing support for hypotheses exploration and CISpaces is designed to address this gap.

6.2.2. Provenance reasoning

As discussed in Section 4.3, the origins of information (including information from the crowd), and how and by whom this information is interpreted during analysis are important to establish the credibility of hypotheses. Provenance can be used to annotate how, where, when and by whom some information was produced (Moreau and Missier, 2013). CISpaces is almost unique in its ability of supporting reasoning about provenance to the point of having the possibility of semi-automatically refute hypotheses on the basis of provenance information. Among the few other tools addressing this issue, TRELLIS (Gil and Ratnakar, 2002) enables information received from different sources to be suitably annotated, highlighting contradictions and how these relate to the trustworthiness of sources.

6.2.3. Collaboration with other analysts

Although we did not stress it in this paper, CISpaces allows for the creation of shared canvases so to enable multiple analysts to operate on a same view of the analysis. Among other tools providing similar capabilities, CACHE is a collaborative ACH environment offering some support during the process of deciding upon the most likely hypothesis (Billman et al., 2006). The CACHE tool provides shared access to enable participants to weigh evidence as a team. The paper describing CACHE (Billman et al., 2006) also includes a user evaluation of the system, studying the effects of group composition in mitigation of biases through computer mediated ACH. The experiment shows the potential for collaborative systems to support the process of analysis and have influenced both our research and evaluation. Both CISpaces and CACHE appear to be useful for the work of analysts. Among other tools, Entity Workspace (Bier et al., 2008) supports collaboration in comparing and deciding upon the most likely hypothesis. POLESTAR (Pioch and Everett, 2006), instead, allows the sharing of an individual portfolio of analysis, enabling different users to make suggestions/critiques.

6.2.4. Information requirements, crowdsourcing and social sensing

To make sense of a situation, analysts need to rapidly analyse and link this information to other contextual evidence, to identify explanations of the environment. Gathering additional information is necessary to avoid the rejection of hypotheses on the basis of insufficient evidence (Heuer, 1999) as also highlighted by analysts in our evaluation (Section 5). Information requirements can be targeted by evaluating the value of information and argumentation frameworks similar to that presented by CISpaces may be suitable for this purpose (Robinson and Pardoe, 2021).

The variety of sources that analysts must take into account has recently changed significantly in particular for what concerns open source intelligence (OSINT) such as social media. There are real challenges concerning how to exploit OSINT in effective and reliable ways; for example, the nature of social media sources is such that it is often difficult to distinguish between witness information and hearsay. Reliability of sources and reports is an important concern in these settings, CISpaces can be extended to include more complex aggregations of results to mitigate these issues (Ouyang et al., 2016a; 2016b) or with automated support in detecting those responsible for propagating misinformation (Paredes et al., 2021). Further, this should not be seen solely as analysts passively consuming open source intelligence, but utilising networks of contributors through crowd-sourced queries. For instance, public platforms have been shown to be useful sources of information in disaster response, for example in mapping the geography of Haiti after the 2010 earthquake (Zook et al., 2012). Social networks have created greater opportunities to leverage social sensing as methods to collect data about the environment, and in Toniolo et al. (2016) we demonstrated how conversational interfaces can be linked to CISpaces. In social sensing, people act as sensors, share information within a network or respond to data or opinion requests (Burke et al., 2006).

There are several other research directions for crowdsourcing in intelligence analysis. The role of crowd-sourced intelligence and its classification within traditional or new parameters, is itself an active research topic (Stottlemyre, 2015). Recently, the IARPA CREATE project (IARPA, 2017) led the development of the SWARM Systems (Sinnott et al., 2019), a collection of platforms for integrating analytics techniques and informal argumentation to support analysis through crowd-sourced intelligence. The SWARM interface provides a portal which combines capabilities of question-answering platforms and

shared document editing systems. Groups in the public and professional domains can contribute with draft reports and opinions to intelligence requirements. A recent evaluation of SWARM demonstrates improvements in the quality of the reports provided when using the system (van Gelder et al., 2020) highlighting the potential for systems to include crowdsourced contributions.

6.2.5. Additional capabilities and CISpaces.org

Developments in natural language text analysis and computational linguistics have enabled greater automation in information extraction. These methods aid in the discovery of criminal groups, patterns of interaction, activity timelines, and so on. Event extraction and characterisation may also be provided from news articles (Lu et al., 2016), and other research has concentrated on the analysis of more complex texts. XIP-Cohere (De Liddo et al., 2012), for example, uses mixed automatic and human annotation to extract and summarise contrasting ideas from documents. Mining arguments from intelligence analysis reports has been studied in Kang and Sinnott (2018) following the significant developments in the area of argumentation mining seen in recent years (Lawrence and Reed, 2020). In CISpaces.org (Cerutti et al., 2018b), a spin-off of the project we are reporting in this paper, we employed natural language processing techniques for automatic information extraction from Twitter, thus demonstrating the effectiveness of linking the system to OSINT sources.

In addition, CISpaces.org (Cerutti et al., 2018b; 2019) also employs natural language generation techniques to produce explanations that could be included as reports. To our knowledge, only the Analyst's Workspace (Hossain et al., 2011) provides similar functionality.

7. Conclusions

In this paper, we illustrate how novel AI methods, based on a combination of argumentation theory, crowdsourcing Bayesian analysis, and provenance recording, advance performance in intelligence analysis. Our research is based on an extensive consultation involving highly-trained, professional intelligence analysts from UK, US, and international agencies in a process of elicitation of requirements, co-design and co-development of CISpaces and finally an evaluation of the approaches.

Recruiting experts in intelligence analysis is highly complex due to the nature of their role, and they are an extremely scarce resource, especially those who are highly-trained and with extensive expertise. Due to this challenge, the number of participants in the two studies is limited, however, it is within the lower end recommendations for qualitative research, specifically concerning purposive homogeneous studies (Guest et al., 2006; Miles et al., 2013), those being highly focused on analysts' needs with similar training background and work objectives.

Our experiments conclude that the novel, principled AI methods implemented in CISpaces may advance performance in intelligence analysis. During the evaluation, CISpaces – despite having being designed to be a basic research prototype (TRL 3) – has been benchmarked against commercial systems being used everyday by analysts. Analysts agree that the AI methods implemented in CISpaces are useful in improving their daily activities, in particular thanks to the perceived improved utility of the outputs CISpaces generates. Analysts suggest that CISpaces has potential particularly for collaborative and complex analysis, training novice analysts and to maintain an audit trail of the formation and selection of hypotheses. The analysts' evaluation highlights drawbacks in CISpaces that, however, lie not in the principled solutions, but rather in its interfaces with the data sources and with the user. For successful adoption CISpaces would need further integration with existing systems and further training. These aspects, while being essential for commercialisation, are beyond the scope of this paper.

In designing CISpaces, we worked closely with professional analysts to design a set of objectives for our system to help structure and record analysis, and to facilitate and improve the quality of analysis through automated support. Our supporting features include reasoning about plausible hypotheses through a small set of computational argumentation concepts, support to analyse provenance of information and aggregate and report crowdsourced information. While CISpaces is unique in bringing these features together, this is a limited set addressing only some aspects of the analysis problems and there are many directions in which a system such as CISpaces can be expanded. For example in Section 6, we noted important current research trends in extracting and analysing large amount of information from open sources such as social media or open crowdsourcing tasks. Current advancement on argument mining may help import arguments from secondary sources. Automatically establishing credibility of this information and likelihood of events would further inform analysis and have potential to improve analysts' tasks. Further and more complex policies for collaboration and integration of analyses would also be beneficial as sharing intelligence is often a critical issue between organisations. Future work may focus on identifying autonomous methods to integrate evidence analysis themes more tailored to specific intelligence requirements, in similar ways as we import crowdsourcing and provenance in CISpaces, for example for geo-spatial data or event causality.

From an evaluation perspective, follow up studies with analysts would be important to better understand the extent of the support that CISpaces provides, and to establish the level of training needed for analysts to model a problem in terms of argument components. Future work in this direction would provide a more in-depth evaluation of the potential for the use of this system as well as further insight into the most suited level of automation for analysis tasks. In future, the restricted set of argument components we have chosen for CISpaces can be extended (e.g., with preferences, strict rules, additional schemes and critical questions, or alternative semantics), supported by studies focussed on understanding tradeoffs between components and training burden required to adopt these new concepts.

CISpaces is devised to support intelligence analysts in the military, however, there is growing need for tools supporting deep thinking and which limit cognitive biases in a variety of disciplines, which may benefit from the CISpaces support from scientific enquiries (Cerutti and Pearson, 2018) to legal analyses (Cerutti et al., 2018a) demonstrating its versatility. CISpaces.org (Cerutti et al., 2018b) shows that a substantial part of the code we developed for CISpaces can be easily adapted to different graphical user interfaces. Our research and evaluation demonstrates the potential of an integrated tool building on state-of-the-art AI and argumentation-based techniques to aid human effort in better interpreting evidence in highly complex environments.

CRediT authorship contribution statement

Alice Toniolo: Conceptualization, Methodology, Software, Writing – review & editing. **Federico Cerutti:** Conceptualization, Methodology, Software, Writing – review & editing. **Timothy J. Norman:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition. **Nir Oren:** Conceptualization, Methodology, Writing – review & editing, Funding acquisition. **John A. Allen:** Conceptualization, Methodology, Funding acquisition, Software, Writing – review & editing. **Mani Srivastava:** Conceptualization, Methodology, Resources, Funding acquisition, Writing – review & editing. **Paul Sullivan:** Conceptualization, Methodology, Resources, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We would particularly like to acknowledge the contribution made by the late Paul Sullivan to this work. Without his expertise, this research would not have been possible. We would like to thank the professional analysts from the UK, US and international agencies for their support in developing this research.

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence and was accomplished under Agreement Number W911NF-06-3-0001. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

Appendix A. Mapping CISpaces to ASPIC+

Here we give the formalisation of the mapping from CISpaces to ASPIC+ (Modgil and Prakken, 2014) argumentation framework, which is restricted to ordinary premises and defeasible rules without preferences, and we discuss how the argumentation schemes are considered in this formalism (see Section 4.1).

A1. An ASPIC-like argumentation framework

Definition 1. An *argumentation system AS* is a tuple $\langle \mathscr{L}, ^-, \mathscr{R} \rangle$ where $\mathscr{L}$ is a logical language, $^-$ is a contrariness function, and $\mathscr{R}$ is a set of defeasible rules. The contrariness function, $^-$, is defined from $\mathscr{L}$ to $2^{\mathscr{L}}$, s.t. given $\varphi \in \overline{\varphi}$ with $\varphi, \phi \in \mathscr{L}$, if $\phi \notin \overline{\varphi}$, φ is called the *contrary* of ϕ, otherwise if $\phi \in \overline{\varphi}$ they are *contradictory* (including classical negation $\neg$). A *defeasible rule* is $\varphi_0, \cdots, \varphi_j \Rightarrow \varphi_n$ where $\varphi_i \in \mathscr{L}$.

We refer to a rule $\alpha \Rightarrow \beta$ as r, where α is the *antecedent* and β is the *consequent*.

Definition 2. A *knowledge-base K* is a subset of the language $\mathscr{L}$. An *argumentation theory* is $AT = \langle K, AS \rangle$.

An *argument Arg* is derived from the knowledge-base K of a theory AT. Let $Prem(Arg)$ indicate the premises of Arg, $Conc(Arg)$ the conclusion, and $Sub(Arg)$ the subarguments:

Definition 3. An argument Arg is defined as:

- $Arg = \{\varphi\}$ with $\varphi \in K$ where $Prem(Arg) = \{\varphi\}$, $Conc(Arg) = \varphi$, $Sub(Arg) = \{\varphi\}$.
- $Arg = \{Arg_1, \cdots, Arg_n \Rightarrow \phi\}$ if there exists a defeasible rule r in AS such that $Conc(Arg_1), \cdots, Conc(Arg_n) \Rightarrow \phi \in \mathscr{R}$ with $Prem(Arg) = Prem(Arg_1) \cup \cdots \cup Prem(Arg_n)$, $Conc(Arg) = \phi$ and $Sub(Arg) = Sub(Arg_1) \cup \cdots \cup Sub(Arg_n) \cup Arg$.

Attacks are defined as those arguments that challenge others, and defeats are those attacks that are successful: we use only rebutting, when two arguments have contradictory conclusions; and undermining, when the conclusion of an argument is the contrary of a premise of another argument. Since we do not consider preferences, attacks are always successful. Moreover, while we do not explicitly encompass undercutting — when the conclusion of an argument is the contrary of a defeasible rule — it can be represented with the introduction of an additional premise, as often considered in literature, see for example Dung et al. (2009), and Čyras and Toni (2016) for a discussion.

Definition 4. An argument Arg_A *defeats* an argument Arg_B iff:

- Arg_A *rebuts* Arg_B on $Arg_{B'}$ iff $Conc(Arg_A) \in \overline{\varphi}$ for $Arg_{B'} \in Sub(Arg_B)$ such that $Arg_{B'} = \{Arg_{B1''}, \cdots, Arg_{Bn''} \Rightarrow \varphi\}$.
- Arg_A *undermines* Arg_B on φ iff $Conc(Arg_A) \in \overline{\varphi}$ such that $\varphi \in Prem(Arg_B)$.

An abstract argumentation framework (Dung, 1995) AF corresponding to an AT includes a set of arguments as defined in Def. 3 and a set of defeats as in Def. 4. Sets of acceptable arguments (aka. extensions) in an AF can be computed according to a semantics. The set of extensions that we consider here is $\widehat{\xi} = \{\xi_1, \ldots, \xi_n\} \cup \{\xi_S\}$ such that each $\xi_i = \{Arg_a, Arg_b, \cdots\}$. The extensions $\xi_1, \ldots, \xi_n$ are the *credulous-preferred extensions* identified via preferred semantics; i.e., maximal wrt. set inclusion extensions that are conflict free (i.e., no arguments in any extension defeat each other), and admissible (i.e., each argument in the extension is defended against defeats from "outside" the extension). The *skeptical-preferred extension* ξ_S is the unique intersection of the credulous-preferred extensions.

A2. CISpaces argumentation theory

In CISpaces the core view where the analysts construct the analysis is called WorkBox. Here, we define the mapping of a WorkBox view to the corresponding AT, called WAT. A Pro-link in the Workbox is textually represented as $[p_1, \cdots, p_n \xrightarrow{p} p_\phi]$ indicating that the Pro-link has $p_1, \cdots, p_n$ as incoming nodes and has the outgoing node p_ϕ.

Definition 5. A *WAS* is an argumentation system $\langle \mathscr{L}, ^-, \mathscr{R} \rangle$ constructed as follows:

i $\mathscr{L}$ is a propositional logic language, and a node corresponds to a proposition $p \in \mathscr{L}$. The WAT set of propositions is $\mathscr{L}_w$.

ii The set $\mathscr{R}$ is formed by rules $r_i \in \mathscr{R}$ corresponding to Pro-links between nodes such that: $[p_1, \cdots, p_n \xrightarrow{p} p_\phi]$ is converted to $r_i : p_1, \cdots, p_n \Rightarrow p_\phi$

iii The contrariness function between elements is defined as: (i) if $[p_1 \xrightarrow{c} p_2]$ and $[p_2 \xrightarrow{c} p_1]$, p_1 and p_2 are contradictory; (ii) $[p_1 \xrightarrow{c} p_2]$ and p_1 is the only premise of the Con-link, then p_1 is a contrary of p_2; and (iii) if $[p_1, p_3 \xrightarrow{c} p_2]$ then a rule is added such that p_1 and p_3 form an argument with conclusion p_h against p_2, $r_i : p_1, p_3 \Rightarrow p_h$ and p_h is a contrary of p_2.

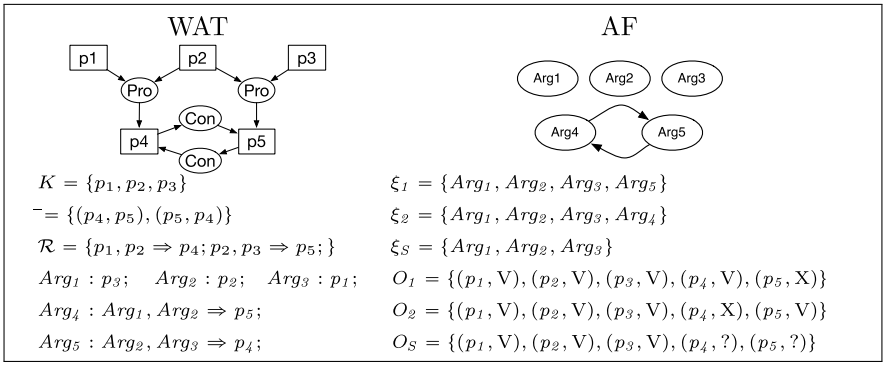


Fig. A.1. WorkBox Argumentation Theory (*WAT*) and its translation first into ASPIC, and then into a Dung's Argumentation Framework to derive preferred extensions and hence intelligence analysis hypotheses.

Definition 6. A *WAT* is a tuple $\langle K, WAS \rangle$ such that K is composed of propositions p_i, $K = \{p_j, p_i, \cdots\}$, where:

i for a set of rules $r_1, \cdots, r_n \in \mathscr{R}$ indicating a cycle (i.e. for all p_i that are consequents of a rule r there exists some r' containing p_i as antecedent), then $p_i \in K$ if p_i is an info-node;
ii otherwise $p_i \in K$ if p_i is not a consequent of any rule $r \in \mathscr{R}$.

The mapping from *WAT* to the ASPIC+ framework is similar to that adopted between OVA+ and various solvers (Reed et al., 2017) with the exception of Con-links mapping (w.r.t. Def. 5(*iii*)) and inference cycles (w.r.t. Def. 6(*i*)). CISpaces stores data in the Argument Interchange Format (Cerutti et al., 2018c). In particular, the option for an analyst to write $p_1, \cdots, p_n$ linked to p_φ with a Con-link is mapped to the argumentation framework with a rule that has a contrary consequent to p_φ, whereas in other frameworks each individual p_i with $i = 1, \cdots, n$ is considered as a contrary to p_φ. Our approach allows the representation of a contrary of a term $(p_\phi, p_\varphi) \in \cdot$ as in other models, for example an unreliable messenger may be a contrary for them to be in a position to deliver a message, however it also permits a compact representation of additional constraints. For example, inspired by Caminada and Wu's Tandem example (Caminada and Wu, 2011), we might consider three gangs, where each gang would only collaborate with another gang if the third is not involved. With a representation where a Con-link is created for a pair of every two gangs against the other, we obtain preferred extensions that include pairs of gangs, rather than a single gang. Additionally, premises of inferences that form a cycle in existing models are not considered part of the knowledge base. In our framework, we are able to distinguish between information and claim nodes, and we chose to consider info-nodes as asserted propositions part of the knowledge-base.

Hypotheses identification. In CISpaces we use a *WAT* as translation of a WorkBox to evaluate plausible conclusions and to show available hypotheses to the user.

Definition 7. Given an *AF* corresponding to a *WAT*, a proposition p_i and an existing extension ξ_j, p_i is acceptable if there is an argument $Arg_i \in \xi_j$ that has conclusion p_i.

CISpaces uses the efficient solver developed by Cerutti et al. (2016) to identify preferred extensions. Given the set of all extensions $\widehat{\xi}$ in the *WAT*, the analyst is presented with n colouring options that indicate when a node contains a statement that can be supported, unsupported or undecided. A node is supported if it contains a piece of information that is acceptable or is defended against its defeaters. A node is unsupported if it is rejected, and undecided if it has insufficient grounds to be either supported or unsupported.

Definition 8. The set of options $\mathscr{O} = \{O_1, \cdots, O_n\}$ for a *WAT* is a set of cardinality $|\mathscr{O}| = |\widehat{\xi}|$ where each option $O = \{(p_i, col_i)$ s.t. $p_i \in \mathscr{L}_w, col_i \in \{\mathrm{V}, \mathrm{X}, ?\}\}$. The assignment of col_i for p_i given an extension $\xi_j \in \widehat{\xi}$ is:

- col_i = V (supported), if p_i is acceptable in ξ_j;
- col_j = X (unsupported), if p_i is a conclusion of an argument Arg_A that is defeated by $Arg_B \in \xi_j$;
- col_j = ? (undecided) otherwise.

The set of supported conclusions consists of the supported elements of an option O_i^V. Each option is available to the analyst for inspection, and represents the semantic mapping from extensions to hypotheses as partial explanations of a world. An example mapping from a WorkBox argumentation theory (*WAT*) to an abstract argumentation framework (*AF*) and the set of options (in this case O_1, O_2 and O_S) is presented in Fig. A.1.

A3. Mapping argumentation schemes

Let us recall once again the *argument from expert opinion* (Walton et al., 2008) from Section 4 completed with implicit premises for the conclusion to hold:

- Source E is an expert in domain S containing proposition A,
- E asserts that proposition A is true,
- Implicit: E is a credible source, E is reliable, there is evidence supporting A
⇒ Therefore, A may plausibly be true.

Critical questions include:[7]

- *CQEO1* "How credible is E as an expert source?";
- *CQEO2* "Is E an expert in the field that A is in?";
- *CQEO3* "Is it the case that E has asserted the claim?";
- *CQEO4* "Is E reliable as a source?";
- *CQEO5* "Is A consistent with the testimony of other experts?";
- *CQEO6* "Is E's assertion based on evidence?".

In a *WAT* each scheme is translated to a rule. Partially inspired by the approach of Modgil and Prakken (2014), and Dung et al. (2009), assume that we give predicate-like labels to propositions contained in each Claim or Info box, a full instantiation of a *WAT* scheme for an expert opinion scheme can be seen as:

$$r_{EO} : expert(E,A), assert(E,A), within(A,S)$$
$$credible(E,S), reliable(E), evidence\_sup(A)$$
$$\Rightarrow hold(A)$$

In the presentation above, while representing a single rule, the first line represents the ordinary scheme, the second line highlights the assumptions and the third line is the conclusion.

When a specific link is tagged in the CISpaces interface with a particular type of inference, e.g., LEO, the premises nodes in turn can be tagged with one of their premises types. This enables a set of critical questions to be used as pointers to other arguments that may challenge this inference. When CQs are enabled, they are showcased on the Workbox as Con-links with a negative answer to the question. This is, however, not sufficient to cover the different types of attacks that a scheme provides, as all attacks will be considered contrary, but we hold that some critical questions point to contradictory conclusions in particular when there might be alternative conclusions, but also in case of alternative premises. In this second case, we have designed undermining critical questions as ways for an analyst to consider alternative views, in a what-if process, and therefore prompt the analyst to provide further evidence to discard an option. Hence, CQs are mapped to a *WAT* according to the type of attack as:

- *undermining CQs* as attacks to premises: a Con-link is mapped to a contradictory relation. For example, CQEO2: $\neg expert(E,A)$
- *undercutting CQs* challenging an exception of the inference rule. In CISpaces, undercuts are contrary underminers to propositions implicit in the scheme; e.g. CQEO1: $non\_credible(E,S) \in \overline{credible(E,S)}$
- *rebutting CQs* as contradictions of the conclusions: a Con-link is mapped to a contradictory relation. For example, CQEO5: $\neg hold(A)$.

Following this approach the two core schemes provided in Fig. 6 can be formally translated in rules as shown in Fig. A.2, by labelling the propositions in CISpaces boxes as premises or assumptions in a predicate-like format as above. In the example in Fig. 5 we see a critical question CQCE6 as an example of an rebutting critical question, asking whether there are other causes to the explosion at the pumping station alternatives to the IED (Improvised Explosion Device). Node and text contained in p_{11} is automatically created when CQCE6 is enabled on p_{13}. This Con-link is then mapped to a contradictory relationship in Fig. 7 leading to the two alternative partial hypotheses of an explosion due to a natural leak of explosive gases from the water system between $p_{11} - p_{13}$.

Appendix B. Crowd-sourced evidence

In this section, we outline the technicalities involved in defining and analysing crowdsourced tasks in Section 4.2.

Definition 9. Given a $WAT = \langle K, WAS \rangle$, a crowdsourcing task, T, is a tuple $\langle p_t, q, Q, d_t, n_t, c_t \rangle$ where p_t is a proposition in K, q_t is the overall question that the task is designed to address, $Q = \{q_1, \cdots, q_n\}$ is a set of sub-questions to be addressed to the crowd, d_t is the deadline, n_t the minimum number of participants, and c_t the target crowd.

When the task is initiated, as in typical crowdsourcing models, the analyst creates a form with questions Q, to be answered by the contributors, such that each question $q_i \in Q$ is a tuple $\langle type_i, text_i, options_i, ev_i \rangle$, where $type_i$ is either *categorical* or *numerical*; $text_i$ defines the question asked to the crowd; $options_i$ indicate the space of possible answers; and ev_i is a function that maps a number/category to its evaluation $\{Pro, Con\}$. If $type_i = categorical$, $options_i = \{cat_{i1}, \cdots, cat_{in}\}$ is the space of possible answers, where for each $cat_{ij} \in options_i$ the analyst chooses $ev_i(cat_{ij}) = Pro$ (or Con) if cat_{ij} is a reason for believing p_t (or $\neg p_t$). If $type_i = numerical$: the answers are real numbers $n \in \mathbb{R}$; analysts define $ev(n) = \{Pro, Con\}$ as specific values for n to be considered as Pro or Con for p_t. We only consider complete reports.

A report $\widehat{\Omega}^j$ for participant j contains an answer $\widehat{\omega}^j_k$ for each question q_k, $\widehat{\Omega}^j = \{\widehat{\omega}^j_1, \cdots, \widehat{\omega}^j_m\}$ as follows:

- For a categorical question q_k with m options, let n_i be the number of participants that reported cat_i s.t. $\widehat{\omega}^j_k = cat_i$, the vector $\boldsymbol{n} = \langle n_1, \cdots, n_m \rangle$ represents the count for s participants, such that $\sum_j n_j = s$.
- For a numerical q_k the report is a number $\widehat{\omega}^j_k = y_i$. A set $Y_k = \{y_1, \cdots, y_s\}$ represents the reports for s participants.

[7] In here we use the ordering of critical questions as presented in Walton et al. (2008) that slightly differs from the simplified presentation we gave in Section 4.

Scheme *LCE*:	- Typically, if C occurs, then E will occur, - In this case, C occurs $\Rightarrow$ In this case E will occur
Rule:	r_{CE} : $rule(R, C, E), occur(C),$ $before(C, E), ruletype(R, causal), noexceptions(R)$ $\Rightarrow occur(E)$

Critical questions and contrary relationships:

CQCE1: Is there evidence for C to occur? $\neg occur(C)$
CQCE2: Is there a general rule for C causing E? $\neg rule(R, C, E)$
CQCE3: Is the relationship between C and E causal? $p_i \in \overline{ruletype(R, causal)}$
CQCE4: Are there any rule exceptions preventing E from occurring? $\neg occur(E)$
CQCE5: Has C happened before E? $p_i \in \overline{before(C, E)}$
CQCE6: Is there any other C' that caused E? $\neg occur(C)$

Scheme *LID*:	- Act_i occurs, and Et_i may be involved - Act_i requires property H - Et_i fits property H $\Rightarrow$ Et_i is associated with Act_i
Rule:	r_{ID} : $occur(Act_i), requiresProp(Act_i, H), entity(Et_i), hasProp(Et_i, H)$ $noexceptions(Et_i, H, Act_i)$ $\Rightarrow association(Et_i, Act_i)$

Critical questions and contrary relationships:

CQID1: Has Act_i happened? $\neg occur(Act_i)$
CQID2: Is H a necessary property for Act_i? $\neg hasProp(Et_i, H)$
CQID3: Does Et_i fit the properties required by Act_i? $\neg requiresProp(Act_i, H)$
CQID4: Are there other entities that fit H? $\neg association(Et_i, Act_i)$
CQID5: Is there an exception to H that undermines the association between Et_i and Act_i? $p_i \in \overline{noexceptions(Et_i, H, Act_i)}$

Fig. A.2. Formal arguments for schemes *LCE* and *LID*.

B1. Analysis of collected results

For categorical data we are interested in knowing the probability distribution π of the categories of a multi-valued answer to question q_k. Since q_k has m possible outcomes, corresponding to the m categories, answers to q_k represent a *discrete* distribution parametrised by the vector $\boldsymbol{\theta} = \langle \theta_1, \ldots, \theta_m \rangle$, where $P(X = j|\boldsymbol{\theta}) = \theta_j$ and $\sum_{j=1}^{m} \theta_j = 1$. Given s the number of participants reporting, $\boldsymbol{X} = \langle X_1, \ldots, X_s \rangle$ is such that $\forall z, X_z \sim$ *discrete*$(\boldsymbol{\theta})$; and $\boldsymbol{n}\big|\boldsymbol{\theta} \sim$ *multinomial*$(\boldsymbol{\theta}, \sum_j n_j)$, such that

$$P(\boldsymbol{n}|\boldsymbol{\theta}) = \frac{s!}{\prod_{j=1}^{m} n_j!} \prod_{j=1}^{m} \theta_j^{n_j}, \text{ with } s = \sum_j n_j \tag{B.1}$$

The vector $\boldsymbol{n}$ is known as a sufficient statistics for $\boldsymbol{\theta}$ because it supplies as much information about $\boldsymbol{\theta}$ as the original vector $\boldsymbol{X}$ does.

From Bayes theorem, $P(\boldsymbol{\theta}|\boldsymbol{n}) \propto P(\boldsymbol{n}|\boldsymbol{\theta}) \cdot P(\boldsymbol{\theta})$. We conveniently choose as prior its conjugate, the Dirichlet distribution parameterised by $\boldsymbol{\alpha} = \langle \alpha_1, \ldots, \alpha_m \rangle$, $\alpha_j > 0$, of the form:

$$dirichlet(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma\left(\sum_{j=1}^{m} \alpha_j\right)}{\prod_{j=1}^{m} \Gamma(\alpha_j)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \tag{B.2}$$

such that

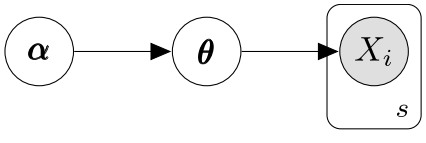


Fig. B.1. Bayesian generative model of the report for a categorical question q_k with m options.

Scheme *LCS*:	- Given that the crowd C was asked q_k and - Answer A is generally accepted as true $\Rightarrow$ Therefore, A may plausibly be true
Rule:	r_{CS} : $request(K, C), crowd(C), assert(C, A),$ $posToKnow(C, K), credible(C, K), evidence\_sup(A)$ $\Rightarrow hold(A)$

Critical questions and contrary relationships:

$CQCS1$: Is the claim A supported by evidence? $p_i \in \overline{evidence\_sup(A)}$
$CQCS2$: Is the group in a position to know about q_k? $p_i \in \overline{posToKnow(C, K)}$
$CQCS3$: Is the claim consistent with others' claims? $\neg hold(A)$
$CQCS4$: Does the group present characteristics appropriate for q_k? $p_i \in \overline{credible(C, K)}$

Fig. B.2. Formal argument for scheme *LCS*.

$$E[X_z] = \frac{\alpha_z}{\sum_j \alpha_j} \tag{B.3}$$

In Fig. B.1, we depict the generative model of the report for a categorical question q_k with m options with a Dirichlet prior, from which we can derive the posterior $P(\boldsymbol{\theta}|\boldsymbol{n}) = dirichlet(\boldsymbol{n} + \boldsymbol{\alpha})$.

As discussed in Toniolo et al. (2015), we considered as prior distribution $dirichlet(C/m, \ldots, C/m)$ with $C = 2$. With this, (B.3) becomes

$$E[X_z] = \frac{n_z + C/m}{C + \sum_j n_j} \tag{B.4}$$

In this paper, independently of the chosen prior, the vector $\boldsymbol{\epsilon}_k = (E[X_1], \ldots, E[X_m])$ refers to the resulting expected values for the m categories of question q_k.

For numerical data, we consider a weighted mean of the s collected reports Y_k for q_k. In the simplest case, weights w_i are assumed to be 1, although these may vary according to features of the reports as for the prior probability. Then, for question q_k we consider the weighted average of answers:

$$\mu_k = \frac{\sum_{i=1}^{s} w_i y_i}{\sum_{i=1}^{s} w_i} \tag{B.5}$$

After the aggregation of responses, the results are introduced in the analysis using an adapted argument scheme from generally accepted opinion of which the formalisation is provided in Fig. B.2.

Appendix C. Provenance

In this section, we explain the formalism underpinning the recording and exploration of provenance of information presented in Section 4.3.

The underpinning language for provenance we use in this research is the W3C standard PROV Model (PROV-DM, Moreau and Missier, 2013). PROV-DM records provenance in terms of *entities, activities*, and *agents* that have caused an entity to be and it defines seven relationships between these elements (Fig. C.1). We refer to those with a prefix *p-*. An entity is a physical or conceptual thing such as a report or a piece of information; an entity may be derived from other entities. An activity represents a process that acts upon entities; e.g., extracting, creating entities. Entities are generated by an activity, and they represent resources that can be consumed (used) by other activities. An activity may inform another activity by triggering it to take place. An agent is something or someone responsible for an activity taking place such as a person, or a software tool. An agent may author an entity or it may act on behalf of other agents.

A record of provenance is formed by nodes (p-entities, p-agents, p-activities) and directed relationships between these nodes. Such a record can be represented as a directed acyclic graph. We may then explore these graphs using OPQL (Lim et al., 2013), a provenance query language that supports lineage queries. Our extension of OPQL for dealing with PROV-DM is presented in Toniolo et al. (2014), here we recall the main elements of this

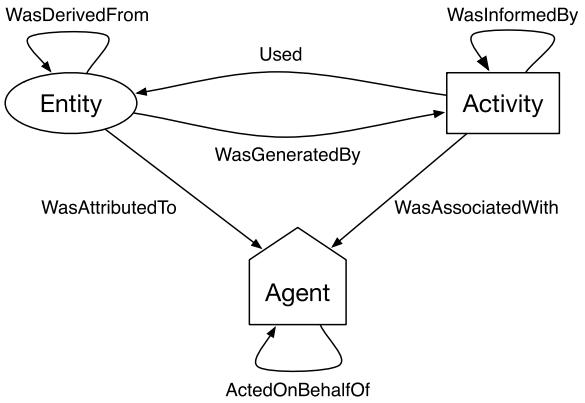


Fig. C.1. The PROV-DM core (Moreau and Missier, 2013).

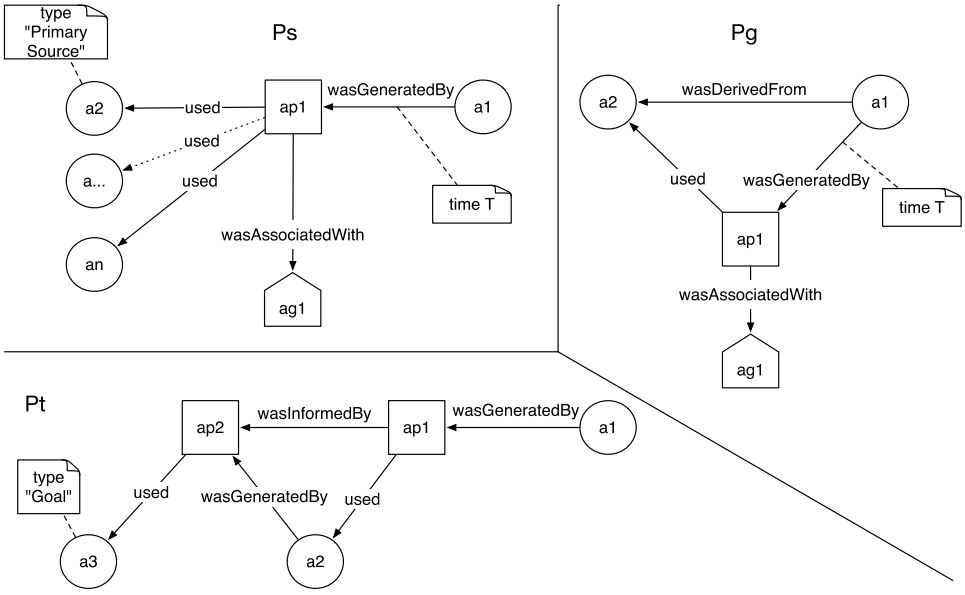


Fig. C.2. Patterns used to query a provenance chain of nodes p_i in CISpaces.

Scheme LPV:
- Given information p_j
- The provenance chain $G_P(p_j)$ of p_j includes pattern P_m of p-entities A_{pv}, p-activities P_{pv}, p-agents Ag_{pv} involved in producing p_j
- P_m is a reason to believe that information p_j is true

$\Rightarrow$ Therefore, p_j may plausibly be true

Rule: $r_{PV} : info(J), prov\_chain(GP, J), pm(PM, GP, El_1, \ldots, El_n), believable(PM, J), no\_missing\_elements(GP, J), evidence\_sup(J) \Rightarrow hold(J)$

Critical questions and contrary relationships:

$CQPV1$: Is p_j consistent with other information? $\neg hold(J)$
$CQPV2$: Is p_j supported by evidence? $p_i \in \overline{evidence\_sup(J)}$
$CQPV3$: Does $G_P(p_j)$ contain p-elements that lead us not to believe p_j? $p_i \in \overline{believable(PM, J)}$
$CQPV4$: Is there any other p-element that should have been included in $G_P(p_j)$ to infer that p_j is true? $p_i \in \overline{no\_missing\_elements(GP, J)}$

Fig. C.3. Formal argument for scheme *LPV*.

formalism.

A Provenance Graph is a graph $G_P = (N, E)$ where a node n can be of type p-entity a (from a set A_{pv}), p-activity ap (from a set P_{pv}), or p-agent ag (from a set Ag_{pv}). The set N is composed by $N = A_{pv} \cup P_{pv} \cup Ag_{pv} = \{n_1, n_2, \cdots\}$. Similarly an edge is labelled with the type of relationship among those defined in Fig. C.1 forming a set $E = E_u \cup E_g \cup E_d \cup E_i \cup E_{aw} \cup E_{at} \cup E_b$ respectively representing: a p-activity *ap used a* (E_u); a p-entity *a was generated by ap* (E_g); a p-entity a_1 *was derived by* a_2 (E_d); a p-activity ap_1 *was informed by* ap_2 (E_i); a p-activity *ap was associated with ag* (E_{aw}); a p-entity *a was attributed to ag* (E_{at}); and a p-agent ag_1 *acted on behalf of* ag_2 (E_b).

Nodes n and edges e comprise a set of attribute-value pairs. Given a set of attributes $Att = \{attribute_1, attribute_2, \cdots\}$ and a set of corresponding values $Val = \{value_1, value_2, \cdots\}$, a mapping function $att : E \cup N \times Att \rightarrow Val$ associates a value to an attribute of an edge or a node. For example, the name Inf_1 of an entity a_1 is $att(a_1, name) = "Inf_1"$ the time associated with a generation edge $e_1 = (a, ap)$ is $att(e_1, time)$ = "2020-01-22:T11-51-00".

In CISpaces, we have two datasets available $\mathscr{I}$ and $\mathscr{P}$. The dataset $\mathscr{I} = \{\cdots\}$ includes pieces of information which corresponds to the information contained on a node p_i created in the WorkBox. $\mathscr{P}$ contains a graph of provenance data for information in $\mathscr{I}$.

Usual operations and properties of a graph apply, in particular, the union of two subgraphs is represented as $G_{P1} \cup G_{P2}$ whereby $G_{P1} = (N_1, E_1)$ and $G_{P2} = (N_2, E_2)$. A directed path is represented as $D_P(n_0, n_k) = (N, E)$ with nodes $N = \{n_0, \cdots, n_k\}$ and edges $E = \{e_0, \cdots, e_{k-1}\}$ such that e_i is an edge directed from n_i to n_{i+1}, for all $i < k$, and a shortest directed path is one where the cardinality of the edge set is the minimum.

Definition 10. (*Provenance chain*) A provenance chain of a node n_j in $\mathscr{P}$ is a subgraph $G_P(n_j) = (N', E')$ of $G_P = (N, E)$ such that:

$$G_P(n_j) = \bigcup_{n_q \in N : \exists D_P(n_j, n_q), \neg \exists n_l (n_q, n_l) \in E} D_P(n_j, n_q)$$

This means that a provenance chain is a graph $G_P(n_j)$ representing a union between all the paths from node n_j in $\mathscr{P}$ to a node $n_q \in N$ that does not have successors. The provenance chain $G_P(p_j)$ indicates a graph $G_P(n_j)$ of an entity node n_j that is linked to information p_j through $att(n_j, name) = p_j$. Henceforth, for convenience we will refer to $G_P(p_j)$ in general discussion, but the formalisation is presented in terms of the correspondent graph $G_P(n_j)$.

Given a provenance graph $G_P(n_j)$, a query to the provenance dataset $\mathscr{P}$ in OPQL is made by using graph patterns and pattern matching.

Definition 11. A *graph pattern* is a pair $P_m = (G_M, C)$, where $G_M = (N_M, E_M)$ is a graph motif and C is a predicate[8] on the attributes of the motif. A *graph motif* G_M is a graph with a certain structure but where nodes and edges are identified by a variable.

A graph pattern $P_m = (G_M, C)$ is *matched* with a graph $G_P = (N, E)$ if there exists an injective mapping $\phi : N_M \rightarrow N$ such that:

i) $\forall e(n_1, n_2) \in E_M$, the mapping $(\phi(n_1), \phi(n_2))$ is an edge in $E \in G_P$
ii) predicate C holds in the mapping of G_M in G_P

The matched graph is a graph identified by $\langle \phi, P_m, G_P \rangle$ and referred to as $\phi_{P_m}[G_P]$.

A graph pattern is a variable that permits the extraction of the structure required by the pattern. A 1-node pattern extracts all nodes that are named with a specific label (e.g.,"Observer"). A 2-node pattern extracts an edge between two nodes. These 1-node or 2-node patterns are used to perform queries in order to extract a named node or a named edge with specific attributes. In CISpaces, we use three composed patterns to record a provenance chain for a piece of information and query it to extract schemes to be included in the analysis:

Extraction of information and updates: A pattern P_g for generating entities takes two entities, a_1 and a_2, whereby a_1 was derived from a_2. Activity ap_1 was responsible for generating entity a_1 using a_2 and it was associated with actor ag_1.
Preparation of a document and primary sources: this is a source pattern P_s where the centre of the provenance record is an activity ap_1 that generates the document recorded in entity a_1 and uses a number of sources $a_2, \cdots, a_n$. An important attribute qualifies an entity as the primary source, where $att(a, type) =$ "Primary Source". Primary sources are those that first reported or created the information.
Intelligence requirement or goal of analysis: this pattern P_t is fundamental for recognising the goal of the analysis. This may also be called an intelligence requirement or a request for information. P_t denotes the triggering activity ap_2 that caused activity ap_1 to be executed. Goals are marked with attribute $C : att(a_3, type) = "Goal"$.

The structure of these three patterns is represented in Fig. C.2. In CISpaces, when the provenance of a node p_j in the WorkBox is inspected, the system queries its provenance chain $G_P(n_j)$ by finding all correspondent matched parts of the graph $\phi_{P_m}[G_P(n_j)]$ for each of these three patterns $P_m \in \{P_g, P_t, P_s\}$. The resulting matched patterns are shown to the analyst who can choose to bring a specific matched pattern of interest in the WorkBox in the form of an instantiated argument scheme. In Fig. C.3 we provide a formalisation of this scheme using predicate-like labels as above.

Appendix D. Study material

In this section, we provide further information on the material of the studies.

D1. Focus group

Below we list the guiding questions of our focus group (Section 2).

1. Could you describe typical day-to-day activities of an analyst?
2. What is the timeline for analysis? How long is the process?
3. What is most useful analytical tools to help analysts to make connections?
4. For new analysts on the job, how is the ground knowledge about a topic formed? How do you get feedback on the quality of work done?
5. What are the techniques to identify new information requirements? What are the criteria to distribute the new queries?
6. What sort of biases can affect the analysis? How would an analyst prevent such biases?
7. Is trustworthiness of information sources important? What are other factors that lead an analyst to consider a hypothesis to be more reliable than others? How is previous analysis used for new tasks?
8. What kind of collaborations would an analyst be involved in? What is making collaboration effective and how do you communicate?
9. How much of the current role is assisted by technology? Where do you see the most significant areas for improvement using technologies that should be possible today?

D2. TAM Questionnaire

Below is a list of closed questions used for the experiment described in Section 5 following our TAM-A model. Closed questions required analysts to respond to a 5-points Likert scale (Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree). Questions were provided to the analysts in a semi-randomised order, shown by the question numbers. We indicate questions that have been reported in the analysis with an inverted scale, meaning that it is the negated indicator that we would expect would provide a positive contribution to the general factor.

Group 1: Experience

14 During training, job related activities, or personal experience, I have previously encountered...
...GEX0: computer mediated analytical tools.

[8] Intuitively, C is similar to the SQL condition "WHERE" in a "SELECT" query.

...GEX1: argument/mind mapping tools.
...GEX2: tools to record provenance.
...GEX3: collaborative analytical tools.
...GEX4: crowdsourcing tools.
13 In my job I regularly use...
...GEX5: computer mediated analytical tools.
...GEX6: argument/mind mapping tools.
...GEX7: tools to record provenance.
...GEX8: collaborative analytical tools.
...GEX9: crowdsourcing tools.

Group 2: Features

27.OQL0: CISpaces may help reduce the time of understanding the events and circumstances.
28.OQL1: CISpaces may improve the robustness of analyses.
29.OQL2: CISpaces may improve the confidence in the accuracy of analyses I produce
30.OQL3: CISpaces may help express my thoughts during analysis
31.OQL4: CISpaces may facilitate better decision-making over plausible hypotheses
22.GRE0: In my job, usage of CISpaces may be important.
18.GRE1: In my job, usage of CISpaces may be relevant.
16.GRE2: The use of CISpaces is pertinent to my various job-related tasks.
10.GRE3: I would have no difficulty telling others about the results of using CISpaces.
8.GRE4: I believe I could communicate to others the advantages of using CISpaces.
3.GRE5: I would have difficulty explaining why using CISpaces may or may not be beneficial. (answers rotated considering "I would have no difficulty...")
19.GPS0: I could complete the job using CISpaces if I had just the built in help facility for assistance.
15.GPS1: I could complete the job using CISpaces if I had some training first.
17.GPS2: Assuming I had resources, opportunities and knowledge it takes to use CISpaces, it would be easy for me to use this system.
21.GPS3: CISpaces is not compatible with other systems I use. (answers rotated considering "CISpaces is compatible...")
20.GPS4: The use of CISpaces would completely change the way I work. (answers rotated considering "...CISpaces would not completely change...")

Group 3: Acceptability

1.PU0: Using CISpaces would facilitate the performance of tasks in my job.
5.PU1: Using CISpaces in my job would increase my productivity.
11.PU2: Using CISpaces would enhance my effectiveness in my job.
7.PU3: I find that CISpaces would be useful in my job.
2.PEOU0: Interaction with CISpaces is clear and understandable.
12.PEOU1: I find that CISpaces would be easy to use.
4.PEOU2: I find that it would be easy to use CISpaces for achieving my goals.
6.BI0 Assuming I had access to CISpaces, I intend to use it.
9.BI1 Assuming I had access to CISpaces, I predict that I would like to use it.

Below is a list of open questions which followed from the previous closed questions. Each question is tagged with the specific factor the question belongs to.

Group 1: Experience

25. GEX: Do you see any similarities between CISpaces and other analytical tools?
26. GEX: What are the strengths and weaknesses of CISpaces in respect to the ACH (Analysis of Competing Hypotheses) tool to perform hypotheses testing?

Group 2: Features

23. GPS: What do you see as a training burden to use CISpaces?
32. OQL: Do you think the output quality criteria listed are relevant criteria?
33. OQL: What is robustness of analysis for you?
34. OQL: Is there any other critical criterion upon which the analysis could or should be assessed?
35. OQL: What parts of CISpaces could address each of the criteria? Why?
36. OQL: To what extent would the structuring of the analysis using con/pro help sharing the reasoning process with other teams? How would CISpaces affect collaboration? [Video at 5'.47"-7'.16"]
37. OQL: To what extent would the crowdsourcing service help to bring new information into the analysis? Assuming that all the permissions to send such requests are fulfilled, would you see it as a useful approach? Do you see any limitations in this approach? [Video at 4'.42"-5'.46"]
38. OQL: To what extent would the recording of provenance inform more robust analysis? Would you see the automatic provenance import as a useful approach? Do you see any limitations in this approach? [Video at 7'.17"-9'.29"]

Group 3: Acceptability

24. BI: What would impede the adoption of the tool?
39. BI: If, from the analysis, you could automatically generate a text report (e.g. in PDF or Word) that summarizes the hypotheses, would you see this as an advantage?
40. BI: What else would you like to see in the tool?
41. BI: What would you see as main applications of CISpaces?

Analysts were given a demonstration video to watch before answering the questions. The video is included in the Supplementary material.

Appendix E. In-depth Study Results

In this section, we provide further information on the results of the study discussed in Section 5.

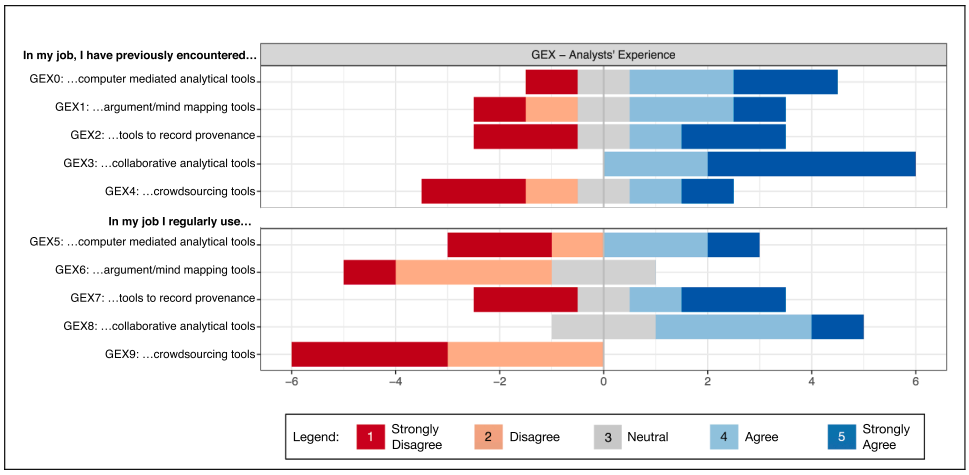


Fig. E.1. Results for Group 1 indicators: Experience.

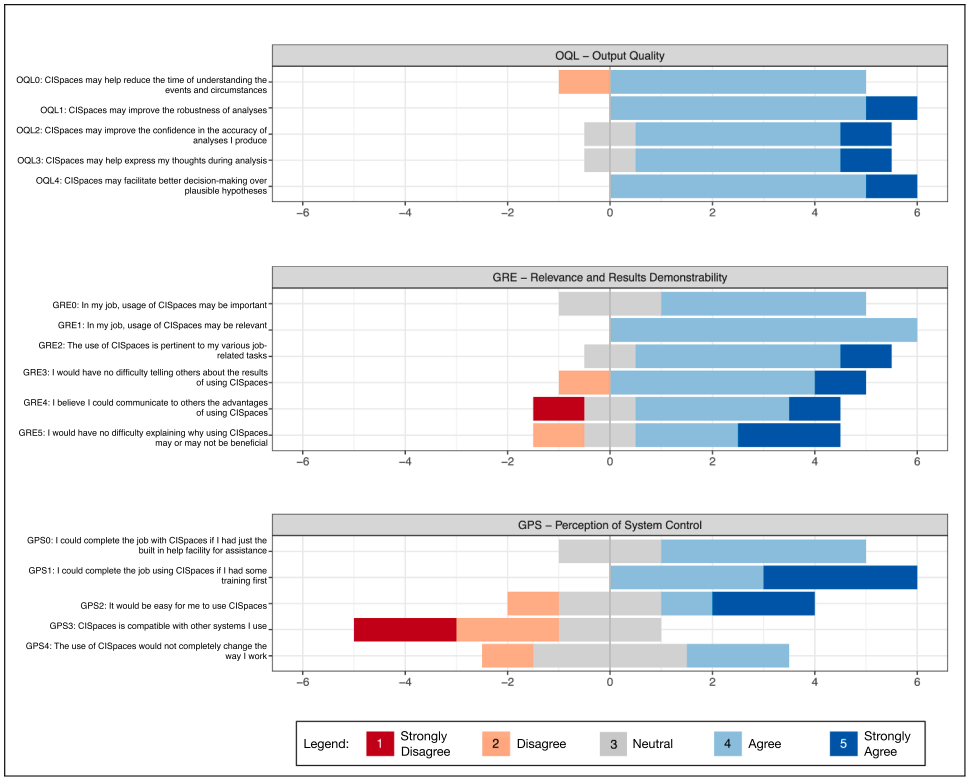


Fig. E.2. Results for Group 2 indicators: Utility.

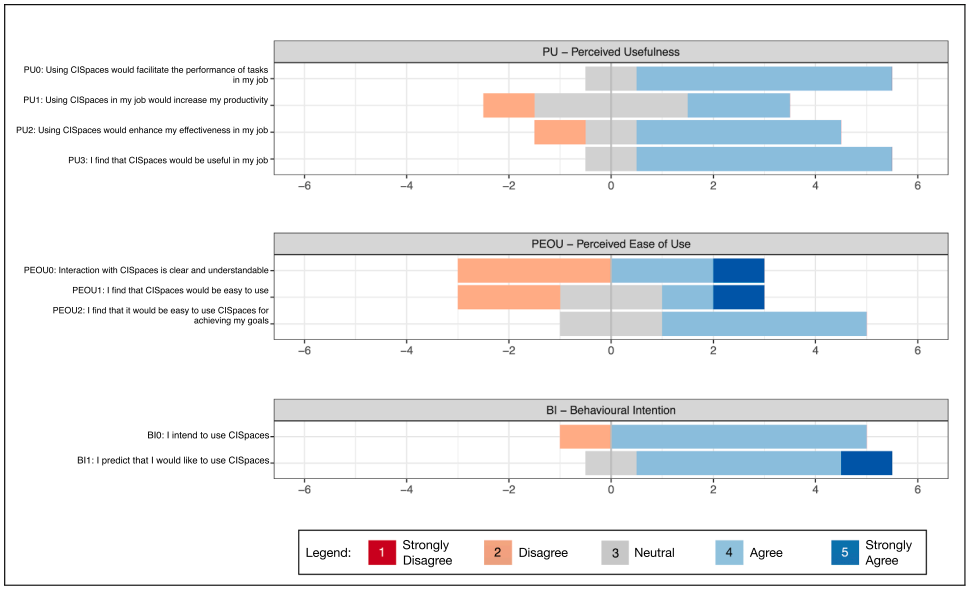


Fig. E.3. Results for Group 3 indicators: Adoption.

E1. TAM-A Detailed results

Results for questions of each group are reported in Figs. E.1, E.2, and E.3, expanding on Fig. 11.

E2. PLS-PM Analysis

PLS-PM runs in two phases, the first to establish the viability of the measurement model and evaluate the correlations between the indicators and their represented factors, and the second to evaluate the structural model, the relationships between factors hypothesised (Chin, 2010; Sanchez, 2013). Factors are measured in a reflective way, which means indicators are consequences of the factor.

We note first that, as discussed in Section 5, the analysis is run on a small number of participants and, therefore, correlations are sensitive to small variations and hard to generalise.

The first phase of PLS-PM focussed on determining how well indicators represent their factors. The results obtained by loadings and cross-loadings are analysed to ensure unidimensionality of the factors, which indicates whether a factor is well represented by its indicators. We have removed indicators with zero variance, as they provide no information on the analysis (GRE1). We have removed indicators too loosely correlated with their factors where loading was nearly zero (GRE0). Three cases where the indicator was highly inversely correlated with its respective factor were rotated (GEX3, GEX8 and GPS1). In this set of results, it is likely that lack of experience with collaborative analytical tools would better represent the level of experience with general tools particularly in current experience (GEX8) and for consistency in previous experience (GEX3). Furthermore, disagreement with the ability to completing the job using CISpaces with due training correlates better with the perception of system control (GPS).

Following the relevant literature, in Table E.1 we report on the values used to assess the measurement model and correlation values for all factors. For unidimensionality of the factors we include: α, Cronbach's Alpha Cronbach (1951), (recommended to be > 0.7); ρ, Dillon Goldstein's rho, (recommended to be > 0.7); 1-ei, 1st eigenvalue (> 1); and 2-ei, 2nd eigenvalue (< 1).

Most of the values indicate homogeneity of indicators according to the recommended values in parenthesis but GPS is the most problematic with low values of α and ρ, and GEX with low values of ρ showing poor internal consistency and, therefore, poor unidimensionality.

The analysis of loadings and cross-loadings in Table E.2 shows that nearly 22/33 indicators have factors loadings above 0.7 (the recommended threshold), and in addition, 23/33 load correctly to their factors, while the others load better to other factors, obtaining around 65% of good representative indicators. In particular, those problematic are part of factors GPS and GEX, and which explain the low values of ρ and α. On the contrary, it is also noticeable that PEOU, PU, BI are well represented by their factors, indicating that it is likely that our factor grouping at the roots (or exogenous factors) may need improvement but the core TAM model is well represented on the other hand.

The second phase of the analysis focuses on constructing and evaluating the structural model, or the strengths and relationship between different factors using multiple regressions. We note that most of the factors present an average variance extracted (AVE) > 0.5 (reported on the diagonals in Fig. E.1) meaning that more than 50% variance of the indicators is accounted for. This is with the exception of GPS and GEX which as noted above are not well represented factors. The overall prediction performance given by the Goodness of fit (GoF) is 0.76 slightly over the recommended value of 0.7.

Table E.1
Unidimensionality measures, correlations of factors with AVE on the diagonal. Cells in italics indicate values outwith the recommended thresholds.

Fac.	α	ρ	1ei	2ei	OQL	GRE	GPS	GEX	PEOU	PU	BI
OQL	0.86	0.91	3.56	*1.05*	**0.63**						
GRE	0.89	0.93	3.10	0.76	0.66	**0.76**					
GPS	*0.21*	*0.55*	2.45	*1.11*	-0.55	-0.72	***0.37***				
GEX	0.85	*0.02*	1.24	0.90	0.78	0.75	-0.41	***0.40***			
PEOU	0.91	0.95	2.56	0.34	0.67	0.84	-0.87	0.77	**0.85**		
PU	0.96	0.97	3.54	0.32	0.42	0.92	-0.51	0.53	0.59	**0.88**	
BI	0.87	0.94	1.77	0.23	0.27	0.85	-0.47	0.45	0.51	0.97	**0.89**

Table E.2
Loadings and cross Cross-loadings for the measurement model where bold values indicate the factors, and ∗ shows when the indicator loads with its factor.

	OQL	GRE	GPS	GEX	PEOU	PU	BI
OQL1	**-0.22**	-0.25	0.16	-0.08	0.02	-0.34	-0.48
OQL2	**0.89***	0.58	-0.38	0.75	0.62	0.34	0.12
OQL3	**0.86***	0.49	-0.71	0.66	0.77	0.13	0.00
OQL4	**0.86***	0.49	-0.71	0.66	0.77	0.13	0.00
OQL5	**0.89***	0.58	-0.38	0.75	0.62	0.34	0.12
GRE0	0.86	**0.49**	-0.71	0.66	0.77	0.13	0.00
GRE2	0.56	**0.96***	-0.55	0.70	0.73	0.94	0.84
GRE3	0.60	**0.99***	-0.74	0.72	0.86	0.92	0.85
GRE4	0.70	**0.95***	-0.76	0.71	0.81	0.86	0.83
GPS0	0.00	-0.28	**0.56***	0.30	-0.14	-0.38	-0.47
GPS1	-0.67	-0.70	**0.82***	-0.28	-0.59	-0.61	-0.53
GPS2	0.15	0.22	**0.25**	0.42	0.16	0.21	0.04
GPS3	0.10	0.58	**-0.82**	0.29	0.77	0.44	0.49
GPS4	0.60	-0.06	**-0.34**	0.30	0.34	-0.41	-0.52
GEX0	0.37	0.25	0.41	**0.63***	0.01	0.28	0.23
GEX1	0.62	0.52	0.11	**0.75***	0.23	0.51	0.43
GEX2	0.65	-0.04	-0.07	**0.46**	0.18	-0.33	-0.37
GEX3	0.00	0.28	-0.19	**0.35**	0.26	0.27	0.47
GEX4	0.50	-0.06	0.10	**0.51***	0.22	-0.35	-0.51
GEX5	0.42	0.56	0.04	**0.80***	0.43	0.50	0.40
GEX6	0.49	0.31	0.04	**0.85***	0.42	0.11	0.06
GEX7	0.55	-0.07	-0.05	**0.53**	0.27	-0.40	-0.45
GEX8	0.49	0.84	-0.50	**0.86***	0.84	0.71	0.65
GEX9	-0.67	-0.70	0.82	**-0.28**	-0.59	-0.61	-0.53
PEOU0	0.55	0.67	-0.74	0.63	**0.91***	0.40	0.26
PEOU1	0.78	0.81	-0.71	0.91	**0.95***	0.53	0.42
PEOU2	0.53	0.83	-0.93	0.58	**0.90***	0.65	0.66
PU0	0.22	0.87	-0.47	0.47	0.57	**0.96***	0.96
PU1	0.73	0.92	-0.55	0.71	0.63	**0.89**	0.84
PU2	0.33	0.79	-0.42	0.28	0.39	**0.94***	0.87
PU3	0.22	0.87	-0.47	0.47	0.57	**0.96***	0.96
BI0	0.22	0.87	-0.47	0.47	0.57	0.96	**0.96***
BI1	0.29	0.72	-0.41	0.35	0.35	0.84	**0.93***

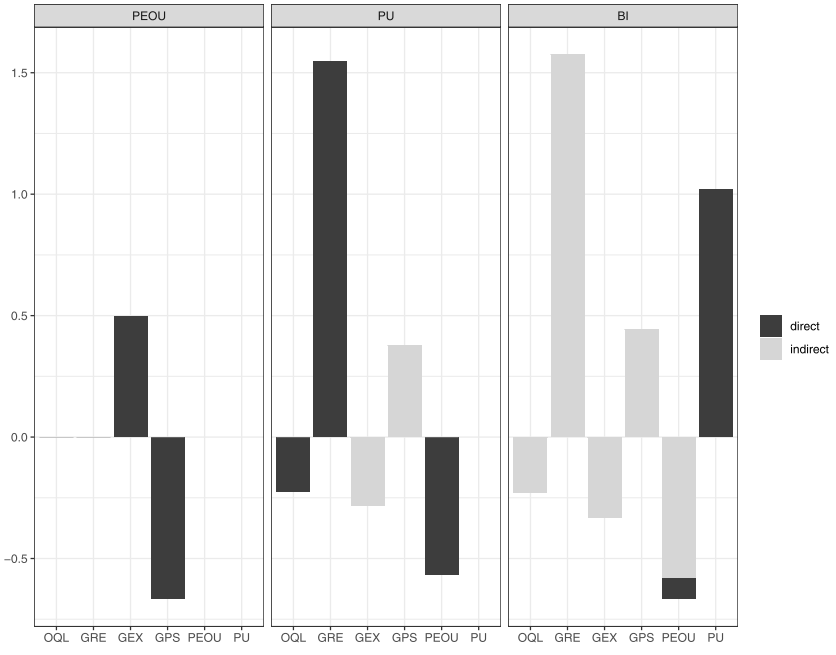


Fig. E.4. Direct and Indirect effects on the core TAM factors BI, PEOU, PU.

As above the results we can draw are limited, and validation through bootstrapping cannot be achieved due to low numbers of participants and limited variances in some of the indicators, here our discussion is limited to consider what factors may contribute to or hinder the intention of using CISpaces.

For this, we can consider the strength of the relationships, which are also presented in Section 5. The structural model resulting from the partial least squares analysis is shown in Fig. 13, reporting the regression weights, and the coefficients of determination of the factors, R^2. R^2 indicates the proportion of the variation in one factor that is dependent on the variation of the other factor. The effects between all variables but PEOU to BI are statistically significant at $p < 0.05$, and high values of R^2 indicate that most of the variance in PU, PEOU, BI can be explained by their independent

factors. In particular, we obtain the following relationships:

- H2.1: OQL→PU$^{\downarrow}$, OQL has a negative effect on PU ($-0.224, p = 0.020$)
- H2.2: GRE→PU$^{\uparrow}$, GRE has a positive effect on PU ($1.547, p = 0.001$)
- H2.3: GPS→PEOU$^{\downarrow}$, GPS has a negative effect on PEOU ($-0.665, p = 0.014$)
- H2.4: GEX→PEOU$^{\uparrow}$, GEX has a positive effect on PEOU ($0.497, p = 0.030$)
- H2.5: PEOU→PU$^{\downarrow}$, PEOU has a negative effect on PU ($-0.567, p = 0.006$)
- H2.6: PU→BI$^{\uparrow}$, PU has a positive effect on BI ($1.018, p = 0.011$)
- H2.7: there is insufficient evidence for any effect between PEOU and BI ($-0.090, p = 0.467$)

As mentioned above, these significance values cannot be validated through bootstrapping due to the limited number of participants. We also explore the indirect effects of multiple paths on the factors as we are interested in what contributes most to the results obtained. Fig. E.4 shows these contributions.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.iswa.2022.200151.

References

Baber, C., Attfield, S., Conway, G., Rooney, C., & Kodagoda, N. (2016). Collaborative sense-making during simulated intelligence analysis exercises. *International Journal of Human-Computer Studies, 86*, 94–108. https://doi.org/10.1016/j.ijhcs.2015.10.001

Bex, F., Prakken, H., Reed, C. A., & Walton, D. (2003). Towards a formal account of reasoning about evidence: Argumentation schemes and generalisations. *Artificial Intelligence and Law, 11*(2–3), 125–165. https://doi.org/10.1023/B:ARTI.0000046007.11806.9a

Bex, F., & Verheij, B. (2012). Solving a murder case by asking critical questions: An approach to fact-finding in terms of argumentation and story schemes. *Argumentation, 26*(3), 325–353. https://doi.org/10.1007/s10503-011-9257-0

Bier, E., Card, S., & Bodnar, J. (2008). Entity-based collaboration tools for intelligence analysis. *Proceedings of the IEEE symposium on visual analytics science and technology*. https://doi.org/10.1109/VAST.2008.4677362

Billman, D., Convertino, G., Shrager, J., Pirolli, P., & Massar, J. (2006). Collaborative intelligence analysis with CACHE and its effects on information gathering and cognitive bias. *Proceedings of the human computer interaction consortium workshop*.

Brabham, D. C. (2008). Crowdsourcing as a model for problem solving an introduction and cases. *Convergence, 14*(1), 75–90. https://doi.org/10.1177/1354856507084420

Burke, J. A., Estrin, D., Hansen, M., Parker, A., Ramanathan, N., Reddy, S., & Srivastava, M. B. (2006). Participatory sensing. *Proceedings of the ACM SenSys world sensor web workshop*.

Burton, M., & Knowles, J. (2010). Open source ACH. https://www.github.com/Burton/Analysis-of-Competing-Hypotheses, previously competinghypotheses.org [Last Accessed 2022].

Caminada, M., & Wu, Y. (2011). On the limitations of abstract argumentation. *Proceedings of the 23rd benelux conference on artificial intelligence*.

Carneiro, G., Toniolo, A., Ncenta, M. A., & Quigley, A. J. (2021). Text vs. graphs in argument analysis. *2021 IEEE symposium on visual languages and human-centric computing (VL/HCC)* (pp. 1–9). https://doi.org/10.1109/VL/HCC51201.2021.9576493

Cerutti, F., Norman, T. J., & Toniolo, A. (2018a). A tool to highlight weaknesses and strengthen cases: CISpaces.org. *Proceedings of the 31st annual conference on legal knowledge and information systems* (pp. 186–189). https://doi.org/10.3233/978-1-61499-935-5-186

Cerutti, F., Norman, T. J., Toniolo, A., & Middleton, S. E. (2018b). CISpaces.org: From fact extraction to report generation, . In *Frontiers in Artificial Intelligence and Applications: vol. 305. Computational models of argument* (pp. 269–280). IOS Press. https://doi.org/10.3233/978-1-61499-906-5-269

Cerutti, F., & Pearson, G. (2018). Supporting scientific enquiry with uncertain sources. *Proceedings of the 21st international conference on information fusion* (pp. 1–8). https://doi.org/10.23919/ICIF.2018.8455649

Cerutti, F., Tintarev, N., & Oren, N. (2014). Formal arguments, preferences, and natural language interfaces to humans: An empirical evaluation. *Proceedings of the 21st European conference on artificial intelligence* (pp. 207–212). IOS Press. https://doi.org/10.3233/978-1-61499-419-0-207

Cerutti, F., Toniolo, A., & Norman, T. J. (2019). On natural language generation of formal argumentation, . In *CEUR Workshop Proceedings: vol. 2528. Proceedings of the 3rd workshop on advances in argumentation in artificial intelligence* (pp. 15–29).

Cerutti, F., Toniolo, A., Norman, T. J., Rahwan, I., & Reed, C. (2018c). AIF-EL - an OWL2-EL-compliant AIF ontology. *Computational models of argument* (pp. 455–456). https://doi.org/10.3233/978-1-61499-906-5-455

Cerutti, F., Vallati, M., & Giacomin, M. (2016). An efficient java-based solver for abstract argumentation frameworks: jArgSemSAT. *International Journal on Artificial Intelligence Tools, 26*(2). https://doi.org/10.1142/S0218213017500026

Chilcot, J. (2016). The report of the Iraq inquiry, executive summary. https://www.webarchive.nationalarchives.gov.uk/20171123122743/http://www.iraqinquiry.org.uk/the-report/.

Chin, W. W. (2010). How to write up and report PLS analyses, . In *Handbooks of Computational StatisticsHandbook of partial least squares.* (pp. 655–690). Springer. https://doi.org/10.1007/978-3-540-32827-8_29

Chorley, A., Edwards, P., Hielkema, F., Philip, L., & Farrington, J. (2008). Supporting provenance and argumentation in evidence-based policy assessment. *Proceedings of the oxford e-research conference*.

Cramer, M., & Guillaume, M. (2019). Empirical study on human evaluation of complex argumentation frameworks. In F. Calimeri, N. Leone, & M. Manna (Eds.), *Lecture notes in computer science: Logics in artificial intelligence* (pp. 102–115). Springer International Publishing. https://doi.org/10.1007/978-3-030-19570-0_7.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334. https://doi.org/10.1007/BF02310555

Čyras, K., & Toni, F. (2016). ABA+: Assumption-based argumentation with preferences. *Proceedings of the 15th international conference on principles of knowledge representation and reasoning* (pp. 553–556).

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

De Liddo, A., Sándor, A., & Buckingham-Shum, S. (2012). Contested collective intelligence: Rationale, technologies, and a human-machine annotation study. *Computer Supported Cooperative Work, 21*, 417–448. https://doi.org/10.1007/s10606-011-9155-x

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence, 77*(2), 321–357. https://doi.org/10.1016/0004-3702(94)00041-X

Dung, P. M., Kowalski, R. A., & Toni, F. (2009). Assumption-based argumentation. In G. R. Simari, & I. Rahwan (Eds.), *Argumentation in artificial intelligence* (pp. 199–218). Springer US. https://doi.org/10.1007/978-0-387-98197-0_10.

Etuk, A., Norman, T. J., Şensoy, M., Bisdikian, C., & Srivatsa, M. (2013). TIDY: A trust-based approach to information fusion through diversity. *Proceedings of the 16th international conference on information fusion* (pp. 1188–1195).

García, A. J., & Simari, G. R. (2004). Defeasible logic programming: An argumentative approach. *Theory and Practice of Logic Programming, 4*(1–2), 95–138. https://doi.org/10.1017/S1471068403001674

Gil, Y., & Ratnakar, V. (2002). TRELLIS: An interactive tool for capturing information analysis and decision making. In A. Gómez-Pérez, & V. R. Benjamins (Eds.), *Lecture Notes in Computer Science: vol. 2473. Knowledge engineering and knowledge management: Ontologies and the semantic web* (pp. 37–42). Springer. https://doi.org/10.1007/3-540-45810-7_6.

Guest, G., Bunce, A., & Johnson, L. (2006). How many interviews are enough? An experiment with data saturation and variability. *Field methods, 18*(1), 59–82. https://doi.org/10.1177/1525822X05279903

Hartig, O., & Zhao, J. (2009). Using web data provenance for quality assessment. *Proceedings of the 1st international workshop on the role of semantic web in provenance management*.

Heuer, R. J. (1999). *Psychology of intelligence analysis*. US Government Printing Office.

Hossain, M. S., Andrews, C., Ramakrishnan, N., & North, C. (2011). Helping intelligence analysts make connections. *AAAI workshop on scalable integration of analytics and visualization* (pp. 22–31).

IARPA (2017). CREATE program: Crowdsourcing evidence, argumentation, thinking and evaluation. https://www.iarpa.gov/research-programs/create[Last Accessed 2022].

IBM (2017). i2 Analyst's Notebook. https://www.ibm.com/downloads/cas/QNGO6RNA Last available as part of the i2 Intelligence Analysis Portfolio, release 9.2.2 https://www.ibm.com/docs/en/i2-iap/9.2.2 [Last Accessed 2022].

Jøsang, A., & Haller, J. (2007). Dirichlet reputation systems. *Proceedings of the 2nd IEEE international conference on availability, reliability and security.* (pp. 112–119). https://doi.org/10.1109/ARES.2007.71

Kamar, E., Hacker, S., & Horvitz, E. (2012). Combining human and machine intelligence in large-scale crowdsourcing. *Proceedings of the 11th international conference on autonomous agents and multiagent systems* (pp. 467–474).

Kang, K., & Sinnott, R. O. (2018). Improving online argumentation through deep learning. In O. Gervasi, B. Murgante, S. Misra, E. Stankova, C. M. Torre, A. M. A. Rocha, D. Taniar, B. O. Apduhan, E. Tarantino, & Y. Ryu (Eds.), *Lecture Notes in Computer ScienceComputational science and its applications* (pp. 376–391). Springer. https://doi.org/10.1007/978-3-319-95162-1_26.

Kang, Y., & Stasko, J. (2011). Characterizing the intelligence analysis process: Informing visual analytics design through a longitudinal field study. *Proceedings of the IEEE conference on visual analytics science and technology* (pp. 21–30). https://doi.org/10.1177/1473871612468877

Klein, G., Moon, B., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems, 21*(5), 88–92. https://doi.org/10.1109/MIS.2006.100

Lahneman, W. J., & Arcos, R. (2014). *The art of intelligence: Simulations, exercises, and games*. Rowman & Littlefield.

Lawrence, J., & Reed, C. (2020). Argument mining: A survey. *Computational Linguistics, 45*(4), 765–818. https://doi.org/10.1162/coli_a_00364

Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. *Information & Management, 40*(3), 191–204. https://doi.org/10.1016/S0378-7206(01)00143-4

Leiva, M. A., Simari, G. I., Gottifredi, S., García, A. J., & Simari, G. R. (2019). DAQAP: Defeasible argumentation query answering platform. In A. Cuzzocrea, S. Greco, H. L. Larsen, D. Saccà, T. Andreasen, & H. Christiansen (Eds.), *Flexible query answering systems* (pp. 126–138). Springer International Publishing. https://doi.org/10.1007/978-3-030-27629-4_14.

Lim, C., Lu, S., Chebotko, A., Fotouhi, F., & Kashlev, A. (2013). OPQL: Querying scientific workflow provenance at the graph level. *Data & Knowledge Engineering, 88*, 37–59. https://doi.org/10.1016/j.datak.2013.08.008

Llinas, J. (2013). Challenges in information fusion technology capabilities for modern intelligence and security problems. *Proceedings of the IEEE european conference on intelligence and security informatics conference* (pp. 89–95).

Lohmöller, J.-B. (1989). *Latent variable path modeling with partial least squares*. Springer Science & Business Media.

Lu, D., Voss, C. R., Tao, F., Ren, X., Guan, R., Korolov, R., Zhang, T., Wang, D., Li, H., Cassidy, T., et al. (2016). Cross-media event extraction and recommendation. *Proceedings of the conference of the north american chapter of the association for computational linguistics* (pp. 72–76). https://doi.org/10.18653/v1/N16-3015

Mahyar, N., & Tory, M. (2014). Supporting communication and coordination in collaborative sensemaking. *IEEE Transactions on Visualization and Computer Graphics, 20*(12), 1633–1642. https://doi.org/10.1109/TVCG.2014.2346573

Miles, M. B., Huberman, A. M., & Saldana, J. (2013). *Qualitative data analysis*. Sage publications.

Modgil, S., & Prakken, H. (2014). The ASPIC+ framework for structured argumentation: A tutorial. *Argument & Computation, 5*(1), 31–62. https://doi.org/10.1080/19462166.2013.869766

Oracle (1996). Java. https://www.java.com/ [Last Accessed: 2022].

Ecma International (2017). Json - JavaScript Object Notation. Industry association for standardizing information and communication systems. https://www.ecma-international.org/publications-and-standards/standards/ecma-404/ [Last Accessed: 2022].

Moreau, L., & Missier, P. (2013). PROV-DM: The PROV data model. http://www.w3.org/TR/prov-dm/ [Last Accessed 2022].

Ouyang, R. W., Kaplan, L. M., Toniolo, A., Srivastava, M., & Norman, T. J. (2016a). Aggregating crowdsourced quantitative claims: Additive and multiplicative models. *IEEE Transactions on Knowledge and Data Engineering*, (99)https://doi.org/10.1109/TKDE.2016.2535383

Ouyang, R. W., Srivastava, M., Toniolo, A., & Norman, T. J. (2016b). Truth discovery in crowdsourced detection of spatial events. *IEEE Transactions on Knowledge and Data Engineering, 28*(4), 1047–1060. https://doi.org/10.1109/TKDE.2015.2504928

Paredes, J. N., Simari, G. I., Martinez, M. V., & Falappa, M. A. (2021). Detecting malicious behavior in social platforms via hybrid knowledge- and data-driven systems. *Future Generation Computer Systems, 125*, 232–246. https://doi.org/10.1016/j.future.2021.06.033

Park, S. Y. (2009). An analysis of the technology acceptance model in understanding university students' behavioral intention to use e-learning. *Journal of Educational Technology & Society, 12*(3), 150–162.

Parsons, S., Tang, Y., Sklar, E., McBurney, P., & Cai, K. (2011). Argumentation-based reasoning in agents with varying degrees of trust. *Proceedings of the 10th international conference on autonomous agents and multiagent systems* (pp. 879–886).https://www.dl.acm.org/doi/abs/10.5555/2031678.2031743

Pioch, N. J., & Everett, J. O. (2006). POLESTAR: Collaborative knowledge management and sensemaking tools for intelligence analysts. *Proceedings of the 15th international conference on information and knowledge management* (pp. 513–521). https://doi.org/10.1145/1183614.1183688

Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data quality assessment. *Communications of the ACM, 45*(4), 211–218. https://doi.org/10.1145/505248.506010

Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of the international conference on intelligence analysis*.

Prakken, H. (2010). An abstract framework for argumentation with structured arguments. *Argument and Computation, 1*(2), 93–124. https://doi.org/10.1080/19462160903564592

PROV Working Group (2013). PROV-O: The PROV ontology. https://www.w3.org/TR/prov-o/ [Last Accessed 2022].

Prunckun, H. (2010). *Handbook of scientific methods of inquiry for intelligence analysis*. Scarecrow Press.

QSR International (1999). NVivo (qualitative data analysis software). Version 12. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/[Last Accessed: 2022].

Reed, C., Budzynska, K., Duthie, R., Janier, M., Konat, B., Lawrence, J., Pease, A., & Snaith, M. (2017). The argument web: An online ecosystem of tools, systems and services for argumentation. *Philosophy & Technology, 30*(2), 137–160. https://doi.org/10.1007/s13347-017-0260-8

Reed, C., & Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools, 13*(4), 961–979. https://doi.org/10.1142/S0218213004001922

Robinson, T., & Pardoe, L. (2021). Value based collection in intelligence analysis. *2021 International conference on military communication and information systems (ICMCIS)* (pp. 1–6). https://doi.org/10.1109/ICMCIS52405.2021.9486414

Rooney, C., Attfield, S., Wong, B. L. W., & Choudhury, S. (2014). INVISQUE as a tool for intelligence analysis: The construction of explanatory narratives. *International Journal of Human-Computer Interaction, 30*(9), 703–717. https://doi.org/10.1080/10447318.2014.905422

Saletta, M., Kruger, A., Primoratz, T., Barnett, A., van Gelder, T., & Horn, R. E. (2020). The role of narrative in collaborative reasoning and intelligence analysis: A case study. *PLOS ONE, 15*(1), 1–17. https://doi.org/10.1371/journal.pone.0226981

Sanchez, G. (2013). PLS path modelling with R. https://www.gastonsanchez.com/PLS_Path_Modeling_with_R.pdf[Last Accessed:2022].

Sanchez, G., Trinchera, L., & Russolillo, G. (2015). plspm package: Tools for partial least squares path modeling (PLS-PM). Version 0.4.9. https://www.github.com/gastonstat/plspm [Last Accessed:2022].

Schrag, R., McIntyre, J., Richey, M., Laskey, K. B., Wright, E., Kerr, R., Johnson, R., Ware, B., & Hoffman, R. (2016). Probabilistic argument maps for intelligence analysis: Completed capabilities. *Proceeding of the international workshop on computational models of natural arguments* (pp. 34–39).

Sinnott, R., Bayliss, C., Guest, C., G. Jayaputera, G. K., Kim, J., Pan, Y., Susanto, R., Vu, D., Widjaja, I., Zhao, Z., de Rozario, R., Silver, E., Thomman, S., van Gelder, T., Ades, Y., Dwyer, T., Marriott, K., & Schwarz, M. (2019). The design and development of a cloud-based platform supporting team-oriented evidence-based reasoning: SWARM systems paper. *Proceedings of the 52nd hawaii international conference on system sciences*. https://doi.org/10.24251/HICSS.2019.050

Stasko, J., Görg, C., & Liu, Z. (2008). Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization, 7*(2), 118–132. https://doi.org/10.1109/VAST.2007.4389006

Stefik, M. J. (2014). Xerox PARC ACH tool. Palo Alto Research Center Incorporated. https://www.markstefik.com/?page_id=702, previously https://www2.parc.com/istl/projects/ach/ach.html [Last Accessed 2022].

Stottlemyre, S. A. (2015). HUMINT, OSINT, or something new? Defining crowdsourced intelligence. *International Journal of Intelligence and Counter Intelligence, 28*(3), 578–589. https://doi.org/10.1080/08850607.2015.992760

Tecuci, G., Schum, D., Boicu, M., Marcu, D., & Hamilton, B. (2010). Intelligence analysis as agent-assisted discovery of evidence, hypotheses and arguments. *Advances in intelligent decision technologies* (pp. 1–10). Springer. https://doi.org/10.1007/978-3-642-14616-9_1

The Apache Software Foundation (2002). The Apache Tomcat Project. https://www.tomcat.apache.org [Last Accessed: 2022].

The Apache Software Foundation (2010). Apache Jena. https://www.jena.apache.org [Last Accessed: 2022].

The Kivy Community (2011). Kivy: Cross-platform python framework for nui development. https://www.kivy.org/ [Last Accessed: 2022].

The R Foundation (2004). The R Project for statistical computing. Version 4.1.2. https://www.r-project.org [Last Accessed: 2022].

The ZeroMQ Community (2007). ZeroMQ - an open-source universal messaging library. https://www.zeromq.org [Last Accessed: 2022].

Toniolo, A., Braines, D., Preece, A. D., Webberley, W., Norman, T. J., Sullivan, P., & Dropps, T. (2016). Conversational intelligence analysis. *Proceedings of the 1st international workshop on understanding situations through multimodal sensing*. https://doi.org/10.1145/2833312.2849568

Toniolo, A., Cerutti, F., Oren, N., Norman, T. J., & Sycara, K. (2014). Making informed decisions with provenance and argumentation schemes. *Proceedings of the 11th international workshop on argumentation in multi-agent systems*.

Toniolo, A., Norman, T., & Oren, N. (2018). Enumerating preferred extensions: A case study of human reasoning, . In *Lecture Notes in Computer ScienceTheory and applications of formal argumentation - 4th international workshop, revised selected papers* (pp. 192–210). Springer-Verlag. https://doi.org/10.1007/978-3-319-75553-3_14

Toniolo, A., Norman, T. J., Etuk, A., Cerutti, F., Ouyang, R. W., Srivastava, M., Oren, N., Dropps, T., Allen, J. A., & Sullivan, P. (2015). Supporting reasoning with different types of evidence in intelligence analysis. *Proceedings of the 14th international conference on autonomous agents and multiagent systems* (pp. 781–789).

United Nations (2011). Criminal intelligence: Manual for analysts. www.unodc.org/documents/organized-crime/Law-Enforcement/Criminal-Intelligence-for-Analysts.pdf [Last Accessed 2022].

US Army (2006). Field Manual 2–22.3: Human Intelligence Collector Operations. https://www.armypubs.army.mil/ProductMaps/PubForm/Details.aspx?PUB_ID=82535.

US Army (2020). Army techniques publication TC 2–33.4. https://www.armypubs.army.mil/ProductMaps/PubForm/Details.aspx?PUB_ID=1008410.

van Gelder, T. (2007). The rationale for RationaleTM. *Law, Probability and Risk, 6*(1–4), 23–42. https://doi.org/10.1093/lpr/mgm032

van Gelder, T., Kruger, A., Thomman, S., de Rozario, R., Silver, E., Saletta, M., Barnett, A., Sinnott, R. O., Jayaputera, G. T., & Burgman, M. (2020). Improving analytic reasoning via crowdsourcing and structured analytic techniques. *Journal of*

Cognitive Engineering and Decision Making, 14(3), 195–217. https://doi.org/10.1177/1555343420926287

Venkatesh, V., & Bala, H. (2008). Technology Acceptance Model 3 and a research agenda on interventions. *Decision Sciences, 39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Network science for military coalition operations. (2010). In Verma. D. (Ed.), (2010). IGI Global.

Visual Analytics Community (2006). Visual analytics science and technology (VAST) challenge. http://www.vacommunity.org/About+the+VAST+Challenge[Last Accessed 2022].

Vyvyan, D., Dantressangle, P., & Bent, G. (2015). The Gaian database. https://www.github.com/gaiandb/gaiandb [Last Accessed: 2022].

Walton, D., Reed, C. A., & Macagno, F. (2008). *Argumentation schemes*. Cambridge University Press.

Waltz, E. (2003). *Knowledge management in the intelligence enterprise*. Artech House.

Whitehill, J., Ruvolo, P., Wu, T., Bergsma, J., & Movellan, J. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Proceedings of the advances in neural information processing systems* (pp. 2035–2043).

Wright, W., Schroh, D., Proulx, P., Skaburskis, A., & Cort, B. (2006). The sandbox for analysis: Concepts and methods. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 801–810). https://doi.org/10.1145/1124772.1124890

Wu, A., Convertino, G., Ganoe, C., Carroll, J. M., & Zhang, X. (2013). Supporting collaborative sense-making in emergency management through geo-visualization. *International Journal of Human-Computer Studies, 71*(1), 4–23. https://doi.org/10.1016/j.ijhcs.2012.07.007

Wu, J. H., & Wang, S. C. (2005). What drives mobile commerce?: An empirical evaluation of the revised technology acceptance model. *Information & Management, 42*(5), 719–729. https://doi.org/10.1016/j.im.2004.07.001

Zhang, J., & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive Science, 18*(1), 87–122. https://doi.org/10.1207/s15516709cog1801_3

Zook, M., Graham, M., Shelton, T., & Gorman, S. (2012). Volunteered geographic information and crowdsourcing disaster relief: A case study of the Haitian earthquake. *World Medical & Health Policy, 2*. https://doi.org/10.2202/1948-4682.1069