Investigations within investigations: a recursive framework for scalable sensemaking support

Simon Attfield
UCL Interaction Centre*
s.attfield@cs.ucl.ac.uk

Ann Blandford
UCL Interaction Centre*
a.blandford@cs.ucl.ac.uk

Stephen De Gabrielle
UCL Interaction Centre*
s.degabrielle@ucl.ac.uk

* University College London Interaction Centre, Remax House, 31/32 Alfred Place, London, WC1E 7DP

ABSTRACT

Using a case-study of a fraud investigation we show that a major challenge in large-scale, collaborative sensemaking arises from the simultaneous decomposition and integration of elements of the task. Given the potential for sensemaking systems to support users in addressing this challenge, we analyse decomposition and integration within the case-study and provide a recursive framework of entities associated with a line of enquiry at any level of description. The framework includes: *theories*, *questions*, *information seeking strategies*, *evidence and evidence collections*, *knowledge*, *assigned investigators* and *lower-level lines of enquiry*. We develop the case for how systems built around such a framework can help address the decomposition/integration challenge.

Author Keywords

Sensemaking, Legal Investigations, Collaboration

ACM Classification Keywords

INTRODUCTION

Lawyers involved in corporate litigations and investigations face an immense sensemaking challenge. At the centre of any corporate litigation or investigation is an evidence-base of documents which is typically vast. Electronic discovery requests for email alone can result in thousands to millions and even tens of millions of documents [1]. Once documents have been obtained a process begins during which lawyers, frequently working in teams, search, review, and re-represent information in order to make sense of facts relevant to a case.

The size of such collections is a result of the rapid increase in the volume of electronically stored information within modern enterprises. But whilst technological advances have created this challenge, they also offer the opportunity for addressing it. The development of technologies to assist in investigative sensemaking, however, must be predicated upon an understanding of the sensemaking processes of the people who perform such investigations. Some work has already been done in the area of researching and modelling investigative sensemaking and considering implications for design [e.g. 4, 2]. Our own work has been to engage explicitly with problems associated with the tractability of large investigations in team-based settings.

We have conducted a case-study of a large fraud investigation based on 10 interviews with 9 members of the team of investigating lawyers. The investigators used a collection of documentary evidence equivalent in size to around 8.5 million novels. This, in addition to witness interviews, provided source material for constructing integrated representations of facts relevant to the case. Our interviews sought to understand the ways in which the investigators individually and collectively worked with this information to support interpretation and decision making.

From our study we have identified two major challenges to the tractability and effectiveness of such large investigations. These are:

- Decomposition: To enable the distribution of labour and specialisation, investigators need to decompose an investigation into multiple, meaningful and tractable chunks of enquiry. Decomposition, however, is gradual and relies upon knowledge uncovered through investigation work;
- Integration: The significance of information revealed by any low-level line of enquiry is frequently only apparent in the light of information within others. Hence low-level lines of enquiry must continually be viewed in terms of a ‘bigger picture’;

These challenges reflect the paradox of the hermeneutic circle—that interpreting the whole depends on the interpretation of the parts, and yet interpreting the parts depends on an understanding of the whole [5]. Within large investigations this presents a particular challenge where investigating is distributed across multiple members of an investigation team.

Investigators need to establish effective ways of managing decomposition and coordinated integration, and the systems that they use can and should play an important role in this. Systems designed to support sensemaking, whether this be searching, filtering, extracting, constructing schematic representations, presenting a story [4], or integrated combinations of these, need to reflect the way that users naturally structure their problems. They need to support users in making sense of parts, and in making sense of the whole.

This has led us to explore the way that the investigators in our study structurally decomposed their problem. This approach contrasts with and complements other models of sensemaking which have tended to focus on describing process [e.g. 4, 2, 3].

The result is a framework that can be used to model decomposition and which lends itself to the design of interactive systems for supporting the challenges of decomposition and integration. We are currently using the framework as the core for an integrated system prototype for supporting large scale, collaborative investigation work.

In this paper we present the framework and its rationale. In the next section we give some background to the investigation we studied. We then provide examples drawn from our data which illustrate some structural aspects and describe the framework.

THE INVESTIGATION

The investigation was carried out over three months by a team of around 30 lawyers within a large corporate law firm. The objective was to discover whether fraud had occurred within a company, and if so, who had been complicit. This objective was principally focused by allegations relating to specific business activities.

A major source of information from which this determination was to be made was the collection of documents (document universe) recovered using data-forensics. Other sources included telephone records and interviews with witnesses and suspects.

The objective of the investigation was to assess the truth behind the allegations and, by extrapolation, to explore the possibility of similar conduct within associated business activities. Based on a review of some initial documents, the investigators decomposed their task into five separate lines of enquiry or 'issues', and split into separate sub-teams to address them .

Figure 1 shows a very simple schematic of the investigation process. In many ways this corresponds with Pirolli and Card's notional model of analyst sensemaking [4]. The recovered documents were added to a server and were searchable. Queries were devised to retrieve documents relevant to each sub-team's issue (document selection) and the resulting documents were reviewed by the relevant sub-team and electronically coded for relevance to any of the issues currently active within the investigation (document review and classification). This had the effect of forming collections of relevant documents on which further work could be performed.

Important information was then extracted from the relevant documents and re-represented within integrated analyses (schematisation). The most important of these were chronologies which represented sequences of significant events, including details of meetings and email communications. Each team created one or more 'issue' chronologies, and as these evolved important content was selected and consolidated into a single master chronology.

Within the master chronology, issue coding inherited from the source documents was maintained and functionality was included such that it could be collapsed to show entries relevant only to a chosen combination of issue codes. Ultimately, the master chronology ran to around 13,000 entries.

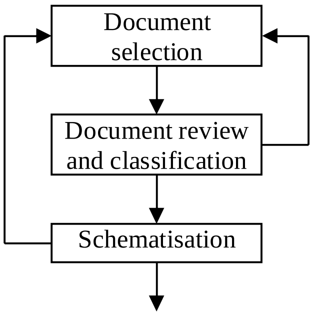

Figure 1. An overview of the investigation process

INVESTIGATIONS WITHIN INVESTIGATIONS

In this section we focus on examples from the case-study data which illustrate the way in which lines of enquiry gave rise to recursively embedded lines of enquiry.

From contract class to specific contracts

One of the investigators' high-level objectives was to explore the possibility of fraud within a particular class of contract. The team dealing with this issue, however, had the initial problem of not knowing what contracts there were within this class:

> P4: Well actually what *[class]* contracts does the company have? And no one in the company knows or can tell you so you're then trying to piece that together.

As evidence was reviewed and related questions were asked in interviews, so the investigators discovered details of each contract. Once this was known, each contract could be defined as an investigation problem in its own right and allocated to an individual researcher. This decomposition was reflected in new searches designed to retrieve information specific to each contract and the creation of new sub-codes for classifying relevant documents and event representations.

From contract focus to a time-period focus

As the investigators responsible for each contract built their chronologies, so particular periods of interest came to light:

> P5: …we'd be thinking, well if we're right on this, this is a really important build up […]. Or, we think money must have been sucked out of this business around this time. […] [Junior Partner] selected certain periods and posed certain questions in relation to those periods. And we would go back and interrogate the information further.

This identification and the focus it provided enabled the investigators to use new strategies for document retrieval typically involving date-delimitation.

> P5: If for example, three days were going to be really important, then we wouldn't worry about search terms. […] We would just say, give me every document that bears this date, created, edited, sent – anything. […]

Other information seeking strategies that took advantage of the identification of particular periods included the examination of telephone records and expense records within certain time-windows. These could provide suggestive evidence for the kinds of activities the investigators were concerned with, but only in the context of the periods in question.

From issue focus to event focus

Working on any of the issues involved the investigators in reviewing retrieved evidence and from these drawing inferences about events, such as meetings and significant communications, and generating event represntations where appropriate. Evidence for an event would often take the form of an email proposing a meeting. But this in itself would not be conclusive.

> P4: […] So you put an entry down for November 20th and then you'd start looking for documents which relates, which might give evidence that that happened, that it actually happened […] and if it did happen who else was involved, who were they meeting, what were they doing, what were they saying to each other?

Faced with this situation, an investigator might focus in on this event and search for further evidence, or they might record the event as a conjecture, and continue to review documents in the hope that they would come across further evidence.

Summary

These cases illustrate the way that discoveries prompted decomposition and refinement of investigation problems. They have some common features which we will briefly explore:

1. Researching issues brought information to light that acted as a cue for more focused enquiry. Without this knowledge focused lines of enquiry would have been impossible;
2. New lines of enquiry were not complete departures but acted as sub-problems. Hence, knowledge outcomes propagated up to inform the outcomes of superordinate issues;
3. Despite 2, each new line of enquiry was independent insofar as it posed new questions and gave rise to new research strategies;

This discussion of the decomposition and focusing of research issues (i.e. investigations within investigations), however, is incomplete without considering how in practice separate issues could inform each other. We have touched on the idea of knowledge outcomes propagating up from sub-issues to issues. The investigators frequently discussed the importance of mechanisms for 'escalating' significant information to team leaders responsible for superordinate problems. An important escalation mechanism was the integration of separate issue chronologies into the master chronology which could then be filtered to align issue sub-sets:

> P6: And the master chronology is actually the key to that because if you're trying to understand what a group of people are doing over a period of time, there are a whole different series of things going on, but it's only really when you start to line some of those things up that you begin to ask yourself questions about, well, "What was really happening in this period of a week?"

In addition to vertical information flow, it was also seen as essential for investigators to frequently discuss findings and theories and exchange information. Multiple mechanisms were put in place to promote this including review meetings, informal 'huddles' and ad hoc document passing. As one lawyer explained:

> P4: The amount of communication that has to go on in order to make that work is phenomenal.

The organisation of the investigation around embedded lines of enquiry has led us to consider how this structure might be reflected within systems for supporting large-scale collaborative sensemaking activities in an effort to address challenges of scalability and collaboration. Of particular interest is a means by which users could define and select from a set of 'contexts' within the investigation information, and so at any time eliminate information extraneous to their interests.

In order to explore this idea we have examined the structure of a line of enquiry more deeply to provide a detailed account of its conceptual elements.

To do this we performed a Grounded Theory [6] analysis of our data using the concept of a line of enquiry as a core category with the aim of developing an ontological framework of entities associated with a line of enquiry at any level of granularity. We describe this framework in the following section.

THE LINE OF ENQUIRY FRAMEWORK

The framework takes a *line-of-enquiry* as a primary object. This is associated with *theories*, *questions*, *information seeking strategies*, *evidence and evidence collections*, *knowledge, assigned investigators* and *lower-level lines of enquiry*. Each line of enquiry included these elements. Significantly, lines of enquiry recursively embed such that the knowledge generated by each can give rise to one or more *sub-problems*, each with similar structure.

Theories

Our data shows that theories or conjectures were central to any line of enquiry.

> P4: Well it's the theories that then define the issues you are coding for and looking for. […] we had lots of sub-issues and theories, well sub-theories that were helping to define the issues […]

A theory would be triggered by a cue. This could be an allegation, or knowledge arising from the investigation. The examples above show how identifying a business activity of a certain type, a key time period or evidence of an event could provoke a more focused line of enquiry. Each of these was associated with a theory, however broad, about what could have been the case.

Theories were systematically investigated and eliminated when the evidence found was contradictory or unsupportive. When all the theories associated with a line-of enquiry were eliminated then the issue would become inactive.

Questions

The investigators made a natural move from theories to research questions, and in many cases these were explicitly recorded. Research questions specified requirements for information that would test theories or simply elaborate their focus. This elaboration could then provide cues for further decomposition or could yield other unexpected findings.

Information seeking strategies

Given the questions, each line of enquiry would have a set of bespoke information seeking strategies. These might include keyword searches over the document collection, the examination of telephone records or interviews with witnesses.

Evidence and evidence collections

Searches provided the investigators with collections of potentially relevant documents. A line of enquiry could have multiple associated searches and these could be repeated periodically as new documents were added to the collection. Results set collections were manually reviewed for relevance and relevant documents were tagged to associate them with the line of enquiry. This then created a further collection of documents which was used for generating knowledge representations.

Knowledge

Within each line of enquiry the investigators continually reviewed and collated evidence and recorded the inferences they drew from it within different forms of knowledge representation.

The representations they created were organised around two types of concept. The first was people and their relationships. It was important to discover and maintain a record of the known key players within a line of enquiry. The investigators created profiles of individuals and in some cases drew link charts to represent relationships.

The investigators also constructed chronologies of events. These then provided a basis from which theories were evaluated and written briefings created. Event records included a date and time, a summary description, a list of people involved in the event, and references to the supporting evidential documents.

Assigned investigators

Given the team setting, any line of enquiry could be allocated to one or more investigators. Hence, from the perspective of the investigating team, these assignments formed part of the concepts associated with each line of enquiry.

Lower level lines of enquiry

Finally, knowledge associated with a line of enquiry could give rise to any number of more focussed problems. These lower-level issues featured more focussed theories, questions and information seeking strategies and gave rise to their own knowledge. They could be assigned to a smaller sub-set of investigators, or they could act as small scale deviations for a single investigator.

DESIGN

The framework provides an ontology of concepts associated with any given line of enquiry. We have found these elements to occur irrespective of granularity. In some cases, a line of enquiry might concern a single relationship or a single event, whilst the investigation as a whole can be considered a single line of enquiry.

When instantiated, the framework gives rise to a hierarchy of enquiry nodes, with relevant elements represented at each node. By implementing this framework within a sensemaking support system we anticipate a number of advantages centring around the decomposition and integration of multiple strands of an enquiry.

By allowing investigators to selectively access information associated with a particular line of enquiry, the framework would support the isolation of associated information for focussed analysis. Users would be able to eliminate extraneous information and focus on elements of interest to them.

Conversely, with outcomes propagating up within the hierarchy, they could also achieve an integrated overview at any higher level. This has implications for sensemaking representations such as chronologies and link charts. By associating the component elements of such representations with framework nodes, users could use node selection to view these different strands of the investigation in different combinations, thus enabling them to easily explore links between apparently separate issues.

Finally, integrating data and user-generated knowledge representations from multiple aspects of a collaborative investigation provides an opportunity for the system itself to identify links between disparate parts of a large investigation. For example, the system might automatically match common characters or travel locations across apparently unrelated lines of enquiry. Investigators alerted to these could then explore the extent to which they offer explanatory leverage. The details of this matching would depend upon specific user-needs and the details of data and knowledge representations within the system. However, the opportunity for automated matching may itself dictate requirements on how information is represented within the system.

CONCLUSION

To address the challenge of decomposition and integration presented by large scale, collaborative investigations we have analysed the structural composition of one such investigation and developed a framework which describes recurring elements associated with multiple, embedded lines of enquiry. This recursive framework structures large-scale sensemaking challenges as ‘investigations within investigations’. Understanding this can enable system design which allows users to reflect on and develop theories, questions, information seeking strategies, evidence and knowledge that are relevant to them at multiple levels of description. The framework effectively defines a line of enquiry as a context for a particular strand of a larger sensemaking activity.

There may be factors which delimit the generalisability of this framework either as a descriptive model of practice or a normative model for sensemaking support tools. We intend to explore these questions through further fieldwork, and by implementing the framework within an integrated sensemaking system. But we hypothesise that the framework generalises well to sensemaking tasks of a given size and complexity in which multiple emerging aspects need to be both explored in detail and understood collectively. In particular we believe that the framework provides a valuable normative model for collaborative sensemaking applications.

ACKNOWLEDGMENTS

We would like to thank Freshfields Bruckhaus Deringer for their kind help with this study. The work was funded under EPSRC grant EP/D056268.

REFERENCES

1. Baron, J., Braman, R., Withers, K., Allman, T., Daley, M., and Paul, G., The Sedona Conference best practice commentary on the use of search and information retrieval methods in e-discovery. *The Sedona Conference Journal 8* (2007), 189-223
2. Bodnar, J.W. Making Sense of Massive Data by Hypothesis Testing. In *Proc. International Conference on Intelligence Analysis*, (2005).
3. Klein, G., Moon, B. & Hoffman, R. Making Sense of Sensemaking 2: A Macrocognitive Model. *IEEE Intelligent Systems 21*, 5(2006), 88-92.
4. Pirolli, P. and Card, S., The Sensemaking Process and Leverage Points for Analyst Technology as Identified Through Cognitive Task Analysis, In *Proc. International Conference on Intelligence Analysis*, (2005).
5. Schmidt, L.K. *Understanding Hermeneutics*. Acumen Publishing Ltd, (2006).
6. Strauss A, and Corbin J. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 2nd ed. Sage Publications, Inc; 1998.