Visual Analytics: Data, Analytical and Reasoning Provenance

Margaret Varga and Caroline Varga

Abstract Analysts and decision makers are increasingly overloaded with vast amounts of data/information which are often dynamic, complex, disparate, conflicting, incomplete and, at times, uncertain. Furthermore, problems and tasks that require their attention can be ambiguous, i.e. they are ill-defined. In order to make sense of complex data and situations and make informed decisions, they utilize their intuition, knowledge and experience. Provenance is fundamental for the user to capture and exploit effectively the explicit data and implicit knowledge within the decision making process. Provenance can usefully be considered at three conceptual levels, namely: data (what), analytical (how) and reasoning (why). This paper explores visual analytics in the exploitation of provenance within the decision making process.

Keywords Analytical provenance • Data provenance • Hypothesis • Reasoning provenance • Visual analytics • Visualization

1 Introduction

Analysts and decision makers are increasingly overloaded with vast amounts of data/information which are often dynamic, complex, disparate, conflicting, incomplete and, at times, uncertain. Furthermore, problems and tasks that require their attention can be ambiguous, i.e. they are *ill-defined*. In order to make sense of complex data and situations and make informed decisions, decision makers rely on explicit information and their implicit intuition, knowledge and experience. Moreover, to have confidence in a decision making process, it is necessary for them

M. Varga
Seetru Ltd, Bristol, UK,
e-mail: margaret.varga@seetru.com; margaret.varga@oncology.ox.ac.uk

University of Oxford, Oxford, UK

C. Varga
Seetru Ltd, Bristol, UK

V.L. Lemieux (ed.), *Building Trust in Information*, Springer Proceedings in Business and Economics, DOI 10.1007/978-3-319-40226-0_9

to understand the sources of information and thus the value and trust that can be placed on every aspect of the process; i.e. the provenance [1–4].

Provenance is fundamental for the user to capture and exploit effectively the explicit data and implicit knowledge within a decision making process. Provenance can usefully be considered at three conceptual levels, namely: data, analysis and reasoning [5]. In essence it comprises the what (data), how (it was analyzed) and why (reasoning).

This paper explores the application of visual analytics as an effective means of analyzing and understanding provenance in the explicit representation of the analytical and reasoning processes: how and why the data is used.

2 Data, Analytical and Reasoning Provenance

There are three categories of provenance that play a role in visual analytics, namely: data provenance, analytical provenance and reasoning provenance. In order to understand findings/discoveries it is necessary to document the entire analysis process and retain all three types of provenance. Capturing the reasoning processes is by far the most challenging.

- **Data provenance** considers the source of the data, and the link between the source and the system using the data. The data may be intelligence reports, videos, network logs, etc. The provenance of the data must certainly be taken into account in the analytical and visualization approaches and processes when addressing problems/making decisions.
- **Analytical provenance** is concerned with the processes performed on the data; in particular, here, the techniques used to analyze and visualize the data. The analysis conducted has an impact on the nature of the results and how the results can be used. The actions performed during an analysis within a visual analytic system can be captured: i.e. data transformations, events (e.g. key strokes) and actions (e.g. zoom) can easily be logged, and the overall history of interactions can be recorded [4]. VisTrails, for example, supports exploratory computational tasks and also provides a provenance management infrastructure [6].
- **Reasoning provenance** is the most challenging to identify, make explicit and capture; it is concerned with how and why analysts arrive at their conclusions/decisions. It is typically concerned with the application of human experience, knowledge and intuition.

Annotation of the analysis can be used to enable recall and sharing [7]. Externalization can be achieved through think-aloud protocols: this process, however, may alter the nature of the reasoning, reduce task performance, or even risk changing decisions [8, 9]. Furthermore, analysis of such externally captured data is extremely time-consuming and labour intensive.

The results of experience, knowledge and intuition used in a decision making process can be presented in visual (e.g. diagrams), textual or numeric narratives.

These narratives are based on the available data and are used to show salient information and different hypotheses. In the reasoning process, interconnections between narratives, and between information and hypotheses, are developed to support informed decision-making [5].

Visual representations of the reasoning space through networks of narratives enable the understanding of the reasoning process and considerably improve both the quality of the reasoning process and the efficiency/effectiveness of informed decision making [5, 10, 11].

3 Visual Analytics

Visual Analytics is the science of analytical reasoning facilitated by interactive visual interfaces. It combines automated analysis techniques with interactive visualizations to allow the user to interact with, explore and analyze big and complex data, both dynamically and visually. It thus facilitates data and situational understanding so as to support informed decision-making.

It is necessary to create tools and techniques to help users to derive information and insight from massive and complex data: to detect the expected and/or discover the unexpected. The tools must also support the provision and communication of timely and accurate situation assessments—upon which users can act [11–14].

4 Intelligence Analysis

4.1 Introduction

Across all subject domains, one concern is how to incorporate and make use of provenance to enhance informed decision making—to better understand how and why data is used and decisions are made. This section uses a case study on intelligence analysis to illustrate the ideas.

Intelligence analysis is the application of individual and collective cognitive methods to explore data and test hypotheses. Events and evidence are assessed, for example, to explain/interpret events that 'might' happen; or to decide how best to prevent the occurrence of an adverse event; or to minimize potential damage [15, 16], etc.

Intelligence analysts respond to Intelligence Requests (IRs), which can be precise or ill-defined. Critical thinking is essential in order to provide the 'best possible' answer, within a short time frame. Important elements of critical thinking are to reduce bias and present all possible options to a decision maker [17].

4.2 An Example Case Study

An example case study of an ill-defined IR is discussed here: the IR is—"What are the current threats?". The threats may be of any form and of varying degrees of urgency.

In order to determine potential threats data/information must be gathered and analyzed, and the situation must be assessed. The experience of an intelligence analyst will guide them to filter 'noisy' data and to zoom into potential threats that require further investigation. If it is known, for example, that there is an upcoming state visit or a military supply convoy, the analysts' experience/intuition might lead them to identify such events as potential targets, and thence to hypothesize possible/likely threats. Any of the hypothesized threats may be true so analysts must consider all relevant information to make informed decisions.

For example, in the case of a convoy, an ambush may be identified as the mostly likely threat [18]; the analyst must then hypothesize likely ambush locations and gather evidence to answer question such as:

- Where and when is the convoy going, and what route will it take?
- Where have recent ambushes been?
- Where are insurgents currently known to be operating?
- What types of ambushes are the insurgents capable of? Land or sea? Chemical or biological?
- What is the certainty that this route is not going to change (e.g. commander deciding to change the route, flooding, unexpected roadblock, etc . . .)?
- What is the weather forecast? Hurricane? Snowing?
- . . . .

Many different approaches may be used to narrate, assimilate and analyse hypotheses and evidence. Here, the Wigmore concept [19] is used to demonstrate the generation, representation and analysis of multiple hypotheses, as well as provide an answer to the IR including the representation/presentation of the analytical provenance [18]. The Wigmore chart was developed as a graphical method for the analysis of legal evidence in trials. It was the first diagrammatic system of charting arguments; other approaches include Toulmin [20]. One of the advantages of the Wigmore approach is its handling of the balance of view (cf. bias). In a Wigmore chart, various types and items of evidence supporting and refuting a hypothesis are represented graphically; this allows the strength/weakness of the case to be readily observed. In particular, it is easy to see 'gaps' where additional effort is required to gather evidence, e.g. to minimise uncertainty or danger of self-confirmation, or to strengthen an aspect of the case/hypothesis.

In Fig. 1, the hypothesis introduced suggests that a potential ambush location is south of Village A [18]. The analyst enters hypothesis properties based on their experience/intuition/knowledge and their information sources. Sources are rated in terms of their reliability [15]:

(1) Completely reliable,

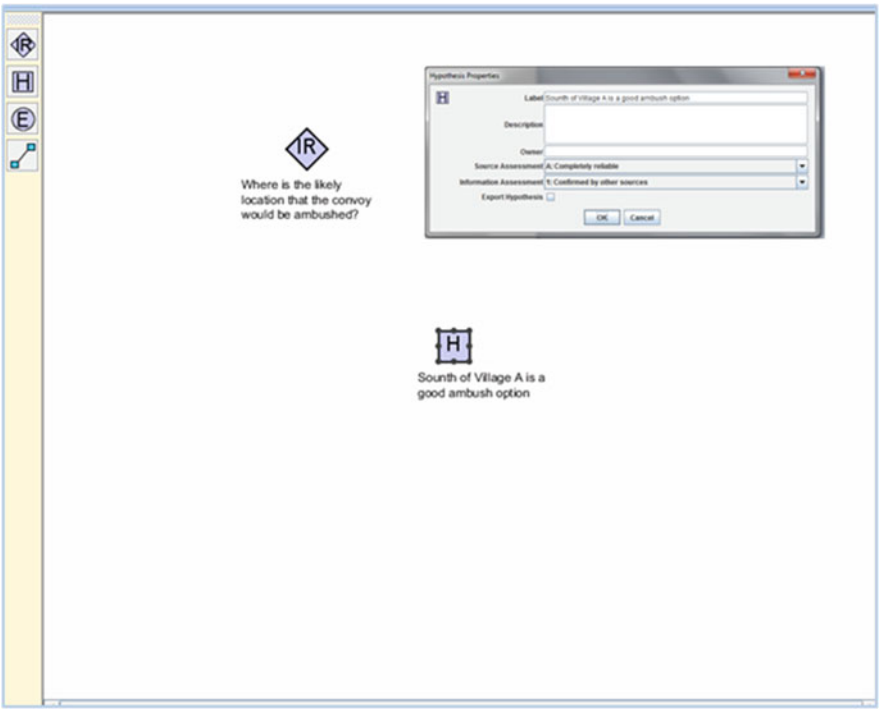

Fig. 1 Hypothesis that South of Village A is a good ambush option and its properties [18]

(2) Usually reliable,
(3) Fairly reliable,
(4) Not usually reliable,
(5) Unreliable, and,
(6) Reliability cannot be judged.

The information is also rated accordingly; see again Fig. 1:

(1) Confirmed by other sources,
(2) Probably true,
(3) Possibly true,
(4) Doubtful,
(5) Improbable, and,
(6) Truth cannot be judged.

In this way, multiple ambush locations (hypotheses) based on the analysts' experience—obvious and less obvious alternatives—can be considered, as opposed to pursuing a biased approach in which, for example, a narrow focus might be pursued.

The provenance of each hypothesis that is being considered can be recorded and assessed; guided by the analyst's intuition, experience, local knowledge and available data. Understanding the processes through which ambush locations—or other hypotheses—are identified and evidence is assessed is important in understanding how and why analysts reason with/about them (i.e. reasoning provenance).

In Fig. 1, the hypothesis (the potential location) source is assessed to be 'completely reliable' and is based upon information 'confirmed by other sources'. The Analyst can export the hypothesis to share with other analysts, make notes about the hypothesis (reasoning), and declare any ownership.

Next, the analyst must gather evidence to support or refute this particular hypothesis (i.e. build a balance of view). The evidence properties are entered using the same rating system as the hypothesis properties. In this example, there is supporting evidence to show that there is good cover for the attackers at Village A, which makes Village A vulnerable. This evidence is believed to be from a 'usually reliable' source and is 'probably true', see Fig. 2. More evidence will be gathered and similar processes will be used for other evidence and other hypotheses.

Different types of evidence can be used; for example, significant trends might emerge from circumstantial evidence that can be correlated with other evidence.

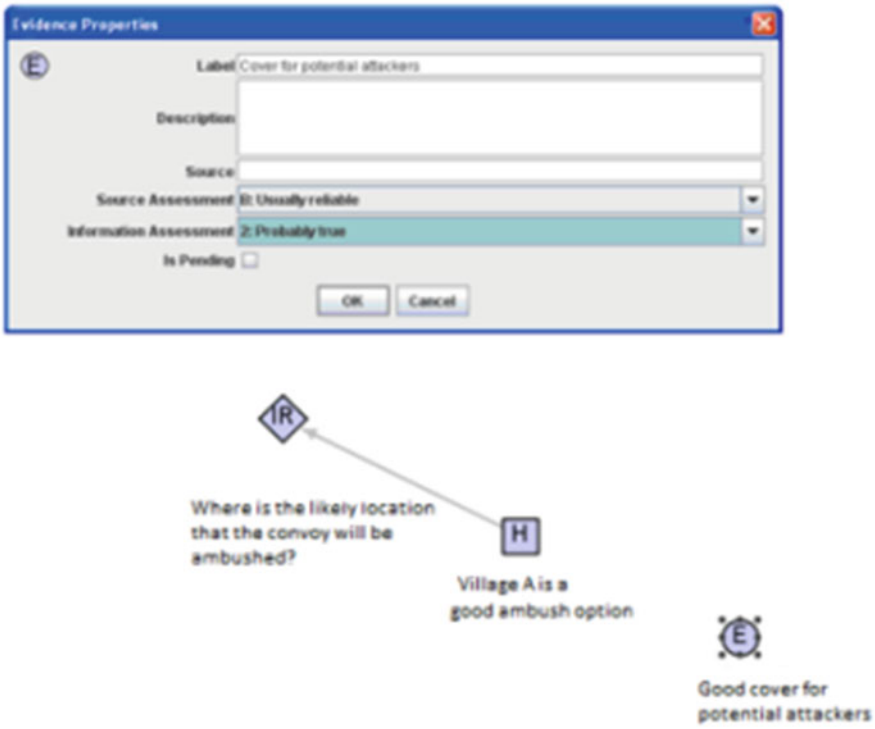

Fig. 2 Properties of the evidence that there is a good cover for the attackers [18]

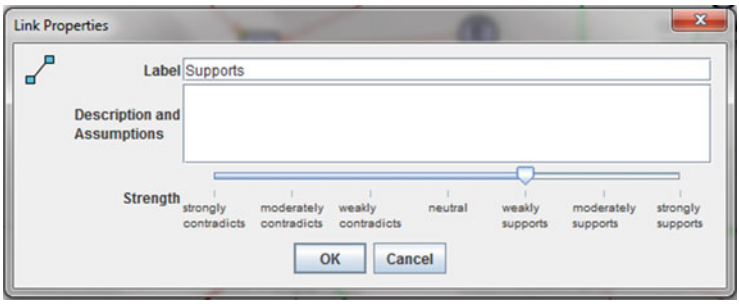

Fig. 3 Link properties [18]

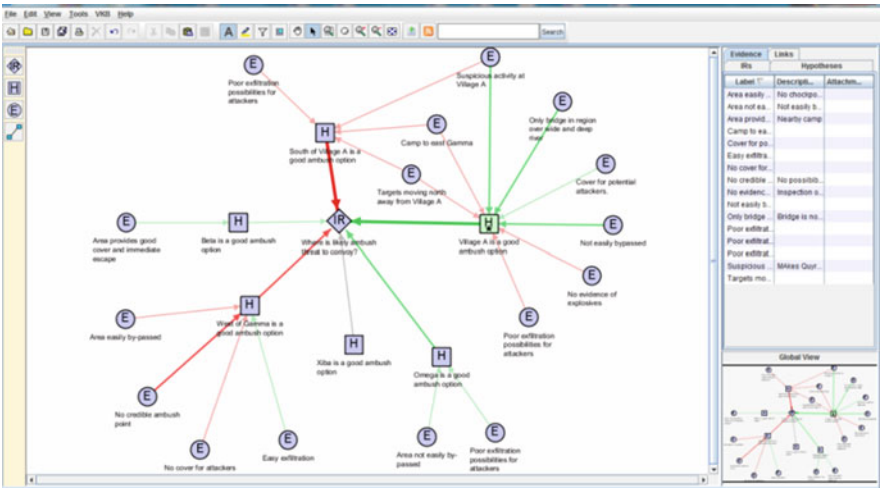

Fig. 4 Network of six hypotheses [18]

The 'Is Pending', checkbox allows the analyst to represent evidence that someone has been tasked to collect, see Fig. 2. This can be added to the graph to be assessed later when the evidence is available. The origin of the evidence is also recorded; that is, its data provenance.

The analyst also enters the strength of the evidence in supporting or refuting the hypothesis; this ranges from strongly contradicts, neutral, to strongly supports, see Fig. 3. This rating is used to determine the role each item of evidence plays in supporting/refuting the hypothesis (it should be noted that the reliability of a source does not determine its relevance or strength).

Figure 4 shows a network of hypotheses (denoted 'H') and evidence (denoted 'E') built up in answering the IR. Green links represent support while red links represent refutation; their strength is represented by the thickness of the link.

Among the six possible locations/hypotheses, no evidence has been collected for the hypothesis that Xiba is a possible ambush location (there are no linked

evidence/'E' circles): either evidence needs to be gathered to address the proposed hypothesis, or it should be removed from the view (it is always possible to display the hypothesis again if evidence becomes available and it is still a viable hypothesis). Conversely, many items of evidence—of varying strengths—are linked to the hypothesis ('H' box) of Village A.

Figure 4 may thus be interpreted as showing a balance of view of all possible ambush locations, as well as illustrating the availability of evidence and the efforts that have been invested in collecting and assessing the evidence to support/refute any hypothesis. The presentation of a balance of views in this manner is an important element in reducing cognitive bias. Inputs from other analysts can also reduce bias.

The upper right hand table on Fig. 4 provides a means by which all evidence and associated information may be examined. The bottom right hand panel provides a global overview facility which can be used to navigate around the network: e.g. when focusing on sub-components of the problem such as Village A. Alternatively, another part of the network may be looking at the threats relating to an imminent state visit; in this case the global overview would show all the possible threats and how they might relate.

The same evidence may be used to support multiple hypotheses, or to support one hypothesis and refute another; it is not necessary to input the same evidence again for different hypotheses. For example, "suspicious activity at Village A" relates to multiple hypotheses, see Fig. 4. The advantage of this is that it avoids creating a misleading impression of the number of evidence items available; that is, the same evidence appearing multiple times for different hypotheses.

The influence of individual and combined evidence is analyzed automatically for the six hypotheses. It is vital that the system can readily be updated to respond to rapidly evolving situations; here, the effect of new information can be visualized instantly. Hypothesis analysis is a dynamic process; new hypotheses can be generated or removed when the situation changes, when new evidence is gathered or when there are changes in existing evidence/situations. Hypotheses can also be saved, re-used or modified for future IRs.

Different analytical processes can be used to analyze the data. In this example, the system calculates the strength of all the evidence relating to all the hypotheses; this reveals that Village A is indeed a likely ambush location. In light of this, the convoy should either alter its route to avoid Village A or prepare for a possible ambush. The display can also be used to brief the commander about all the threats considered, which is the most likely threat, and why. The understanding of the hypotheses and corresponding evidence as well as the analysts' notes give an idea of the reasoning of why the locations were chosen.

This concept can be transferred for use in other applications such as financial risk analysis or medical analysis [18].

5 Conclusions

The case study has illustrated the interactive formation, visualization and analysis of dynamic hypotheses for decision-making in ill-defined problems; with provenance for explicit data, defined analytical process and some implicit knowledge/experience from the analysts. It shows the benefits of interactive analysis and visualization in response to changes in evidence and hypotheses; such as reducing bias and improving efficiency. Furthermore, the approach considers the degree of uncertainty in each piece of evidence and its role in supporting/refuting different hypotheses. The data provenance and the data uncertainty can be expressed by the user based on assessment of its source, assessment of the information, as well as consideration of the links between the evidence and hypotheses. It also provides an audit trail of the analysis in terms of data provenance, analytical provenance, and, to certain extent, reasoning provenance.

This case study shows narratives as the explicit representations of the hypotheses, which include different types of data in their presentation, such as: the explicit data used and its provenance; the processing and manipulation performed; and, the implicit information from the analysts' knowledge and experience.

Systems that allow for the dynamic visualization of hypotheses which develop over time, and change with the arrival of 'new information' or the application of a 'new process', provide invaluable support for informed and dynamic decision making in ill-defined problems. The system in the illustrated case study is an example of this capability. It also provides methods to visualize competing hypotheses or complementary theories (that would support and enhance the strength of a particular argument), each depicting different degrees of certainty.

The case study shows that although many pieces of the puzzle have been found, much research is still needed to further the development of tools to support informed decision making for ill-defined problems. Robust reasoning provenance about how and why analysts make decisions, deduced from implicit data, would complete the audit trail for understanding what, how and why data were used.

References

1. Attfield, S.J., Hara, S.K., Wong, B.L.W.: Sensemaking in visual analytics: processes and challenges. In: Kohlhammer, J., Keim, D. (eds.) EuroVAST 2010: International Symposium on VAST, pp. 1–6. Eurographics Association, Bordeaux, France (2010)
2. Gotz, D., Zhou, M.X.: Characterizing users' visual analytic activity for insight provenance. Inf. Vis. **8**(1), 42–55 (2009)
3. Jankun-Kelly, T.J.: The Case for Visual Analysis Provenance Cases, Workshop on Analytic Provenance: Process + Interaction + Insight, CHI. (2011)
4. Venters, C.C., Austin, J., Dibsdale, C.E., Dimitrova, V., Djemame, K., Fletcher, M., Fores, S., Hobson, S., Lau, L., McAvoy, J., Marshall, A., Townend, P., Taylor, N., Viduto, V., Webster, D.E., Xu, J.: To trust or not to trust? Developing trusted digital spaces through timely reliable and personalized provenance. In: Provenance for Sensemaking. Paris, France (10th November 2014)

5. Roberts, J.C., Keim, D., Hanratty, T., Rowlingson, R., Hall, M., Jacobson, Z., Lavigne, V., Rooney, C., Varga, M.: From Ill-defined Problems to Informed Decisions. EuroVis Workshop on Visual Analytics, UK (2014)
6. Silva, C.T., Freire, J., Callahan, S.: Provenance for visualizations: reproducibility and beyond. Comput. Sci. Eng. **9**(5), 82–90 (2007)
7. Groth, D., Streefkerk, K.: Provenance and annotation for visual exploration systems. IEEE Trans. Vis. Comput. Graph. **12**(6), 1500–1510 (2006)
8. Boren, T., Ramey, J.: Thinking aloud: reconciling theory and practice. IEEE Trans. Prof. Commun. **43**(3), 261–278 (2000)
9. Hertzum, M., Hansen, K.D., Andersen, H.H.K.: Scrutinising usability evaluation: does thinking aloud affect behavior and mental workload? Behav. Inf. Technol. **28**(2), 165–181 (2009)
10. Schacter, D.: Psychology, 2nd edn. Worth Publishers, NY (2009)
11. Thomas, J.J., Cook, K.A.: Illuminating the Path: The Research and Development Agenda for Visual Analytics. National Visualization and Analytics Centre (2005)
12. Keim, D.A., Kohlhammer, J., Ellis, G., Mansmann, F.: Mastering the Information Age-Solving Problems with Visual Analytics. Florian Mansmann (2010)
13. Keim, D.A., Mansmann, F., Thomas, J.: Visual analytics: how much visualization and how much analytics? SIGKDD Explor. **11**(2), 5–8 (2009)
14. Thomas, J.J.: Taxonomy for Visual Analytics: Seeking Feedback. VAC Views (May 2009)
15. FM 2-22.3 (FM 34-52) Human Intelligence Collector Operations, US Department of the Army (September 2006)
16. Johnson, R.: Analytics Culture in the U.S. Intelligence Community: An Ethnographic Study, Centre for the Study of Intelligence. Central Intelligence Agency, Washington (2005)
17. Moore, D.T.: Critical Thinking and Intelligence Analysis, Occasional Paper Number 14. National Defense Intelligence College, Washington (March 2007)
18. Varga, M.J., Adams, K.: Interactive hypothesis visualization. In: NATO Workshop on Visualising Networks: Coping with Change and Uncertainty (October 2010)
19. Anderson, T., Schum, D., Twining, W.: Analysis of Evidence, 2nd edn. Cambridge University Press (2005)
20. Toulmin, S.E.: The Uses of Argument - Updated Edition. Cambridge University Press, Cambridge (2003)