Provenance for Intelligence Analysis using Visual Analytics

B.L. William Wong
Middlesex University
London, UK
w.wong@mdx.ac.uk

Kai Xu
Middlesex University
London, UK
k.xu@mdx.ac.uk

Simon Attfield
Middlesex University
London, UK
s.attfield@mdx.ac.uk

ABSTRACT
In this paper we discuss various aspects in intelligence analysis relating to provenance and the new requirements resulting from the changed nature of terrorist activities. We propose a three-layer provenance model which relates the scope of provenance to the intelligence workflow and the idea of a 'provenance reasoning workspace' for integrating provenance information into visual analytic tools.

Author Keywords
Provenance, Visual Analytics, Intelligence Investigation

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: Miscellaneous

General Terms
Security, Theory

INTELLIGENCE ANALYSIS AND PROVENANCE

Intelligence agencies engaged in anticipating and countering terrorist threats face a complex and uncertain task. During such analysis information is sought and extracted from multiple sources and used to produce new integrating representations (e.g. communication network graphs, chronologies, narratives, argumentation structures) in an emerging and iterative process. Conclusions based on derived from these analyses form the basis of reporting and ultimately decision-making. However, the difficulties in dealing with very large qualities of dynamically changing information from multiple sources in a wide range of forms combine with a need to focus quickly on key threats and to develop analyses with understood levels of reliability.

Whilst intelligence analysis has always been an information-intensive activity, the global environment has changed dramatically since the end of the Cold War. A small number of centralized groups with established methods of operations have been replaced by agile, decentralized terrorist cells who are continually reinventing the nature of the threat. Anticipating attacks through behavior analysis is potentially confounded by the data sources and patterns which are relevant to the analysis changing regularly, forcing security agencies around to continually adapt their data sources and analytic methods in the pursuit of useful insight. This forces a new kind of creativity amidst overwhelming, mixed format multi-sourced data of varying quality and completeness; data which are often confusing and confounding because they are inconsistent and lacking context.

As part of the new agile analysts toolset there is increasing interest in the use of dynamic, interactive visual tools as a means for supporting more effective analyses and faster insights. In the support of analyzing group behavior, visualizations can help with the identification of key players in complex communication networks; show emergent themes in unstructured datasets (such as SMS messages, email and social networking resources); expose visual correlations across multiple parameters; show geo-temporal relationships between key players, locations and events; and support the collaborative construction of high-level reasoning artifacts. Visual analytics is the science of analytical reasoning facilitated by interactive visual interfaces [2]. The use of visual analytics in intelligence analysis has significant potential for tracking and analyzing the behavior of groups. In particular, visual analytics coupled with semantic analysis can support the analysis of intelligence in real-time, providing early warning of imminent attack and the alerting of relevant authorities. In order to be effective, however, the integration of visual analytics into intelligence analysis requires due attention to increasing demands for diligence and accuracy. In the UK, for example, this has been translated into policy that requires analysts to challenge assumptions, review and assess the quality of decisions made at different levels of the command chain and in earlier stages of analysis, and to use multiple points of reference to verify and validate evidence, analyses and the conclusions that are inferred from them. These needs reflect the importance of maintaining a balance between detecting and acting against potential threats in an uncertain environment, and protecting individual human rights and democratic freedoms within the same setting.

Achieving such balance demands processes which can adapt dynamically, following the twists and turns of changing information and emerging insights, hunches and data, whilst forming pathways and conclusion that are explicit and amenable

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

to ongoing review. During intelligence analysis information is sourced, sifted and transformed through multiple stages in an opportunistic blend of top-down and bottom-up reasoning [1]. The need for traceable creativity throughout this process amid high levels of emergent complexity places particular demands on analysts and on the technologies and artifacts at their disposal.

INTELLIGENCE PROVENANCE AND VISUAL ANALYTICS

We focus on this traceability as an essential part of individual and collaborative reflective dialogues with both evolving and completed analyses. Included within it are a need to track the sources of data, their originating contexts and reliability; the need to track and review the analytic process; and the need to track assumptions and background knowledge as these combine with data to support insight and contribute to an assessment. Information that supports this kind of traceability is referred to as provenance information and its potential availability as part of the analysis process is significant not only for assessing the plausibility of conclusions (i.e. the quality question), but through an assessment of analysis strengths and weaknesses, for focusing effort on potentially high-return avenues of inquiry (i.e. the efficiency question).

Visual analytic tools offer new capabilities for displaying and organizing data and complex reasoning. The challenge is to enhance such tools with provenance information so that users can easily track data and data transformations (i.e. data lineage) and reasoning paths in order to help them understand their own complex, emerging analyses and enhance rigor. The aim is to render an otherwise opaque reasoning process transparent such that analysts can view, trace and probe how conclusions came about, and thus query the soundness of reasoning and quality of the data used. This will give analysts greater confidence in generated conclusions and allow them to be evaluated independently. Furthermore, by incorporating a framework of permissions management applicable to different components of an analysis, transparency can be controlled such that aspects of an analysis can be made available to other analysts with differing security clearances either in the same team, in different departments or even in other countries.

Our approach is informed by a conceptual framework illustrated in Figure 1. This shows the problem space divided into three complementary work areas. This framework draws attention to stages of analytic products we believe are always present within the analytic workflow, whether as external (publicly available) or cognitive (private) artifacts. Together they form what we refer to as the intelligence analysis reasoning workspace. The three levels are:

i. The data level, which includes raw data derived from external sources (e.g. communications, financial records, signals intelligence reports, photographs etc.)

ii. The analysis products level, which includes the results of data manipulations resulting in the creation of abstracted representations that draw out key facts about a domain. These might typically be interactive visualization states (temporarily frozen) or manually constructed, semi-formal schemas which select and re-represent data in ways designed to support higher-order reasoning. Critically, these representations sort and structure data but do not go beyond it.

iii. The reasoning products level, which incorporates findings into a framework in which multiple analysis products integrate with high-level reasoning artifacts such as interpretations, hunches, assumptions and questions to form an argumentation structure.

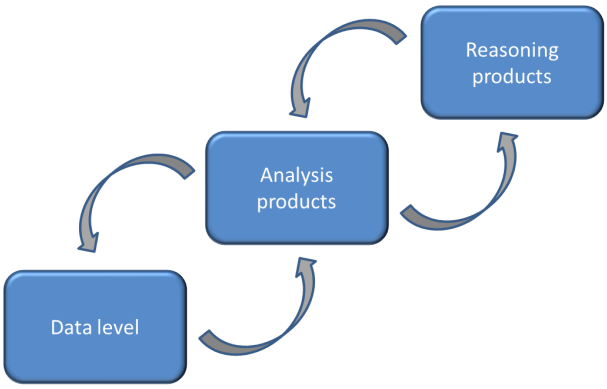

Figure 1. The intelligence analysis reasoning workspace decomposed into data, analysis products and reasoning products.

Provenance operates at these three levels: (1) tracking data and information sources, (2) tracking automated or semi-automated data manipulations and analytical moves (e.g. data integration, summarization, the computation of new indices and rendering of visualizations), and (3) tracking the resulting lines of reasoning and argument in which knowledge is used to interpret and draw conclusions, and generate new hypotheses. To address the challenge of tracing provenance in the context of intelligence analysis we propose a conceptual framework. This framework is what we refer to as the provenance reasoning workspace. Conceptually it comprises three spaces: (i) a data space that reveals what is stored or available for consideration and use, (ii) a computation space for carrying out various statistical and/or other transformations on the data in order to discover patterns and other relationships in the data in a semi-automated way, and (iii) a reasoning space to explore hunches and generate hypotheses, and to test them by collating evidence (from the data or from computations) that would support or refute the claims, and to thereby create conclusion pathways that can be tracked and probed for plausibility of the arguments and lines of reasoning, validity and quality (reliability, uncertainty) of the data.

REFERENCES

1. P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, 2005.

2. J. Thomas and K. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. National Visualization and Analytics Center, 2005.