

Analytic provenance for criminal intelligence analysis

Junayed Islam, Kai Xu, William Wong

Department of Computer Science, Middlesex University, London NW4 4BT, United Kingdom

Abstract: In criminal intelligence domain where solution discovery is often serendipitous, it demands techniques to provide transparent evidences of top-down and bottom-up analytical processes of analysts while sifting through or transforming sourced data to provide plausible explanation of the fact. Management and tracing of such security sensitive analytical information flow originated from tightly coupled visualizations into visual analytic system for criminal intelligence that triggers huge amount of analytical information on a single click, involves design and development challenges. In this research paper, we have introduced a system called “PROV” to capture, visualize and utilize analytical information named as analytic provenance by considering such challenges. A video demonstrating its features is available online at <https://streamable.com/r8mlx>. Prior to develop this system for criminal intelligence analysis, we conducted a systematic research to outline the requirements and technical challenges. We gathered such information from real police intelligence analysts through multiple sessions who are the end users of a large heterogeneous event-driven modular Analyst’s User Interface (AUI) of the project VALCRI (Visual Analytics for Sensemaking in Criminal Intelligence), developed by using visual analytic technique. We have proposed a semantic analytic state composition technique to trigger new insight by schematizing captured reasoning states. To evaluate the system we carried out few subjective feedback sessions with the end-users of the project and found very positive feedback. We also have tested our event triggered analytic state capturing protocol with an external geospatial and temporal crime analysis system and found that our proposed technique works generically for both small and large complex visual analytic systems.

Key words: analytic provenance, visual analytics, transparency, visualization design and sensemaking

doi: 10.11959/j.issn.2096-109x.2018016

1 Introduction

Provenance is a broad topic that has many meanings in different contexts. The Oxford English Dictionary defines provenance as “the source or origin of an object; its history and pedigree; a record of the ultimate derivation and passage of an item through its various owners.” In scientific experiments, provenance helps to interpret and understand results by examining the sequence of steps that led to a result; we can gain insights into the chain of reasoning used in its production, verify that the experiment was performed according to acceptable procedures, identify the experiment’s inputs, and, in

some cases, reproduce the result. In criminal intelligence it has a greater impact to understand the process by which the decision has been made. Nowadays the large and complex event-driven systems around us are computationally intense where data flows from one process to another as it is transformed, filtered, fused, and used in complex models in which computations are triggered in response to events. Provenance capturing and representing to support judgmental process of criminal intelligence

analysis by using such computation systems with hundreds of interconnected services that creates huge volume of data at a single run is a matter

Received data: 2017-12-04, **Revised date:** 2018-01-15

Corresponding author: Kai XU, k.xu@mdx.ac.uk

of obvious challenge.

Wong, et al^[1] proposes a three-layer provenance model which describes the relationship between the provenance and the intelligence process i.e, the data provenance, process provenance and reasoning provenance. The process and reasoning provenance are termed as “Analytic Provenance” at it’s broader category. Criminal intelligence analysts are likely to benefit from the ability to review the way in which the data they collected evolved during the intelligence process (data provenance) and from the possibility to track back various activities in which they engage (analytic provenance). This is likely to help them in dealing with the complexity of the intelligence process considering the limited capability of the human mind to store all relevant details. As explained by Shrinivasan, et al^[2] to keep track of the data exploration process and insights, visual analytics systems need to offer history tracking and knowledge externalization to the analyst. This will reduce the cognitive overload imposed on the analyst and by freeing essential mental resources and offering a new perspective on the recorded information. According to North, et al^[3] provenance “has demonstrated great potential in becoming a foundation of the science of visual analytics”.

Visual analytics for systematic scientific analysis of large dataset has opened up a new era for criminal intelligence analysts to understand the process through which a crime or criminal situation has occurred. The project VALCRI^[4] aims to develop a semi-automated visual analytic system that will help find connections in criminal intelligence that often humans miss. It’s provenance recording system will keep track of analytical reasoning processes to minimize human errors. Significant amount of research have been carried out to capture system specific analytical provenance data into science, engineering and medical sectors. A generic platform will be beneficial to handle provenance information into different visual analytic systems.

So, this research work aims to contribute to following developments for criminal intelligence analysis:

1) A systematic qualitative research approach for outlining requirements to capture analytic provenance data and designing underlying system to capture those into large complex visual analytic system for criminal intelligence analysis.

2) Visual analytic techniques to utilize captured analytic provenance data for insight generation.

2 Related work

Significant amount of research have been carried out for developing a usable and manageable provenance tracker along with the user interface for representation, access to provenance information. We describe and summarize few of this research under three different areas to focus on. ① Capturing provenance of analysis process. ② Visualizing captured information. ③ Utilizing visualized provenance.

2.1 Provenance capture and management

Groth, et al^[5] developed an implementation-independent protocol for recording of provenance. They described the protocol in the context of a service-oriented architecture and formalize the entities involved using an abstract state machine or a three-dimensional state transition diagram. To track events processing stream provenance for workflow driven system, Vijayakumar, et al^[6] described an information model and architecture for stream provenance capture, collection and evaluated the provenance service for perturbation and scalability for the LEAD (Linked Environments for Atmospheric Discovery) project. Prov4J is a semantic Web framework for generic provenance, proposed by Freitas, et al^[7]. This work described a framework which used Semantic Web tools and standards to address the core challenges in the construction of a generic provenance management system. The work also discussed key software engineering aspects for provenance capture and consumption and analyzed

the suitability of the framework under the deployment of a real-world scenario. Problem of systematically capturing and managing provenance for computational tasks have been receiving significant attention because of its relevance to a wide range of domains and applications. Freire, et al^[8] have presented a survey on concepts related to provenance management for computational tasks, so that potential users can make informed decisions when selecting or designing a provenance solution.

2.2 Provenance visualization

GeoTime developed by Eccles, et al^[9] is a commercial geo-temporal event visualization tool that can capture a screen shot of the tool and perform text or graphical annotation. It also allows users to construct a report of the analysis. Tableau Public offers a story telling feature, which consists of several pages or story points, each is a captured visualization with annotation. To reuse captured states, the Human Terrain Visual Analytics system (HTVA) proposed by Walker, et al^[10] allows the analyst to drag and drop captured visualizations automatically onto an empty space and add narrative to each visualization to build the story. To visualize captured information, LifeLines developed by Plaisant, et al^[11] is a visualization for personal histories, which uses icons to indicate discrete events and thick horizontal lines for continuous ones. Typically, the system begins with an initial state (node). When the user performs an action, a new node is created for the current state, and a new edge is added to connect the previous node with the current node. VisTrails introduced by Bavoil, et al^[12] colour codes the background of visualization nodes according to when they are created and Aruvi introduced by Shrinivasan, et al^[2] uses the length of edges to represent the distance in terms of time between two states. For visualizing the reasoning process, the Scalable Reasoning System developed by Pike, et al^[13] provides a more formal method to document the reasoning process. A captured visual-

ization can be added into reasoning space to create a node as a miniature and can be tagged as an evidence artefact.

2.3 Provenance utilization

SensePath developed by Nguyen, et al^[14] is a tool for understanding sense making process through analytic provenance. SensePath provides four linked views of i.e, a timeline view that shows all captured sense making actions in temporal order, a browser view that displays the Web page where an action was performed, a replay view that shows the captured screen video & can automatically jump to the starting time of an action when it is selected in an-other view, a transcription view that displays detailed information of selected actions. “Vistories” introduced by Gratzl, et al^[15] is a visual stories based history exploration system by following the CLUE (Capture, Label, Understand, Explain) model. This tool has an authoring mode, a provenance graph view, a story view for showing the history of the analysis and a vistory being created.

3 Analytical provenance requirements

Analytical provenance is the means for providing insight into data processing operation in question. So, for criminal intelligence analysis it is one of the best means to provide necessary support to explain in a clear way how decisions or choices were made, what they were based on, how steps in a selection process were made, provide information grounds to justify and answer claims of bias or discrimination, and show compliance. All these are enabler of fairness and lawfulness of the data processing activities from the legal framework. Transparency in criminal intelligence analysis is an important requirement for maintaining respective LEP (Legal, Ethical, and Privacy) guidelines. This is the property that all operations on data including legal, technical and organizational setting and the correlating decisions based on the results can be understood and reconstructed at any time. So, Transpar-

ency can be regarded as the underlying foundation of the analytical provenance. As well as analytical activities performed by analysts should be recorded for supporting accountability for particular action of analysis process. Analytical provenance data has got greater influence in this regard.

Capturing analytical provenance has also got significant role in criminal intelligence analysis, because the legal directive foresees an obligation to provide competent legal authorities with information about the processing operation upon request. Competent authorities are any public authority or any other entrusted body by national law to exercise public authority and public powers for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security. Analytical provenance data can help to validate the processing operation in such case.

To have a better understanding of the requirements for analytical provenance in criminal intelligence analysis, we organized a focus group discussion with police analyst end users of the project VALCRI^[4]. Based on our initial understanding of capturing analytical provenance, we developed an analytical state capturing prototype and demonstrated to police analysts during the focus group. We adopted the technique proposed by Walker, et al. of saving analytical states as bookmarks for implementing our prototype. The purpose of such prototype demonstration in the focus group was to gather requirements for a much larger system as well as to evaluate the prototype. The focus group involved three groups of police analysts and each group had two people.

We tested two techniques of capturing analytical states by using our developed prototype - 1) Capturing a URI. 2) Capturing event properties to save and restore analytical states automatically. We also tested these techniques on two separate visualizations, using the Canadian Crimes by Cities

1998-2012 dataset for Geo-Spatial Temporal^[16] (GST) crime analysis and VAST Challenge 2015 dataset for Call Data Records (CDRs) analysis^[17]. This system automatically logs information about the user's interaction with system as well as saves corresponding state data into database and shows the preview of the analytical state at front-end along with meta information on tooltips by using which a captured state can be restored again. The event based approach out of these two techniques that we followed to develop our initial prototype, provided us better results for capturing analytical states even at a granular level.

3.1 System requirements (SysReqs)

Based on the prototype development experience and realization from the focus group demonstration, we identified following system requirements for supporting criminal intelligence analysis.

SysReq1 Different techniques should be supported for capturing and recording analytical provenance information.

SysReq2 A standard mechanism should be referred to the discovery of an analytic provenance state object and a representation model should be used.

SysReq3 Different levels of granularity should be used in describing analytical provenance of complex state objects.

SysReq4 Analytical provenance data needs to be stored, logged, and versioned to allow capturing of states.

SysReq5 The system needs to scale with large amounts of recorded analytical provenance data and lots of analyst end-users.

SysReq6 Analytical provenance information needs to be able to be easily queried.

SysReq7 Different levels of security are needed to provide access to analytical provenance data.

3.2 Police analyst requirements (AnaReqs)

The police analysts currently record their

thoughts in their diaries or spreadsheet manually and found this process cumbersome and ineffective. The police analysts found the demonstrated concept of analytical states capture and restore, and automatic state suggestion system could be effective for their work-flow. Based on the focus group we identified five potential end-users for an analytical provenance capturing system to support criminal intelligence. These include police analysts, analyst trainers, researchers, managers, and auditors. We now outline the identified requirements of the five-end users based on the focus group.

AnaReq1 Analysts need to see different representation techniques for visualizing analytical provenance data.

AnaReq2 Analysts need to be able to compare different analytical provenance information.

AnaReq3 Analysts need to validate whether captured analytical provenance information is of adequate quality for evidence.

AnaReq4 The provenance information needs to show whether laws, rules and regulations have been correctly adhered to.

AnaReq5 Analysts must be able to step-back and step-forward through the states they have captured in the past to see what actions they performed in the system.

AnaReq6 Analysts need to be able to record a set of macro states to perform a collection of operations on different sets of data. We also call this Repetitive Replicating Playback (RRP).

AnaReq7 Analysts need to be able to annotate provenance information about different states.

AnaReq8 Analysts (based on role) must be able to turn off automatic logging of the provenance capture method.

AnaReq9 Trainers should be able to use the system to train new analysts.

AnaReq10 Auditors should be able to use the system to examine the kinds of activities analysts are performing and to generate reports.

AnaReq11 Managers need to be able to monitor what their police analyst colleagues are working on and see summaries of information.

AnaReq12 Researchers need to be able to use the system in conjunction with analysts to understand how to effectively perform criminal intelligence analysis.

4 System design

To meet different system requirements (SysReqs), we progressed developing a widget named as PROV and proposed a protocol for it to capture analytical provenance data from AUI which was developed by following modular software design technique, consisting of many widgets and heterogeneous platforms. In modular architecture, functionalities are separated into independent, interchangeable modules such that each contains all necessities for its own execution for distinct purposes. The proposed protocol as shown in Figure 1 supports such system to generically capture/restore analytical provenance states or workflows both automatically and manually. The whole architecture has been divided into following functional sub-sections.

AUI Widgets The widgets are analyst's visual interface for their scientific computations mostly built using different Javascript libraries on Google Web Toolkit (GWT) framework by following MVP (Model, View, Presenter) design pattern. They have been integrated into shell presenter of the AUI system that inherits widget attachment information from an *Abstract Presenter*, so that attachment of widgets can be tracked at any time. Few groups of widgets support interactive cross-filtering among themselves for computation purpose as shown in Figure 1.

Data Channel The AUI system has been built by using Errai GWT-based framework for supporting uniform, asynchronous messaging services across the client and server end through its Remote Procedure Call (RPC) service. The data

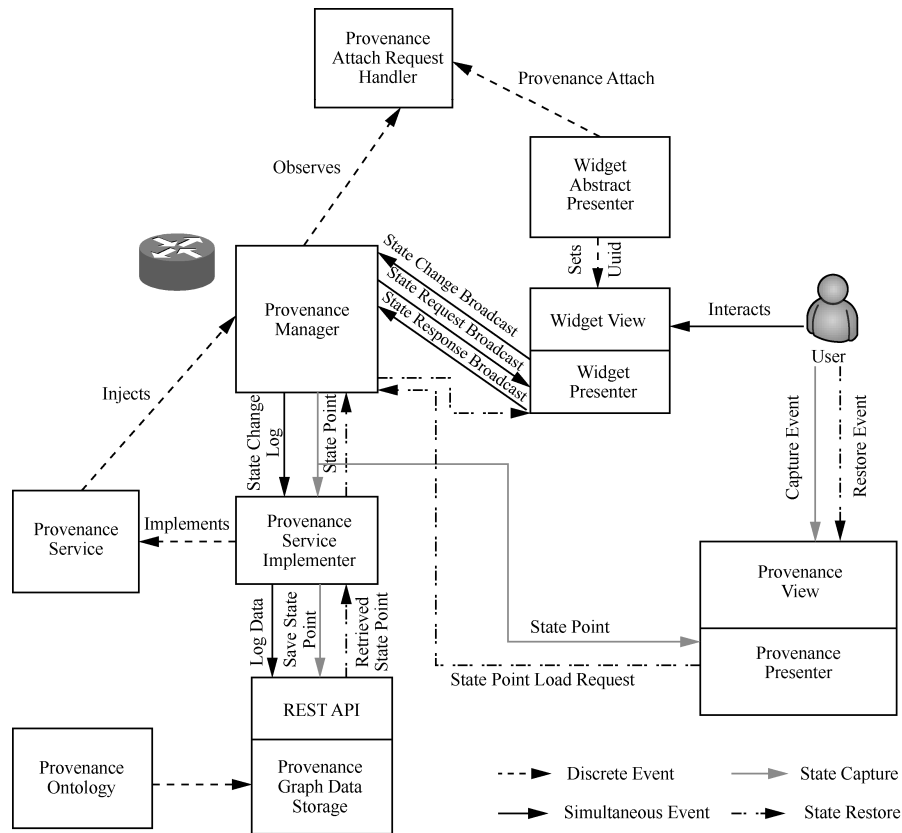


Figure 1 data flow diagram of the provenance visualization system for analyst's user interface (AUI)

channel is the presenter of messages generated by the interactions during the analysis process. These messages are consisting of two types of data i.e, metadata (MDATA) generated upon user's interactions and statedata (SDATA) are accumulated states data of different widgets after interacting.

Provenance Service As shown in Figure 1, provenance service is the middle-tier server which co-ordinates with the tier-1 requests from clients and tier-3 data storage system. Provenance Service has got two vital roles: 1) Provenance Service Implementer. 2) Provenance Manager.

Provenance Service Implementer SAVE/QUERY SERVICES for provenance data i.e, log data and state point information as shown in Figure 1, are implemented by this role player into our system.

Provenance Manager While user interactions on AUI widgets occur, the interacted widgets initiate provenance service by broadcasting

Statechange message to Provenance Manager. A Staterequest message is broadcasted by the provenance manager to receive state & state change information from different widget presenters through a stateresponse broadcast message. This is how provenance manager becomes aware of the state changes of AUI system. Not only state changes but also provenance manager observes attachment requests into provenance system from different widgets through a request handler so that it can provide information on demand. These are all discrete events not dependent on user interactions as shown into Figure 1.

State Point Capture The analyst fires an event for capturing his/her intended analysis state. The most recent state point received into Provenance Presenter from Provenance Manager gets saved into data storage and creates an image as state point preview into provenance view prov as shown in Figure 2.

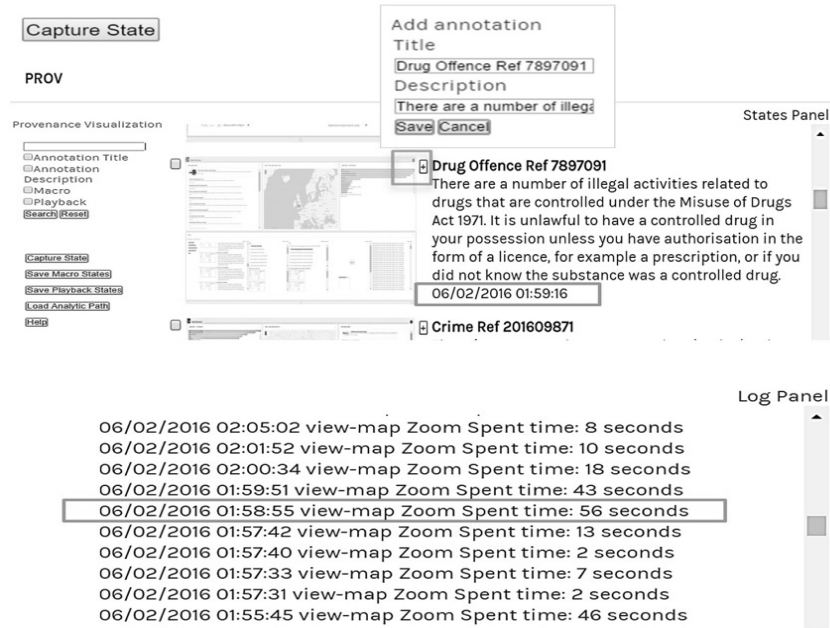


Figure 2 Manually captured states panel with annotation add/edit and an automatic log panel for analyst's user interface (AUI)

State Point Restore The analyst clicks on the state point preview to restore his/her previous analysis. A State Point loadrequest with its corresponding id is sent to After receiving enquired state from data storage, Provenance Manager broadcasts this as a stateprevious message (Figure 1) so that it is received by the widget to restore the analysis state back to the analyst.

Provenance Data-Storage All provenance data are currently stored into and queried from virtuoso universal server as RDF graph data format by using our developed REST (Representational State Transfer) API for the AUI system. Currently provenance data is stored along with our preliminary version of developed analytical provenance ontology.

5 Provenance visualizations

The PROV system supports police intelligence analysts (end-users) requirements (AnaReqs) for their visual judgmental process during crime analysis. This system has several interactive visual panels for captured analytical states representation, data querying in multi-ways and saving/replaying different workflows. Following sub-sections describe

these visualizations which have been developed by using our proposed data manipulating protocol to query/access database and event based technique to capture different analytical states.

5.1 States representation

To meet police analyst requirement AnaReq1, our developed provenance visualization system can capture different analytical states of AUI (Analyst's User Interface) and saves them as snapshots to show their previews (Figure 2). As well as to meet AnaReq7 and AnaReq8, currently annotations can be added and viewed again on tooltips upon interactions with saved analytic states (Figure 2). Provenance data can be captured either manually by the analysts or automatically by the system as log.

5.2 Repetitive replicating playback (RRP) system

The gathered requirements from police analyst end users as described into AnaReq2 and AnaReq6 were to record step-by-step analytical processes and use them for different other dataset to compare results. During the focus group that we arranged, they said "...we would like to be able to record a number of actions if some tasks are more standardized. (start/stop recording). Therefore want to use the

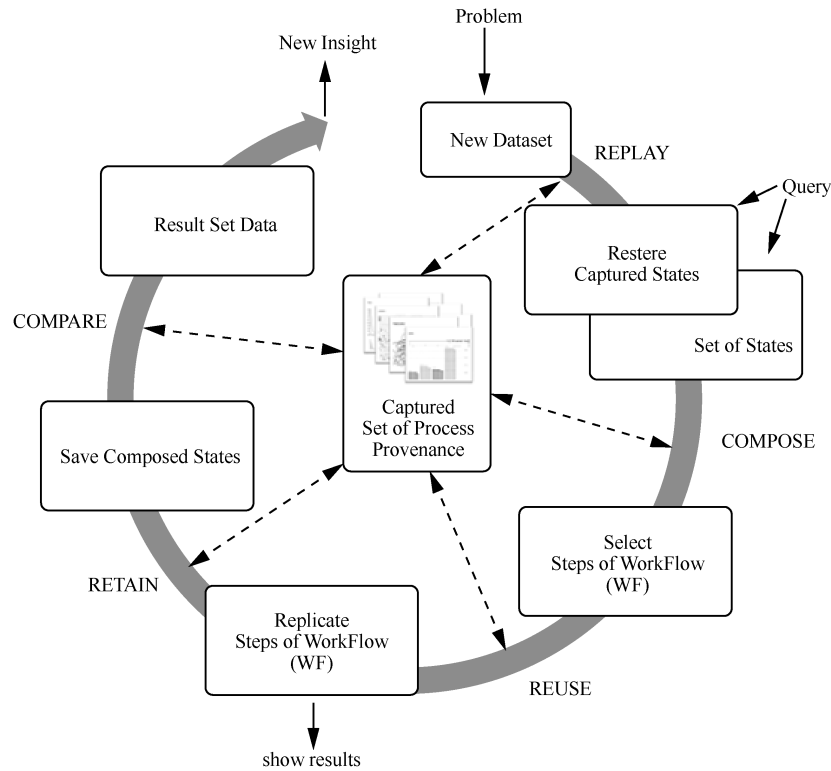


Figure 3 The RRP Cycle

same tasks/ask the same questions for another case with similar data...”. To develop such record/replay system we propose a model as shown in Figure 3 to support of analyst’s requirements. We call this Repetitive Replicating Playback (RRP) system as shown in Figure 4.

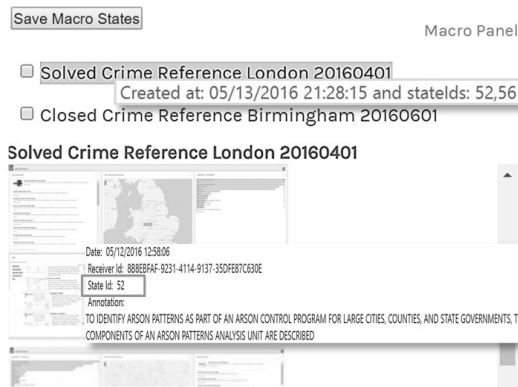


Figure 4 Repetitive Replicating Playback (RRP) System shows results with source state id information after running macro operation on saved group of states

The Analyst’s User Interface (AUI) is consisting of many widgets developed under heterogene-

ous platforms. Therefore challenges are there to find out methods while reapplying insights to a new dataset into such environment.

North et al^[3] identified that analysts often utilize multiple tools simultaneously which renders the use of existing methods inadequate. Our proposed RRP supports this problem very well for recording/replaying WorkFlow (WF) in a heterogeneous environment. The RRP cycle is consisting of replay, compose, reuse, retain and compare steps. We tested this model by implementing into AUI for replicating WFs on different set of crime data and compare result set with selected previous states to gain new insight. After “Replaying” each RRP state (consisting of previously captured group of states), we find automatically captured states (as result) to “Compare” with their corresponding previous states. Composition of such RRP states can be modified by adding, deleting or reshuffling all of their contained previously captured analytic states. We can “Compose” new RRP state by making selections from

States and RRP panels and save (“Retain”) them for future use.

One of the challenges to “Reuse” captured analytical states is to be able to formulate queries that retrieve and employ traces in order to fulfill an analyst’s information needs in a user-friendly way. It involves formulating ad-hoc traceability queries, allowing interactive filtering of retrieved analytic states and ad-hoc query refinement. We visualized all RRP states (consisting of previously captured analytic states) on a timeglider as a group of states organized in a temporal order (Figure 5). States can be searched/filtered by types and/or users. As well as states sequences can be represented (highlighted in yellow colour) as shown in Figure 5 to show the analysis steps temporally. Also states can be traced back by using temporal information (gliding the timeline or using calendar). Our developed RRP (Repetitive Replicating Playback) system includes W3C PROV-AQ: Provenance Access and Query standard^[18] factors i.e., Recording – represent, denote; Querying – identify, pingback; Accessibility – locate, retrieve.

5.3 Analytic path

To understand the relationships among reasoning steps we implemented “Analytic Path” (Figure 6)

as a tool for visualizing analyst’s activities through interactions with the visualizations. Intelligence analysis is not practiced exclusively as a solitary activity. So, in a collaborative environment of criminal intelligence analytic provenance can add considerable value, where it must be communicated and shared among teams. Additionally, by allowing communication and sharing of information, visual representations of analytic provenance data will support analyst’s ability to identify and work with the desired information. So far the application of analytic provenance system supports sensemaking for individuals. In case of more than one analyst working together for a specific problem, automatically recorded interactions can help to understand their thinking processes.

The RRP (repetitive replicating playback) panel supports to create a composition of captured analytic logical states which can be applied back again on different other scenarios. All such captured visual analytic states can be replayed back again and visualized as a colour coded users’ actions network known as “Analytic Path” to show analysts’ higher level subtasks (Gotz, et al^[19]) through low level action sequences i.e., $Seq.\#001 A \rightarrow B \rightarrow D \rightarrow E \rightarrow G$,

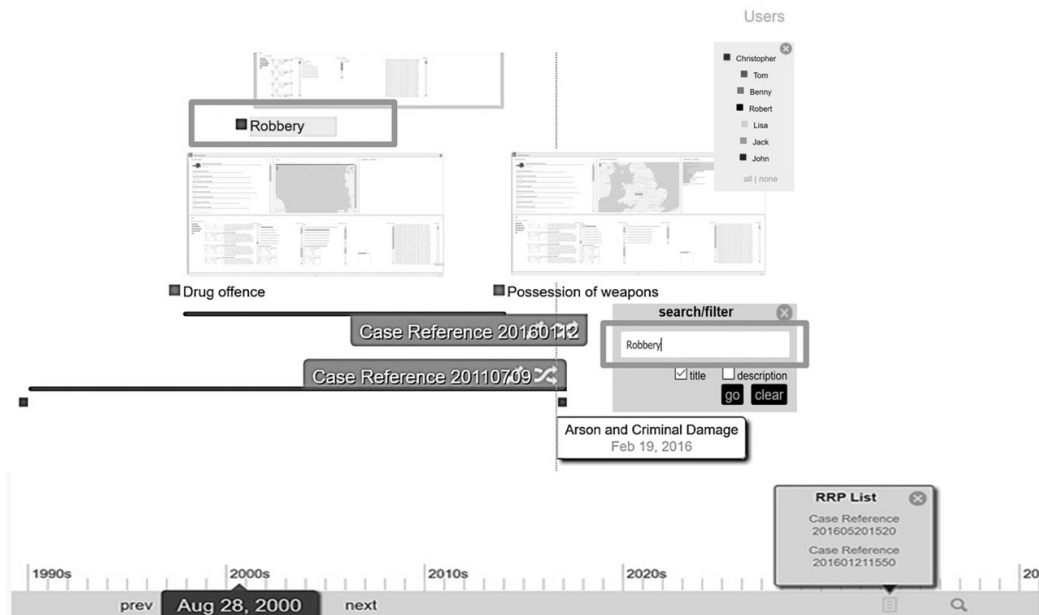


Figure 5 Visual representation of saved RRP states into macro panel to trace states back by time gliding, colour coded users (analysts) filtering, keyword searching and selecting from RRP list

Seq. #002 $A \rightarrow C \rightarrow I, \dots$, *Seq. #N* as shown in Figure 5.

The incidence nodes of “Analytic Path” network can be computed by following our proposed compositionally reductive framework for the contextual information of complex analysis. To illustrate the idea, let’s assume $P(S)$ is a semantic state composition function $P(S)$, where S is an analytic state.

So, $P(S) = S$.

For the *Seq. #001*, it can expressed as –

$$P(S_A) = S_A$$

$$P(S_B) = S_B$$

$$P(S_D) = S_D$$

...

$$P(S_n) = S_n, \text{ where } n \text{ is the number of nodes.}$$

Thus we computed the n th state S_n as P :

$$S_{A,B,D,\dots,n-1} \rightarrow S_n.$$

Composition function of different analytic states can be expressed as

$$P(S_A) \circ P(S_B) = P \circ P(S_A, S_B) = \{S_A, S_B\} = S_{A,B}$$

$$P: S_A \rightarrow S_B$$

$$P(S_B) \circ P(S_D) = P \circ P(S_B, S_D) = \{S_B, S_D\} = S_{B,D}$$

$$P: S_B \rightarrow S_D$$

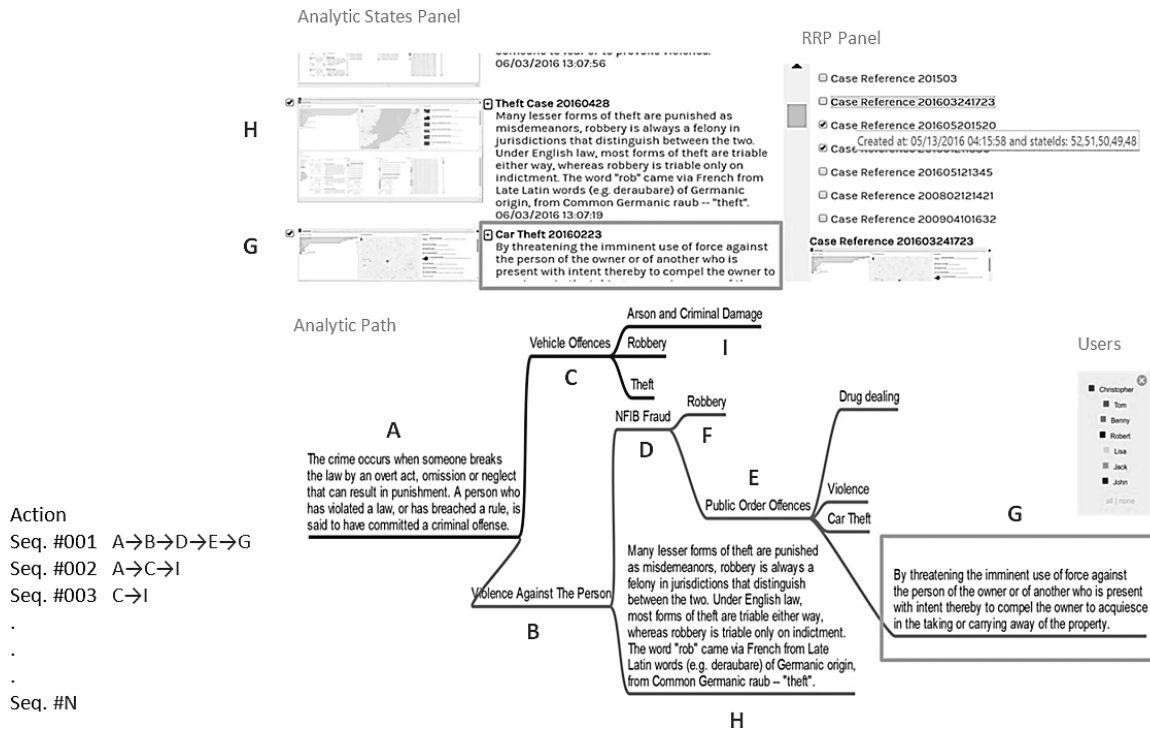
...

$$P \circ P(S_{A,B,D,\dots,n-1}, S_n) = \{S_A, S_B, \dots, S_n\}$$

$P: S_{A,B,D,\dots,n-1} \rightarrow S_n = S_{ST}$, where S_{ST} is a Sub-Task State (Gotz, et al^[19] 2008) through low level actions or events. All other low level action sequences *Seq. #002*, *Seq. #003*, ..., *Seq. #N* can be computed in the same way.

The tool “Analytic Path” supports saving of mapped analytical states into and loading back from data storage. It allows combining such multiple maps together to make a visual story of the group analysis process. It also supports adding, deleting, editing or rearranging different branches with users’ colour codes, consisting of annotations set by analysts along with captured states. This is known as “Schematization” of the reasoning process as shown in Figure 7.

Schematization is the process of organizing findings in some way that can trigger new insight. Pirolli & Card^[20] suggest the need for tools to facilitate the schematization such as visualizing findings based on their temporal or spatial information. Ac-



cording to their sense making model, analysts seek information, search and filter for more relevant one, read and extract evidential information, and organize it into some schema. However, we find that the sense making loop is well elaborated through different sense making activities in the “Data-Frame Model” proposed by Klein, et al^[21]. These sensemaking activities are connect data to a frame, elaborate a frame, question a frame, preserve a frame, and reframe. According to data-frame model’s terminology, the analyst tries to match some data to create an initial frame. When encountering new data, the analyst can either add it to the frame to elaborate the frame (if it fits to the frame) or remove existing data (if it cannot fit the frame any more).

The analyst starts questioning the frame when they detect inconsistencies between data, or poor quality data in the frame. Then, they need to decide

between preserving the frame by looking for more data, or reframing it by comparing it with other frames, or seeking a completely new frame (Klein et al.^[21]). Our developed analytic Path supports schematization in a visuo-spatially manner for iterative and dynamic nature. Takken, et al (2013) found that when people directly manipulate data, for example, by moving individual pieces of information to create temporary groups or sequences, or eliminating pieces of information from a group; this can enhance their sense-making and analytical reasoning ability by helping them discover new explanatory relationships created by the rearranged pieces of information. They have named such technique as “Tactile Reasoning” which is an interaction technique that supports analytical reasoning by direct manipulation of information objects into graphical user interface (GUI).

We have developed Analytic Path (Figure 6) in

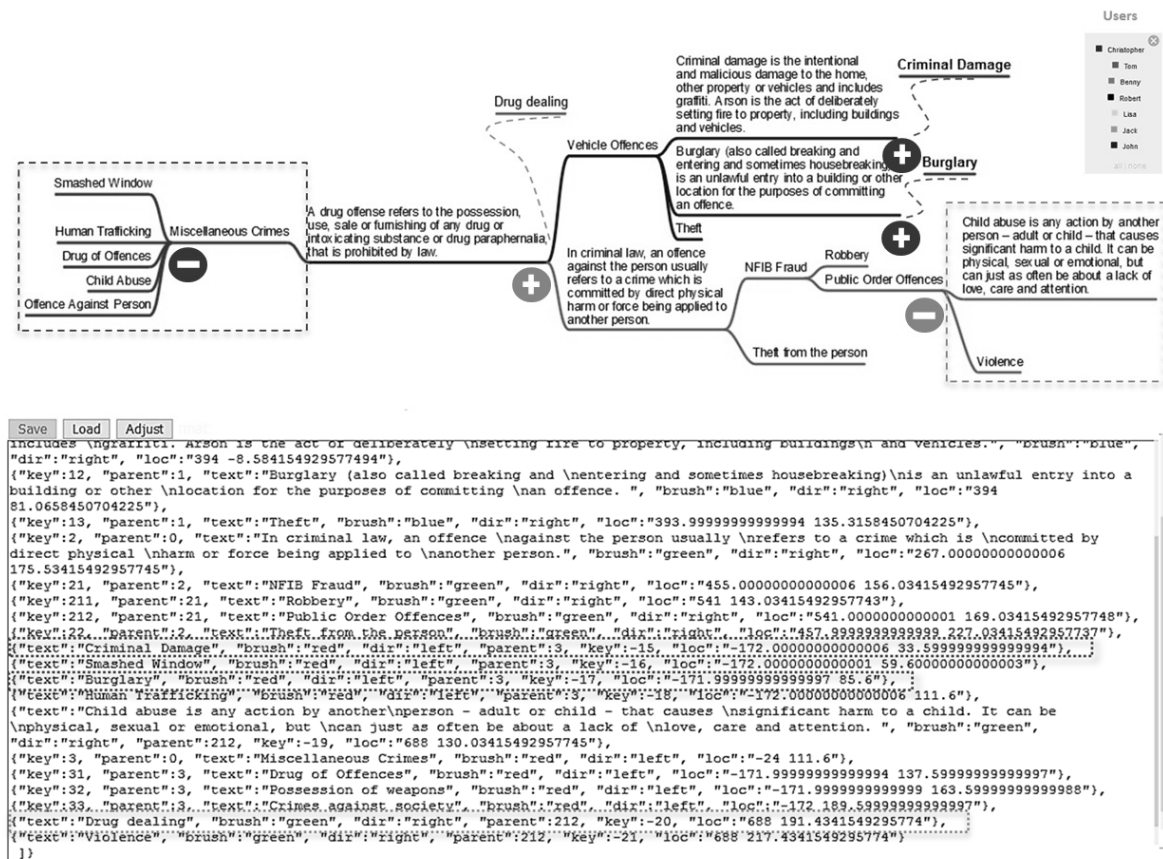


Figure 7 Schematization of Analytic Path in a visuo-spatially manner.

a visuo-spatial manner is because it enhances the ability to process and interpret visual information about where data objects are in space and overcomes the shortcomings of existing tools to visually represent and support schematization in a collaborative environment. We visualized all of these analysts' reasoning provenance information (annotations) as bookmarks on a time glider to provide their temporal information (Figure 5). The colour coded lines help to distinguish each analyst's reasoning contribution. As shown in Figure 7, analysts can integrate   and disintegrate   data to create a New Frame. They can drag the data and add it to an existing frame (Elaborate Frame). As well as they can disintegrate a piece of annotation data and drop it onto the void space (Preserving a Frame). The Dotted Rectangles into Figure 7 shows the Parent Nodes of data pieces "Drug Dealing", "Criminal Damage" and "Burglary" into Analysis Space and more about their spatial information into "Data Space". As well as "Dotted Curves" into analysis space shows their new nodes to be added with. As well as analysts can release an existing frame onto void space to make all it's attached pieces of data free for rearrangement to add/delete new/existing pieces of annotation data. This how if the analyst thinks about any frame that it is completely wrong, he/she can construct a new frame (Reframing). The analysis space can panned, zoomed in/out and adjusted to scale and view the whole analysis in detail by using keyboard short keys. As well as schematization of annotation data can be saved into data space.

6 Evaluation

We conducted an evaluation with our police analyst end-users to elicit subjective feedback on our prototype. We wanted to evaluate how the provenance visualizations support analysis and reasoning about data for deriving relevant knowledge in criminal intelligence. The evaluation involved

qualitative focus groups. We had three groups of analysts who participated in pairs. Each pair was from a different police organization. The procedure of the focus group involved demonstrating the prototype, illustrating the visualizations for different tasks, and obtaining feedback. Each group had 30 minutes for the demonstration and feedback. We had separate observers during the focus groups that recorded notes, ideas, and feedback from the end-users. We now report on the feedback as recorded by the observers, based on five questions as follows.

Question 1: What is the purpose and value of the component?

P1: The goal of the tool is to track the analyst's work and save what has been done enabling the later retrieval. It will capture the data from the interface and information on the users' interactions. The analyst will be able to "bookmark" a particular stage in the process and save it as part of the provenance record.

P2: Track and capture what has actual been done. Bookmark and save as a provenance component The idea is to be able to capture data from the interface and user interactions.

P3: Capture different states in time. Retrace earlier states system and user provenance.

P4: Capturing analyst's use of the system. Capturing anything happening from analyst's point of view and data point of view. Provenance refers to a historical record of the process taken or employed. Tool offers a capture mechanism.

Question 2: Is the purpose and value clear to End Users?

P1: Yes. The end users confirmed it is useful for the analyst to save the current state and be able to get back to it. Also, it is helpful if the analyst can automatically save the progress and get back to states they didn't save. Finally, the tool would allow the supervisor to trace the reasoning of the analyst.

P3: Once conceptual architecture, was displayed and explained, the Users really understood

the component.

Question 3: What do End Users like or dislike about the component?

P1: The end users enjoyed the ability to get back to a certain significant point in time. They are interested in process playback not typical video playback (i.e. the re-run of the process which provides clear understanding of its various stages and components, not an overall view of what has happened and when).

P2: Feel that it could be of use to them as they don't do something like this already. Able to go back through the provenance log to a specific point and save is good. Would like to be able to record a number of actions if some tasks are more standardized. (start/stop recording). Therefore want to use the same tasks/ask the same questions for another case with similar data (i.e. Macros). Want to be able to see previous action recommendation system could be useful.

P3: Generally very useful. Particularly if you have several pieces of work in progress and you may not work on one for 2/3 weeks and have to come back to it. "Adding notation would be brilliant". Good basis for training for inexperienced analysts. Being able to pause the training and show what an experienced analyst would do. Ability to find new information and go back to and older state and add the information. The ability to demonstrate that you have previously saved states that are similar.

P4: Analysts have no way of capturing the data except in log-book. This is difficult to do given workload so tool is very helpful to this end. Like the ability to annotate Guides Tool offers a good basis for training; follow in the footsteps of an analyst using the provenance tools. Like the tool they can reconstruct the path the analyst took over time.

Question 4: What features or functionalities would End Users like added, changed or removed?

P1: The end users suggested the recoding of the set of actions, not states, and then applying it to

a new case in the same data set, the feature resembling macros in Excel.

P2: Would like to be able to record a number of actions if some tasks are more standardized. (start/stop recording). Therefore want to use the same tasks/ask the same questions for another case with similar data (i.e. Macros). Want to be able to see previous action recommendation system could be useful.

P3: Ability to "play" through the saved states

P4: Analysts log in for system of auditing and performance.

Question 5: Overall, is the End User group's assessment positive, negative or neutral?

P1: The assessment is positive. The ability to record the macro would be useful for repetitive tasks. The ability to review the history of the analytic process would provide useful insight into the evolving judgment.

P2: If you're doing the same task everyday then a Macro would be useful, but Guy wouldn't find the history side of things as useful as Mark would. Essentially a hypothesis flow Playback definition – process playback – re-run the process.

P3: Overall, very positive.

P4: Positive

6.1 Purpose and value

All the end-users understood the purpose of the prototype to log and track analytic workflow in the VALCRI^[4] system. They claimed that this task would add value to what their current workflow processes are, as it would allow them to track what they were doing on a daily basis and analyze what they had done previously. They deemed these tasks to add value, as it is necessary to explore different analytic pathways, or to even pick up and validate the work of others.

6.2 Strengths and weaknesses

The biggest strength as reported by all the end-users was that the tool tracked the tasks they were performing as well as the ability to bookmark

certain parts of the interface they were working with. The tracking and book-marking feature was found to be useful as they could come back to a previous state where they had been working and continue to work from that state.

6.3 Improvements

Different features were suggested by some of the analysts for addition to the prototype. They would like to see a team leader login part, which can monitor the activities of all analysts. The purpose for doing so is they can see at what stage an analyst is working on within a crime investigation and to get reporting features based on the progress of analysts. They would also like to be able to add outcome reports to different stages of the analytical path. Being able to summarize information through annotations and free text will enable analysts to record some of their thoughts when investigating a crime.

6.4 Satisfaction

The overall assessment of the prototype by all the end-users was very positive and they were satisfied with the progress of the prototype. All the analysts felt that the different provenance features could add value to what they are currently doing and to help make more effective decision making for criminal intelligence analysis.

7 Discussion and conclusion

The key to this research on analytic provenance is the belief that by capturing user's interactions with a visual interface, some aspects of the transparency of user's reasoning processes can be retrieved. To correlate analyst's interactions with the visualizations for his/her reasoning process, the analytic provenance research needs to start with the understanding of how information is perceived by the user. We conducted a focus group discussion meeting with the police analysts to understand their needs for analytic provenance visualization. As the user interacts with visualization, the series of inter-

actions can be considered as a linear sequence of actions. So, how can these analytic provenance information be captured — is still an open challenge. We have implemented our proposed protocol for managing huge analytic provenance dataflow for a large complex system like Analyst's User Interface (AUI) of the project VALCRI^[4]. Once the user's provenance data has been captured, the challenge becomes making sense of the provenance. As noted by Jankun-Kelly et al.^[22] history alone is not sufficient for analyzing the analytical process with visualization tools. Often, there are relationships between the results and other elements of the analysis process which are vital to understanding analytic provenance. Our provenance visualization system can also capture analytical relationships automatically. We have developed an analytic process mapping system named as "Analytic Path" to visualize those related process sequences for multiple analysts working in a group. One of the research goals in analytic provenance is to be able to automatically reapply a user's insights to a new data or domain. It refers to the utilization of specific knowledge of previously experienced, concrete problem situations or cases. By employing such repetitive process, the analyst can solve a new problem by finding a similar past case, and reuse it in the new problem situation. We have developed a Repetitive Replicating Playback (RRP) system, where analysts can use their previously saved group of analytic states, apply to new dataset and see the results. We have tested our proposed way of capturing event-driven analytical provenance by developing visualization prototypes based on police intelligence analysts' requirements and found it supports the challenges of five interrelated stages of analytic provenance generically, as suggested by North, et al^[3] i.e., perceive, capture, encode, recover and reuse.

According to Gotz, et al's^[19] hierarchy of analytic behaviour, the sub-tasks at higher-level have more concrete states with rich semantics into prov-

enance-aware analytic process comprising of interactions for understanding human intention and computational elicitation. Semantics of interactions that occur during switching among multiple visualizations hasn't been addressed into this work. Also this work hasn't addressed the coupling between cognition and computation through interactions during analytic processes. As well as for sensemaking or computational problem solving during crime analysis in criminal intelligence and the analytic processes, require insightful alignment with the visualizations for supporting analyst's thought processes. The current developed visualizations have got limited support in this regard, which we are still working on.

Our future endeavour for this work is to add few more features with our current system i.e., creating case specific new provenance capturing space with pluggable annotation system and tag them. Also developing a document trail system by using attached crime reports with the annotations will be useful as identified by the police intelligence analysts. We also aim to develop an ontology for analytical provenance, which is currently absent into W3C standard^[18] for describing and integrating analytical states from different sources. As well as visualizing evolution of ontology is a crucial issue to understand the way knowledge evolves from one state to another during analyst's analytical process and that is a potential area of research.

Acknowledgement

The research results reported here has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) through Project VALCRI, European Commission Grant Agreement N° FP7-IP-608142, awarded to B.L. William Wong, Middlesex University and partners.

References:

[1] WONG B L W, XU K, ATTFIELD S. Provenance for intelligence

- analysis using visual analytics[C]//Workshop on Analytic Provenance. 2011.
- [2] SHRINIVASAN Y B, WIJK J J V. Supporting the analytical reasoning process in information visualization[C]// The Twenty-Sixth Sigchi Conference on Human Factors in Computing Systems. 2008: 1237-1246.
- [3] NORTH C, CHANG R, ENDERT A, et al. Analytic provenance: process+interaction+insight[C]//The 11th Extended Abstracts on Human Factors in Computing Systems. 2011:33-36.
- [4] WONG B L W, ZHANG L, SHEPHERD I D H. VALCRI: addressing european needs for information exploitation of Large Complex Data in Criminal Intelligence Analysis[C]//European Data Forum. 2014: 19-20.
- [5] GROTH P, LUCK M, MOREAU L. A protocol for recording provenance in service-oriented grids[C]//International Conference on Principles of Distributed Systems. 2004:124-139.
- [6] VIJAYAKUMAR N N, PLALE B. Tracking stream provenance in complex event processing systems for workflow-driven computing[C]//Copyright VLDB Endowment. 2007:23-28.
- [7] FREITAS A, LEGENDRE A, O'RIAIN S, et al. Prov4J: a semantic web framework for generic provenance management[C]//International Workshop on Role of Semantic Web in Provenance Management. 2010.
- [8] FREIRE J, KOOP D, SANTOS E, et al. Provenance for computational tasks: a survey[J]. Computing in Science & Engineering, 2008, 10(3): 11-21.
- [9] ECCLES R, KAPLER T, HARPER R, et al. Stories in GeoTime[J]. Information Visualization, 2007, 7(1):3-17.
- [10] WALKER R, SLINGSBY A, DYKES J, et al. An extensible framework for provenance in human terrain visual analytics[J]. IEEE Transactions on Visualization & Computer Graphics, 2013, 19(12): 2139-2148.
- [11] PLAISANT C, MUSHLIN R, SNYDER A, et al. LifeLines: using visualization to enhance navigation and analysis of patient records[J]. Proc Amia Symp, 1998, 5(1):76-80.
- [12] BAVOIL L, CALLAHAN S P, SCHEIDEGGER C E, et al. VisTrails: enabling interactive multiple-view visualizations[C]//Visualization. 2005: 135-142.
- [13] PIKE W A, BRUCE J, BADDELEY B, et al. The scalable reasoning system: lightweight visualization for distributed analytics[J]. Information Visualization, 2009, 8(1): 71-84.
- [14] NGUYEN P H, XU K, WHEAT A, et al. SensePath: understanding the sensemaking process through analytic provenance[J]. IEEE Transactions on Visualization & Computer Graphics, 2015, 22(1): 41-50.
- [15] GRATZL S, LEX A, GEHLENBORG N, et al. From visual exploration to storytelling and back again[C]// Eurographics / IEEE Vgic Conference on Visualization. Eurographics Association, 2016: 491-500.
- [16] IISLAM J, ANSLOW C, XU K, et al. Towards analytical provenance visualization for criminal intelligence analysis[C]// Conference on Computer Graphics & Visual Computing. Eurographics Association. 2016:17-24.
- [17] WHITING M, COOK K, GRINSTEIN G, et al. VAST Challenge 2015: mayhem at dinofun world[C]// Visual Analytics Science and

Technology. 2015:113-118.

- [18] MOREAU L, HARTIG O, SIMMHAN Y, et al. PROV-AQ: Provenance Access and Query[EB/OL]. <https://www.w3.org/TR/prov-aq/>.
- [19] GOTZ D, ZHOU M X. Characterizing users' visual analytic activity for insight provenance[C]//Visual Analytics Science and Technology.2008:123 - 130.
- [20] PIROLLI P, CARD S. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis[C]//International Conference on Intelligence Analysis. 2005.
- [21] KLEIN G, PHILLIPS J K, RALL E L, et al. A data-frame theory of sensemaking[C]//The Sixth International Conference on Naturalistic Decision Making. 2007:113-155.
- [22] JANKUNKELLY T J, MA K L, GERTZ M. A model and framework for visualization exploration[J]. IEEE Trans Vis Comput Graph, 2007, 13(2):357-369.

About the authors



Junayed Islam is a PhD student at Middlesex University London. His research interests include big data, visualization and visual analytics. Main focus of his current work is to design and develop UI for reconstructing past criminal situation by combining geospatial and temporal visualization techniques, network visualization tools, with argument and narrative structuring techniques to formulate plausible explanations in criminal intelligence.



Kai Xu is an Associate Professor in Data Analytics at Middlesex University London. He has over 15 year experience in data visualization and analytics research in both the academic and industry context. He has extensive experience working with the UK government departments and leading defence companies in data analytics projects. His work has won a few international data visualization awards. More details are here: <https://kaixu.me/>.



William Wong is Professor of Human-Computer Interaction and Head, Interaction Design Centre at Middlesex University London. His research interest is in the representation design of information and the interaction with user interfaces that support decision making in complex dynamic environments. He uses a cognitive engineering approach to understand the nature of expertise and to model the nature of cognitive work in order to design better systems. He is currently investigating the problems of visual analytics in sense making domains such as intelligence analysis, financial systemic risk analysis, and low literacy users. He has received over USD \$25 mil in research grants and together with his students and colleagues have published over 100 refereed articles.