Is Multi-perspective Visualisation recommended for E-discovery Email Investigations?

Mithileysh Sathiyanarayanan
giCentre, City University London, UK
Mithileysh.Sathiyanarayanan@city.ac.uk

Cagatay Turkay
giCentre, City University London, UK
Cagatay.Turkay1@city.ac.uk

ABSTRACT

Problem Statement: To help improve efficiency and reduce costs involved in an electronic discovery[1] (E-discovery) process for email investigations, visualisations can be of great help, and they can change the way analysts/investigators understand contacts, messages in inboxes and their relationship. Though email data is a central resource in E-discovery processes [1,2] but the existing tools such as JigSaw, INSPIRE and DocuBurst are not capable of handling this dynamic, heterogeneous and relational data. As the socio-technical systems have grown in complexity, E-discovery analysts who are not that tech-savvy are looking for a simple and effective visualisation tool to detect, analyse and understand anomaly behaviours in email communication. This project is in close collaboration with the Redsift Limited London who are currently working on E-discovery related projects.

Case Study: Enron [3] scam is a well-known case in the data visualisation field. Enron produced fake profit reports and company's accounts which led to bankruptcy. Most of the top executives were involved in the scam, as they sold their company stock prior to the company's downfall. The Enron email is available for the public to access. In our work, we will be using the Enron data as a test case for designing and user-testing.

Workshop: We conducted couple of workshops to understand analysts requirements. The first workshop was with a legal team of six solicitors in Bangalore, India. They use Excel as a tool for their investigations. They liked the simple visualisations but found the manual search and data arrangements strenuous. The second workshop was with an intelligence analyst who works at the cyber investigation department, Bangalore, India. He uses E-discovery tools such as Jigsaw, Concordance by LexisNexis and/or IN-SPIRE to analyse unstructured data. He finds the visualisations to be complex and difficult to understand.

Workshop Suggestions: The five-point visualisation features summarised for E-discovery email investigation are:

1. *Multi-faceted:* representation must be supported with a multi-faceted search feature to display various granularities.
2. *Multi-modality:* representation must include temporal behaviours, individuals' action, connections and text/topic responses.
3. *Multi-level:* representation must have a drill-down approach (multiple levels) to filter and sort the data based on the multi-modality and present with some visual cues about what to consider and what not to (investigation cueing).
4. *Multi-aggregation:* representation must be systematically organised based on the multiple aggregations from the higher level (top-level) to all the consecutive levels which helps in building visual summaries that can be presented legally.
5. *Multi-juxtaposition:* representations must be effective for displaying multiple relationships and comparison when placed close together or side by side.

Proposed Solution: Based on the workshop suggestions and the limitations of the current tools that generate email visualisations, we propose a multi-perspective approach (shown in the Figure 1) that will generate elementary (simple) and intelligible automated visual representations for displaying the most relevant information from the email data and aid in comparing two subsets of information.

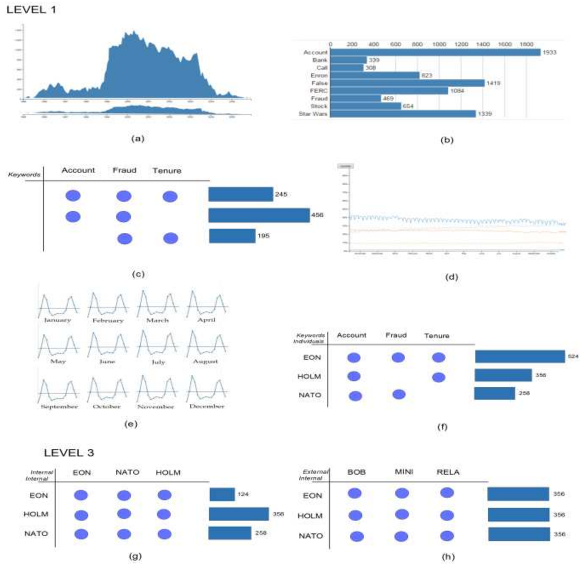

Figure 1: D3 Prototypes: multi-modal and multi-level design - a simple line, bar and matrix charts.

REFERENCES

[1] http://www.radicati.com/wp/wpcontent/uploads/2013/04/Email-Statistics-Report-2013-2017- Executive-Summary.pdf.

[2] D. Lawton and R. Stacey and G. Dodd. Uk home office. https://www.gov.uk/government/uploads/system/uploads/attachment data/file/394779/ediscovery-digital-forensic-investigations-3214.pdf, 2014.

[3] B. Klimt and Y. Yang. The enron corpus: A new dataset for email classification research. In Machine learning: ECML 2004, pages 217–226. Springer, 2004.CHI '00. ACM, New York, NY, 526-531.

[1] it is a process in which electronic data is sought, located, secured, and searched with an intent of using it as evidence in a civil or criminal legal case.