Visual Analysis of E-mail Communication to Support Digital Forensics & E-discovery Investigation in Organisations

MITHILEYSH SATHIYANARAYANAN

A thesis submitted in partial fulfilment of the requirements of the City, University of London for the degree of Doctor of Philosophy (PhD)

Major: Computer Science

Minors: Communication Science, Digital Investigation and Data Visualisation

October, 2020

City, University of London, UK

Supervisors: Dr Cagatay Turkay and Prof Jason Dykes

Acknowledgments

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible without the support and guidance I received from many good hearts.

I would like to say a very big thank you to Rahul Powar & Randal Pinto of Red Sift (project partners), a cyber security company in London, for their insightful comments and enlightening me with their corporate knowledge and supporting me with the latest technologies. A very special gratitude goes to Red Sift for funding the research project. It was fantastic to have the opportunity to work at their facilities - such a cracking place to work!

I would like to express a deep sense of gratitude and thank my supervisors Dr Cagatay & Prof. Jason Dykes for providing an excellent guidance and unbelievable support during the research work and made me realise the potential I have within. It is due to Dr Cagatay that I was introduced to the subject of Data Science and I thank him for encouraging me throughout the research process. My experience working with Dr Cagatay has been an invaluable asset to me, which would be beneficial throughout my career. Without his guidance and constant feedback this PhD would not have been achievable.

I would like to thank Dr Phong Nyugen, who supported me with the engineering aspect of developing the E-mail Visualisation tool and for insightful discussions. I would also like to extend my gratitude towards the members of the giCentre for their timely inputs and guidance. It is due to their constant support and encouragement that this project has been successful.

I would also like to thank my mother, father, mother-in-law, father-in-law and grand-mother for their precious advice and sincere support at all times. They have been a source of encouragement and inspiration throughout this project. I am also grateful to my other friends who have supported me along the way.

Finally, my thanks go to my wife, Dr Sharanya, for her everlasting love and care. There has been more than one occasion where she motivated me when I used to get mentally exhausted.

Abstract

The main aim of the research is to design and develop interactive visual solutions to explore the information in E-mail communication data to support E-discovery compliance in an organisation. The solutions intent to assist the world of digital forensics and investigations, which will enable users/analysts to explore, identify/find/discover interesting communication behaviour and characterise information of interest. In this research, we designed & developed software prototypes through a structured process of abstraction, design and testing, by using a well-known methodology called Design Study Methodology (DSM). We describe our analysis/approach through examples applied within the context of a real-world application domain. Doing so is intended to explore and answer a series of research questions in ways that will improve the role of visualisation in Digital Forensics and E-discovery investigations.

The work identified the knowledge gap, challenges, requirements and tasks in Digital Forensics and E-discovery involving the analysis of E-mail communication data from the unstructured interviews with the organisation domain experts and from the literature. We employed user-centered design (UCD) which involved iterative design process for 3 years and built several visual solutions based on the requirements and tasks. We evaluated the solutions by conducting an empirical study with the experts to understand E-discovery tasks, visual solutions and the interface that can help analyst, to investigate and navigate within communication data, to identify/find/discover various patterns, trends, anomalies and information that might be interesting/relevant to investigation. The solutions were deployed in the collaborator's E-mail platform.

Contents

List of Figures

Visual Analysis of E-mail Communication to Support Digital Forensics & E-discovery Investigation in Organisations

Mithileysh Sathiyanarayanan

October 22, 2020

Chapter 1

Introduction

The aim of the research is to design and develop interactive visual solutions to unravel the information in E-mail communication data to support E-discovery compliance in an organisation, that can assist digital forensics and E-discovery investigations, which will allow users/analysts to explore, identify/find/discover interesting communication behaviour and characterise information of interest.

The use of E-mail as a communication and information sharing medium in large, complex, globally distributed organisations is widespread; yet implications of its use in organisation compliance and its integration with E-discovery has not been properly assessed and developed [143]. The need to assess the role and use of visual analysis for the investigation of email collections is recognised by the industrial partner for this research, Red Sift London, UK (https://redsift.com/) - we had regular iterative interviews with the experts to address the challenges (mentioned in Chapter 4). To support the above argument a discussion of E-mail communication, visualisation and information in the context of Digital Forensics and E-discovery compliance is presented in the following. The limitations of the existing work and requirements are explored followed by a discussion of the research requirement. The research questions, aim and objectives are then posed.

1.1 Rationale

In today's socio-technical environment, electronic mail (E-mail) is still a major digital communication medium, especially in organisations, as data is a central resource in Digital Forensics and E-discovery compliance processes [143]. Electronic Discovery (E-discovery) [54, 55] is an investigation domain where electronic/digital communication data is sought, located, secured, and searched with an intent of using it as an evidence in a civil or criminal legal case. Digital Forensics (E-forensics) [54, 55] is also an investigation domain where electronic/digital devices are analysed in order to recover deliberately modified, deleted or hidden information/evidence and produce it as a legal evidence of a crime or unauthorised action that are in relation to the case.

E-mails collected in organisations are multi-faceted, dynamic, and complex data describing individuals, connections, content exchanges and time of messages exchanged. With E-mail traffic continuing to grow at 5% a year [1] in the business context more companies are now requiring cost effective solutions for Digital Forensics and E-discovery compliance involving E-mail data [143]. Despite the increased importance of Digital Forensics and E-discovery for organisations, it still remains a reactive procedure where, once a company is involved in litigation or receives a request for information, a legal firm is then appointed to review the E-mail archives to produce evidence [63]. Since the whole process is cumbersome, time-consuming and expensive [30], organisations have started working on "Digital Forensics and E-discovery"[63], to respond to regulatory or audit requirements, or, in government, to Freedom of Information Act (FOIA) requests.

The E-discovery compliance for E-mail communication within an organisation is still an under-researched topic, the tools and solutions currently available on the market are based on simple string/keyword search and legal firms manually review E-mails iteratively to come to a conclusion and build a legal case [63]. To do so, the compliance team in an organisation present E-discovery analysts a large archived E-mail dataset of all individual inboxes to manually comb through information in order to characterise information they need, expending large amounts of time, energy, effort and money in the process [63].

For example, characterising what temporal points look like (such as temporal gaps, holiday times) or characterising types of communication such as announcements or intense discussion helps in understanding and finding characteristics of individuals and any other relevant information related to the investigation. The E-discovery investigation results in significant costs for the company or in a number of cases some kind of settlement because they can't afford the costs of E-discovery [143].

Addressing E-discovery requests that involve E-mail data, which nowadays can easily go up to millions, is becoming a task that is unmanageably time consuming [30]. It is not yet clear how to characterise various information or groups of data objects from multiple perspectives, effortlessly and build a strong case [30]. The need for solutions has been highlighted in several papers [30][100] which will help analysts in their E-discovery tasks through interactive and visual analytics and lead to faster and effective processes [100].

Some of the reviewers/audience might have question marks whether E-mail data is still interesting to look into. We argue that E-mail data is a central resource in E-discovery processes [63] and the existing tools are not capable of handling this dynamic, heterogeneous and relational data. This sparks the interest to study and explore this field, as E-discovery tasks require an interactive visualisation solution that can help in better navigation, investigation and facilitate visual evidence to support legal cases.

The research on this topic is of great interest to not only E-discovery researchers [63] but to the wider research community where researchers formulate scientific principles and theories to gain insights about the processes to develop effective visualisation solutions. Since visualisation techniques are used in solving real-world problems in many applications areas, this trend will grow in the next few years, promising to make visualisation a key technology in tomorrow's market. This serves as a motivation to understand the role of visualisation in E-discovery and develop an effective visualisation solutions for analysts to present and understand large E-mail data collected over the years.

1.1.1 Motivation and Relevance

As discussed in the rationale, Digital Forensics, E-discovery and Compliance in an organisation plays an important role to take on any legal actions against individuals working in their organisation or legal actions against other organisations based on the communicated E-mails [143]. As an illustrative example, the real case study and real-world problem in an organisation within E-discovery compliance is explained below (in the blue box). E-discovery requests are mostly conducted by Compliance Officer, Freedom of Information (FoI) Officer, Legal Counsel (E-discovery/legal officer), Human Resource officer, and/or IT Director/Manager. These officers might have many reasons to commence the E-discovery process in an organisation. In any case, organisation must produce data and/or relevant information in a timely and complete manner when necessary during legal proceedings (includes both pretrial and trial), which is part of "Digital Forensics and E-discovery Compliant readiness" [63]. As a legal requirement, a company needs to have an audit process to determine it's E-discovery readiness and litigation preparedness (in our case, E-mail communication data). The results of this audit will provide a company with practices that create a simple and affordable way to quickly present evidences or required information for corporate and legal purposes.

In E-discovery litigations or investigations, practically every analyst/investigator finds a vast and semantically meaningful collection of data, that is uncategorised and unlabelled, in E-mail inboxes to investigate which makes it tedious and time-consuming to classify, identify and/or discover various information [30], eventually making it expensive for an organisation. The tools currently available on the market are based on simple string/keyword search and legal firms charge companies based on the volume of information produced by the search, which is then manually reviewed [63] intensely to find/identify/discover relevant information. Investigators search through E- mails, seek answers to various questions in the reports - who? what? when? - to produce it as an evidence to the judiciary [143]. This is an iterative process to confirm the obtained information, such that strong evidence is produced to build a concrete legal case [143].

Since the whole process is cumbersome to identify/discover various information and relationships between them, there is need of simplifying the investigation process by providing visual solutions [63] and from the preliminary discussions with the Red Sift experts (mentioned in the Appendix A.7). The organisations are in need of visualisations [63], as an E-discovery compliance solution, that can aid in investigating individuals and facilitating the generation of categorisation of communication types (emailing behaviour over threads of discussions), by making the whole process proactive, preventive and/or support legal evidences [63], [120]. Proactive & preventive measures are a way of immediately supporting organisation compliance team and E-discovery legal analysts to effectively explore, find/identify/discover interesting information and produce it as visual evidence to win the case before/after trail [63]. Some of the examples of E-mail communication data are provided in the Figure 1.1.

Illustrative Example: This is a high-level example of what the current problem is about. Corporate scandals such as the Enron case [110] surfaced in 2001, where many top officials and staff were involved in the fraud, which were evident in the E-mails communicated. Once the scandal surfaced, an US legal team had only the temporal information (October 2001) but they did not have any information about the individuals, connections and contents. The legal team had to manually comb through a four-year period of E-mails communicated to find/identify/discover time, individuals, emails/contents exchanged and the relationships between the three features. As stated, the investigation process was complex, tedious, costly and time-consuming. After several rounds of iterative manual investigation, E-discovery analysts identified the top officials (Kenneth Lay, Richard Shapiro, Jef Skilling, Andrew Fastow, etc.) involved in the scam. Scandals of this sort have increased calls for stronger compliance (which means compliancy with laws and regulations) - these laws can have criminal or civil penalties or can be regulations. There is a pressing need for investigation process to be simple, easy, inexpensive and time-saving [63].

The real case study and the real-world problem in an organisation within E-discovery compliance motivated us to establish problem statement, research question, aim and ob-

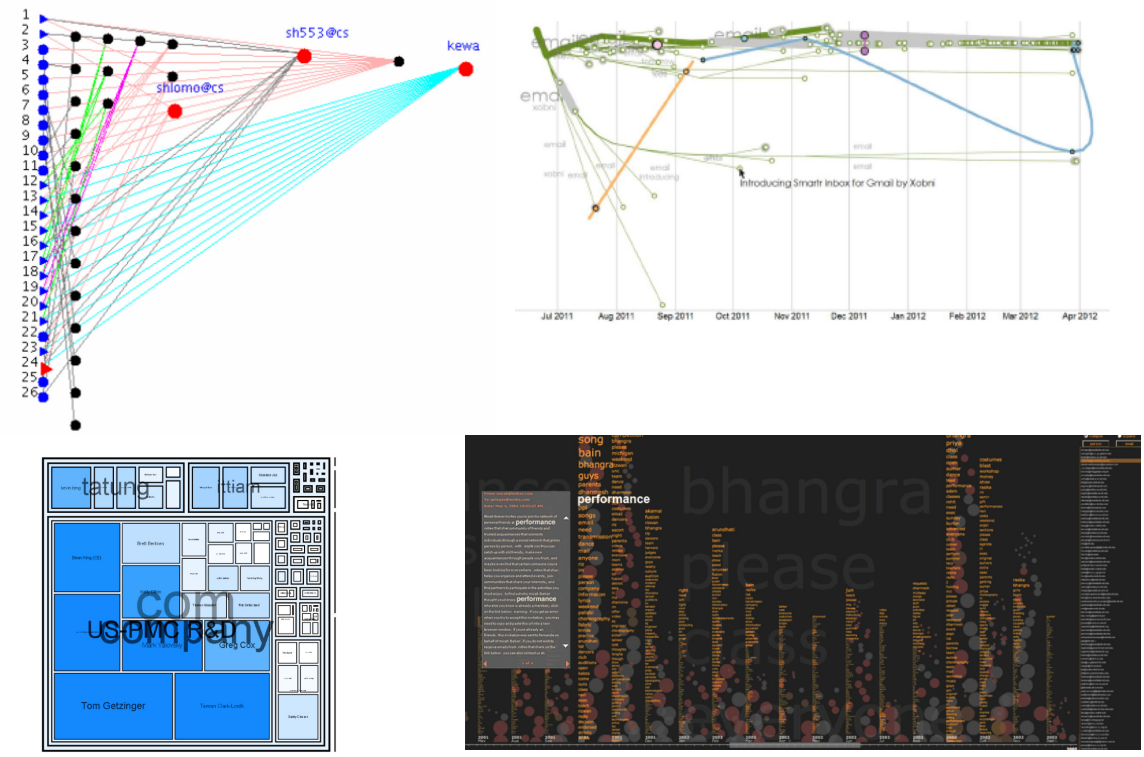


Figure 1.1: Examples of E-mail Communication Data. L-R (a) Email Clique [121] is used for visualising relationship between users and their clique membership. (b)EmailMap [124] is used for visualising frequency of message exchanged during a selected period of time. (c) Treemap [138] is used for visualising an individual for a particular month to understand his contacts in in-groups and out-groups. (d) TheMail Vis [176] is used for visualising an user's email exchange with an another individual during a selected period of time.

jectives.

1.2 Research Context

In this section, the need of E-discovery compliance in an organisation and the need of visual analysis and visual evidences to support E-discovery compliance in an organisation are discussed.

1.2.1 E-discovery compliance in an organisation

E-discovery Compliance in an organisation is one of the chief drivers for organisations to take on any legal actions against individuals working in their organisation or legal actions against other organisations [67]. Organisation compliance team and E-discovery team work in tandem to tackle legal issues (as shown in Figure 1.2). Due to data protection laws and privacy legislation (such as GDPR [2]), organisations are mandated to keep records of all electronics/digital communication data (compliance with government regulations). One of the traditional communication modes in organisations is E-mail system to exchange messages or documents. E-mail data are increasingly called upon as evidences in legal cases, either to protect organisations or even incriminate them. If organisations are unable to produce E-mail data or evidence when called upon by the courts or the authorities, they can face penalties [67]. Hence the need for analytic empowered solutions that cover organisations and leaves them better prepared when records/evidence are requested.

Point: From the discussions with the experts, we understand the branches and the focus of our work (mentioned in the Appendix A.7). In this research, our high-level focus is on "E-discovery Compliance" and a low-level focus is on "E-mail compliance" (E-mail communication discovery), specifically targeted at "Information Discovery" as shown in the Figure 1.3.

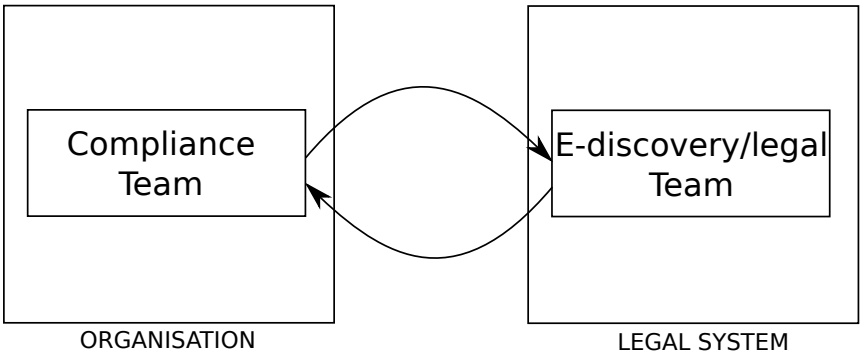


Figure 1.2: Organisation Compliance team and E-discovery team work in tandem to tackle issues.

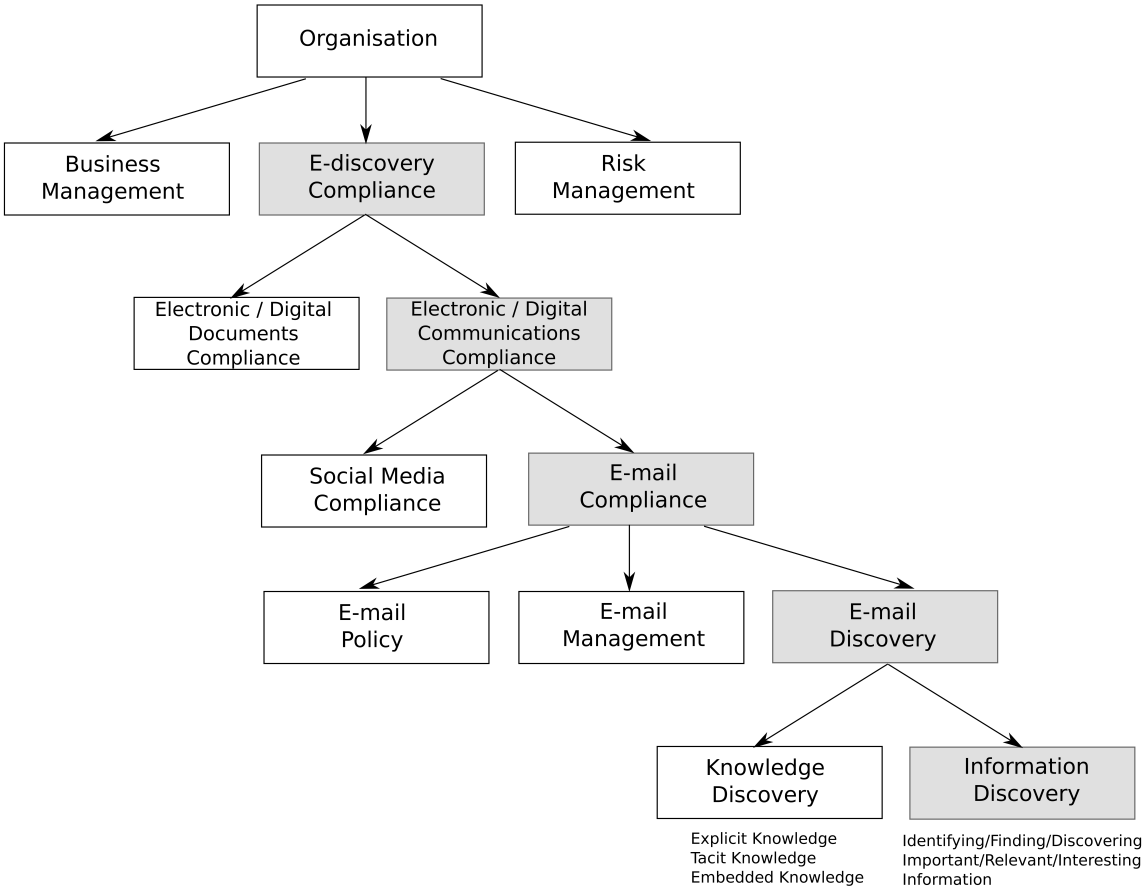


Figure 1.3: The shaded blocks represent the areas we will focus in this work.

1.2.2 Visual analysis to support E-discovery compliance in an organisation

In the non-visual data mining techniques, there is a high probability of irrelevant data being fleshed out which may not be useful from the investigation perspective. Since the communication data is multi-faceted and there is a lack of established classes/labels, it does not always come in a shape that can readily be fed into a classical data mining algorithm, and would need "creative" intervention of an analyst to get into shape. We specifically use visual techniques (visualisations with a combination of computation/analytics), so the communication data can be explored by the analysts involved in the investigation to observe the communication patterns and identify relevant information. The report by the UK Home Office [63] highlights the importance of using visualisations in E-discovery investigations, that can aid in investigating individuals and facilitating the generation of categorisation of communication types (emailing behaviour over threads of discussions). Despite the use of visualisations in various domains, there are no optimal solutions to support organisations in E-discovery compliance [63]. It is argued that visually identifying/finding/discovering various information or groups of data objects from multiple perspectives in E-mail communications have been under- explored and under-investigated [63]. Existing literature indicates a lack of substantial work in the area of email investigation. So, there is a need to analyse the data in efficient ways and visualisation could be a good solution [63] to support the organisation compliance and analysts find important/relevant information such that the whole process saves time. As mentioned earlier, investigating emails is a time-consuming process. The authors of "What Is Visualisation Really For?" [56] have presented a reasoned argument that visualisation helps in accomplishing tasks by saving time, which helps in gaining insight and making routine observations. To secure one's organisation in several ways and to defend any legal cases successfully, we understand visualisations can save time and act as an evidence in supporting the case. This is discussed below.

1.2.3 Visual solutions to strengthen E-discovery compliance cases

There is a well-known idiom "Seeing is believing" which means "only physical or concrete evidence is convincing". Especially, in a court of law, the result of a trial is typically in light of substantive proof combined with visual evidence and a convincing contention to keep jury interested in the case and to explain complex issues in a simple and easy manner which will be understood and retained by the jurors. An original report published in 1963 [180] uncovered that following 72 hours, individuals have a tendency to hold just 10 percent of the information they hear and 20 percent of information they see. At the point when individuals hear and see the same information, they hold 70 percent. This is an extraordinary 700% increase over information that is heard by individuals.

To this argument, many legal firms have started thinking in terms of visual solutions to support and defend their cases [35]. Though visual solution is a vaguely used term in the court of trial, we differentiate ourself clearly from this understanding. We call it as "data visualisation as evidence" – which are visualisations built using data (explained in detail in Chapter 3). In an E-discovery case, an organisation lawyer who wants to win a legal case against the opposition organisation lawyer, data visualisation as evidence can be demonstrated to defend the case, as visuals have the potential to convey more complex meanings and often represent concepts that are challenging to express. A jury that can visualise the case while the expert is testifying will be much more likely to comprehend the occurrence and give a fair judgement. But the question that remains to be answered is "to what extent visualisations can support E-discovery analysts (including jury)". This leads to the formulation of our research question which is discussed in the next section.

1.3 Research Question

Using visual analysis in E-discovery compliance for E-mail communication within an organisation is still an under-researched topic, this research aims to develop interactive visual solutions that can help in the investigation/analysis for analysts to come to a conclusion, build a legal case and support jury, in a way making the whole process proactive and

preventive [63, 120, 3]. For example, if there is a regulatory or compliance request for financial transactions that happened through email communications. A solution is expected to aid in investigation by proactively searching through various communication patterns and preventive measures can be taken immediately to find out what really happened.

The fundamental goal of this research is to investigate critical aspects of E-mail communication within an organisation compliance by designing and developing interactive visual solutions to support E-discovery analysts. More specifically, this research is interested in answering the question *"To what extent visualisations can support analysts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation?"*.

The UK Home Office report on "E-discovery in Digital Forensic Investigations" [63], and the discussions with Red Sift experts have grounded our research. Based on the literature review and discussion with the experts, there is a difference between "finding relevant" and "discovering interesting" information in the email communication data. In the investigation domain, the term "discover" is more pronounced than "find". From the experts' view (mentioned in the Appendix A.7), we consider "discovering interesting" information in the data as identifying/finding/discovering relevant, important, interesting and characterising interesting information to support investigations. We aim to address questions related to discovering interesting temporal information, individuals information, and their behaviour (conversations) in the E-mail communication data.

1.4 Aim and Objectives

Aim of the research

The aim of the research is to design and develop interactive visual solutions to explore and find/discover relevant/interesting information in a corpus of E-mail communications from an investigation perspective to support organisations specialising in Digital Forensics and E-discovery.

Objectives of the research

O1: *Develop design requirements:* understand the E-discovery domain, identify the knowledge gap and develop a rich understanding of challenges, tasks and requirements (specific to E-mail communication data).

O2: *Design and develop visual methods:* design and develop interactive visualisations based on the domain requirements to effectively navigate and explore within data to uncover relevant/interesting information and relationships within the multi-facets such that solutions can be used as an evidence in investigation.

O3: *Validation of visual methods:* validate and re-access the developed interactive visualisation prototypes by conducting validation and empirical studies. Express findings based on the user-centric approach & evaluation that can help analysts to investigate and navigate E-mail data productively and identify various interesting information relevant to the investigation.

1.5 Summary of Contribution

The contributions of this thesis can be summarised as follows:

- **Domain Characterisation & Tasks for E-discovery** - Characterising the domain and tasks in E-discovery investigations is of importance because analyst/investigator finds a vast and semantically meaningful collection of data, that is uncategorised and unlabelled, in E-mail inboxes to discover/find interestingness in the data. Iterative user-centered design approach helped in understanding user requirements (from the experts). The thesis contributes with E- discovery domain, key challenges identified, design requirements, analytical goals and tasks related to E-discovery which helped us characterise the domain and tasks related to E-discovery investigation (discussed in Chapter 4).

- **Knowledge Gap and Overview of the Existing Techniques** - This thesis contributes with a harmonisation of the taxonomy of entities based on the visualisations related to email communication data that makes the entities, the association between them and their limitations explicit. The taxonomy helped in identifying four

main entities (temporal, individuals, contents, and threads) and the limitations of integrating all the four entities. The taxonomy also helped in understanding the state-of-the-art, visual design principles, interaction techniques used, visualisation tasks, methods, techniques etc. to design interactive visualisations for email communication (discussed in Chapter 3).

- **Interactive Visualisation Designs** - Iterative user-centered design approach supported in designing interactive visualisation solutions in an applied context with the experts. We observed that interactive visualisation solutions (multi-faceted exploration and multi-granular analysis) can support analysts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation. Specifically, interactive visualisation assisted active learning helped in classifying communications relevant to investigation. The novel visualisation for a single thread analysis helped in revealing hidden patterns of several individuals who were secretly copied (bcc) when sensitive information was broadcasted. This thesis contributes with insights and reflections on the effectiveness of particular design choices where we learnt conventional visualisations and novel visualisations must be considered based on the design requirements. We argue that using conventional visualisations considering all the features will help in exploring and discovering interesting information that can be relevant to investigations. To find nuances in a communication, novel visualisations can be considered which can help in discovering hidden information relevant to an investigation case (discussed in Chapter 5).

- **Validation** - Iterative user-centered design approach supported in validating our solutions with the experts throughout the study. To evaluate our visualisation and the user experience of our system, we conducted an empirical evaluation to understand how the experts use the system in exploring, discovering and interpreting the features and patterns that can be interesting in a given E-mail dataset. We also made observations on how experts interact with information, how they use interactive clas-

sification methods and discover various information of interest. We draw upon and demonstrate to what extent visualisation can support analysts and how the solutions are effective in an applied context (discussed in Chapter 5).

- **Deploy Solutions** - The interactive visualisations developed are deployed in our partner organisation to analyse their organisation emails and discover interesting patterns related to their business collaborations. The design study helped in critically understanding various design, developing them and understanding its effects through analytical and practical way of inquiring to support email investigation, through the iterative process of deploying solutions in the collaborators platform (discussed in Chapter 5 & 6).

- **Lesson Learnings & Principles** - the principles and lessons learnt are expressed to inform future use of visualisation in E-discovery and Digital Forensics (discussed in Chapter 6).

1.6 Report Layout

In this research, we develop techniques and implement various strategies in software prototypes through a structured process of abstraction, design and testing, by using a well-known methodology called Design Study Methodology (DSM) [156, 61]. Doing so, the remainder of this report is organised as follows:

- **Chapter 2** describes the methodology and methods considered for this research work.

- **Chapter 3** examines the notion of E-mail communication data, E-discovery and Visual Analytics. We examine the current state of the art of the research literature in each of the topics, develop taxonomy of communication data, and enumerate the current available implementations.

- **Chapter 4** describes the requirements and task analysis for this research work, which is based on the interviews with the experts.

- **Chapter 5** describes the framework of the work, designs and sketches developed using paper and Tableau which leads to the development stage, where the data-driven visual solutions are implemented based on the tasks, requirements, designs and sketches. This chapter also discusses on the evaluation of the visualisation and the actual experience of our system. User evaluation was conducted to understand how the participants use the system in exploring and finding/discovering interesting features/patterns/information/points in a given E-mail dataset.

- **Chapter 6** forms the conclusion and reflection. Based on the tasks and requirement analysis, along with the evaluation, we have identified learnings, positive findings, limitations and principles forming key points of discussion.

Chapter 2

Methodology

To address the problems mentioned in the Chapter 1, we consider User-centered Design (UCD) approach, a human factors-based design, which is an iterative process involving task analysis, design, prototype implementation and testing. In UCD, we adopt Design Study Methodology (DSM) [156, 61], which is a three phase (a nine-stage) model, a methodological framework that provides practical guidance for conducting a design study and research. In this methodology, we start by characterising the domain of E-discovery for E-mail Communication. From this analysis, we compile a list of tasks and associated data variables that one would wish to visualise to support these tasks. These tasks can be framed as a set of questions, for instance, 'Which individuals are always part of E-mail communication but not replying to E-mails?'.

As discussed in the introduction, the goal of the project is to develop effective interactive visualisations for analysis of E-mail communication data in an organisation that can help organisation compliance team and E-discovery legal analysts to effectively explore, find/identify/discover relevant, important, interesting and key information relevant to the case. The need to assess the role and use of visual analysis for E-mails, a problem-driven research, is recognised by the industrial partner for this research, Red Sift London, UK. Since the project aims to deliver a proof-of-concept for a robust E-discovery solution which enable companies to make this a proactive and preventive process, the solution will sit on top of Red Sift's computational architecture, and harness powerful techniques that can

carry out key E-discovery tasks more efficiently and create richer visualisations and insights from E-mail data. To support this argument a discussion of the design study and methodology is now presented in the following sections.

According to Sedlmair et al. [156] "A design study is a project in which visualisation researchers analyse a specific real-world problem faced by domain experts, design a visualisation system that supports solving this problem, validate the design, and reflect about lessons learned in order to refine visualisation design guidelines". The design study helps in critically understanding various design, developing them and understanding its effects through analytical and practical way of inquiring. Also the authors [156] explain the difference between methods and methodology. Methods are "techniques or procedures" and a methodology is the "strategy, plan of action, process, or design lying behind the choice and use of particular methods".

There are many models and methods to approach problem-driven research for visual design and evaluation [128]. Some of the mostly cited ones are Multi-dimensional In-depth Long-term Case studies (MILCs) [159], the Nested Model (NM) for Design and Validation [131] and the Nested Blocks and Guidelines Model (NBG) [127, 128]. In these models, there is very little guidance available how to conduct design studies effectively. However, Design Study Methodology (DSM) [156, 61], a nine-stage model, is a methodological framework and provides practical guidance for conducting a design study. This framework has three main phases: pre-condition, condition and post-condition phase. Each phase has three stages, which makes it nine stages in total as shown in Figure. 2.1 & 2.2. The details are discussed in Section 2.2. For each stage in all the phases, the authors have provided practical guidance/advice based on their own experiences and outlined several potential pitfalls. The DSM model is being used in many design studies and some of the papers had collaboration with organisations [46, 80, 140, 155, 153]. We considered three papers between 2016 and 2018 relevant to the investigation that adopted Design Study Methodology (DSM).

Study 1: BubbleNet [126] is a cyber security visualisation dashboard for investigation. The design study adopted DSM [156] with the goal of improving how analysts discover

and present anomalies and patterns within cyber security. The study ran over 2 years where the authors considered user-entered design process to incorporate user feedback, their needs, and workflows throughout the design of solutions. The approach resulted in a successful evaluation of dashboard with the domain experts and further deployed in both research and operational environments. The paper helped us in understanding how design a complete solution and how to bridge the gap between domain and visualisation experts. The authors also helped us in understanding how to use ideas as data sketches and turn them into prototypes.

Study 2: Concept Explorer [97] is a visual comparative case analysis tool for investigation. This work is part of the EU-funded project "Visual Analytics for Sense-making and Criminal Intelligence Analysis (VALCRI) [172]. The authors adopted DSM [156] to iteratively build and refine their approach based on several rounds of expert feedback from the users. The study ran over 2 years considering the user-centered design process, where the authors built effective visualisations such that domain experts can provide feedback to the system and observe the impact of their interactions in the investigation cases. The paper helped us in understanding the complete design process and specifically, visual analytics workflow that can be designed using an iterative approach (interviews with the experts which forms an user-centered design) to build a system that can aid in visualising the clustering patterns with the feature relations and provide feedback to the system.

Study 3: Polimaps [166] is a visual analytic predictive policing tool for investigation. The study followed DSM [156] which were driven by on-site visits, discussions, sketching ideas and feedback rounds with the domain experts. The study ran over 1 year following the user-centered design process which helped the authors integrate machine learning with interactive visualisation for predicting risks at particular period of time in various locations/areas in a map. The paper helped us in understanding how to conduct a design study and also helped in understanding some of the design decisions and features implemented were directly motivated by the iterative user suggestions/feedback.

Based on Sedlmair et al. [156], we understood if there is a specific real-world problem faced by the domain experts using real data, we can design visual solutions in solving

the problem, validate the designs iteratively, and reflect about lessons learnt in order to improvise the visual solutions. These three papers BubbleNet [126], Concept Explorer [97] and Polimaps [166] served as a motivation for us to adopt Design Study Methodology (DSM) for building interactive visual solutions to support Digital Forensics and E-discovery investigations.

2.1 Stage-by-stage Research Approach

Based on the Design Study Methodology (DSM) [156, 61], a three phase (a nine-stage) model, each of the stages below correspond to a specific output and should be read in conjunction with the Diagrammatic Work Plan (provided in the Appendix). This is a methodological framework that provides practical guidance for conducting a design study, where the model involves both theoretical and a practical development.

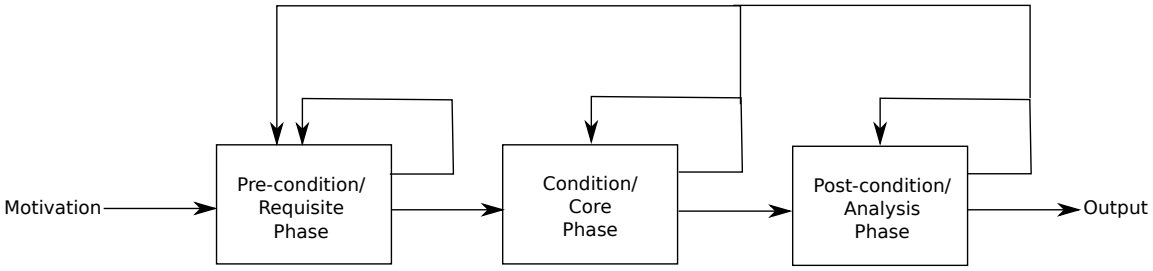


Figure 2.1: A high-level view of the three main phases of the Design Study Methodology (DSM) [156] for Visual Analytic Design and Validation.

The general layout of the framework is linear to suggest that one stage follows another. However, this linearity does not mean that previous stages must be fully completed before advancing to the next. In most of the cases, many stages often overlap and the process is nested and highly iterative, which means the work can get into loops and one can go back to any stage from any other stage. Validation is essential at each stage in the design process and they are classified into three phases: pre-condition, condition and post-condition phase. The details are discussed below.

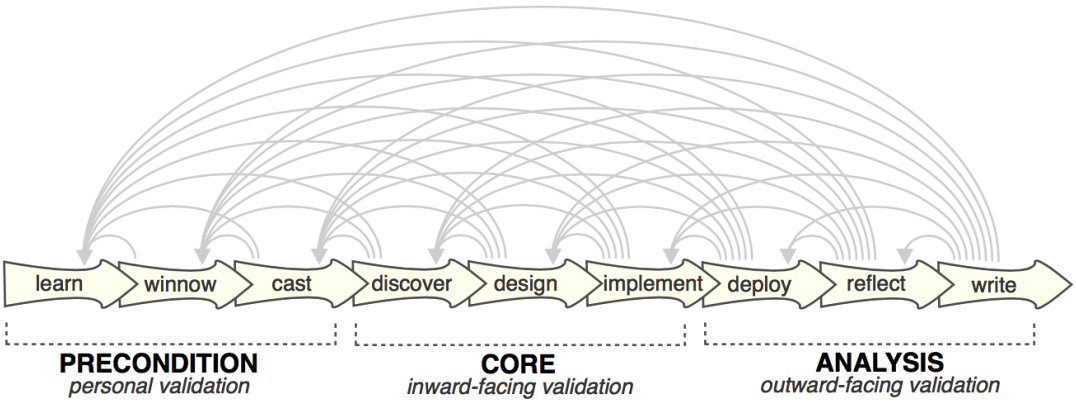


Figure 2.2: A low-level view of all the nine stages of the Design Study Methodology (DSM) [156] for Visual Analytic Design and Validation.

2.1.1 Phase 1: Pre-condition/Requisite Phase

In our design study, the goal is to work with users (E-discovery compliance) to solve their real-world problems (E-mail investigation) by providing visual solutions using real-world data. We consider E-discovery [30] as an investigation domain for searching, finding and discovering information and use it as an evidence in a legal case. Our target domain users are lawyers and analysts in an organisation.

In this phase, state-of-the-art for the research was conducted and reviewed. Followed by identifying a number of real-world problems, case studies, scenarios and/or stories that can serve as a motivation for the research. Also, identified potential collaborators, their engagement in the project, real-data and tasks.

Learning, Winnowing & Casting Stage

Methods: For an effective design study, a solid knowledge on the literature is critical (stage 1) [156]. For example, knowledge of the current state-of-the-art is crucial for comparing and contrasting findings using inclusion and exclusion criteria (as discussed in Chapter 3). The knowledge will provide a good base for all later stages. The goal of the winnowing stage

(stage 2) is to identify potential collaborators who can play a role through the complete lifecycle of the project. In our project, the winnowing was not carried out as it is an industrial-funded project and the experts agreed at the start to be committed to collaborate through the lifecycle of the project. In the casting stage (stage 3), two critical roles in a design study has to be considered. The front-line analyst is the domain experts performing the investigation and the gatekeeper is the person with the power to approve/reject the designs/solutions. In this phase, we consider **personal validation** [156] which focuses on research planning and implementation. We followed these papers [46, 80, 140, 155, 153] for guidance.

Milestones: We worked with real users (E-discovery compliance) to solve their real-world problems (E-mail investigation) by providing visual solutions. Based on the winnowing and casting stage, the front-line analysts are Red Sift CEO and employees; the gatekeeper is the CEO of Red Sift (the complete details are discussed in Chapter 4). Based on the first three stages of the DSM [156, 61], we considered E-discovery [30] as an investigation domain for searching, finding and discovering information and use it as an evidence in a legal case. We conducted state-of-the-art review for E-mail communication data within E-discovery. Investigated how the existing E-discovery (investigation) tools and other visualisations work (finding gaps in the literature). We identified a number of real-world problems/challenges, case studies, use case scenarios and/or stories that can serve as a motivation for the research work. Our target domain users are Red Sift analysts. We used a real case study for investigating the E-mails communicated by the employers and employees and used this publicly available dataset for the design process. The complete stage is discussed in Chapter 4.

2.1.2 Phase 2: Condition/Core Phase

In this phase, research questions, aim and objectives based on the existing problems were framed. The other deliverables are tasks and requirements needed for the project were formed. Several meetings/interviews held with the domain experts in order to understand the current workflow were considered. Unstructured interviews were conducted to gather

tasks, followed by initial requirements and they are validated using mock-ups and sketches. The key output from this phase is a transformed real data set which helps in designing visual solutions. The prototype visualisations with interaction are delivered. In this phase, we consider **inward-facing validation** [156] which focuses on validating/evaluating findings and artefacts with domain experts.

Discovering Stage

The goal of the discovering stage (stage 4) is to characterise problem and abstract task [156]. This stage is also called as requirements analysis in software engineering.

Methods: the general practice in user-centered design is a combination of methods including interviews and observations [40] that can be structured, semi-structured or unstructured. However, *just talking* and *contextual inquiries* [103][132] along with deep literature study helps to provide interesting and relevant information where the researcher observes users working in their real-world context and interrupts to ask questions when clarification is needed, also clarifies many points by referring/conducting literature review. The notes captured are transcribed and coding/thematic analysis is carried out (the methodology is discussed below, in the Evaluation process).

Milestones:

We conducted several meetings/interviews (unstructured) with the Red Sift analysts to understand requirements and how they use visualisation. From the literature and the interviews, the E-discovery analysts find the current way of investigating emails to be complex and difficult to understand, explore, identify outliers (anomalies) and find interesting information. In many investigation cases, most of the domain experts often know *what* they are looking for in their dataset but not sure *how* to get there which can be time-consuming. Since, most of them are not technically sound, they are in need of a simple and effective tool to visualise and identify interesting information in E-mails. The complete stage is discussed in Chapter 4.

Designing and Implementing Stage

The goal of the designing stage (stage 5) is to facilitate data abstraction, visual encoding and interaction mechanisms [156]. The abstraction step is to represent the analysis goals and tasks at a high and generalisable level, map problems and data from the specific domain point of view. And then we consider these analysis goals and tasks to inform designs we carry out in the subsequent steps. After reaching a shared understanding of a problem with domain experts in the discovery phase, we ideate and design a visualisation solution (making visual encoding decisions). In our work, implementing stage (stage 6) is nothing but solution development process. The goal of this stage is to develop/implement designs, prototypes, tool and usability [156].

Design Process / Design Prototypes

Design Prototypes are used for representing ideas and also represent what could be a potential final product. It can be in the form of paper/digital sketches or some level of implementation which helps in exploring the design and its context. Buchenau & Suri mention "prototyping is a key activity within the design of interactive systems" [49]. The design prototyping to be developed through a human-centered and iterative design process. Prototypes allow us to examine the designs and decide whether to move forward with the idea/designs or try alternate approaches [130]. This can be iterative to get the possible solution. The final prototype probably include realistic views and interactions that can be tested for assessment and bringing out insights. The prototypes may have distinct features. A prototype's fidelity relates to how tightly the mockup represents the real completed product. The greater the fidelity, the more it looks to the final product. The more distinguishable it from a final product, the lower the fidelity was mentioned by Nielsen in The Usability Engineering Life Cycle [134].

Based on Tamara Munzner's recommendation [131], we consider **rapid prototyping** (Prototype model) before developing a fully functional software tool. An Incremental

Development approach combined with an Iterative Prototyping approach [117] was found to be the best solution, they work well within the nested model. Rapid iterative Prototyping is about building initial versions of designs/solutions with limited functionality (not a complete version) and collecting feedback from the end users to get a fully functional product. The main goal of this method is to quickly develop throw-away code, which is crucial in design studies. In particular, the more time spent coding a solution the harder it is to throw it away. The tendency is to tweak a given implementation rather than to start over from scratch, which is problematic in cases where a design turns out not to fit the identified needs and problem of the experts, or where the needs have changed. Several tactics for design studies are: start simply, ideally with paper prototypes; quickly write code that can be thrown away; and close user feedback loops with methods such as design interviews and workshops [57], or deploying early versions of a tool as technology probes [95].

Rapid iterative prototyping enables to understand customer/client/user requirements (user specific) at an early stage of development. It helps get valuable feedback in a short time from the users and helps researchers, software designers and developers understand about what exactly is expected from the product under development. In this approach, errors can be detected much earlier, missing functionality and/or confusing/difficult functions can be identified easily. Also, this approach reduces costs, time and risks associated with the projects. This is an iterative process to build a complete version. For example, paper prototypes and wizard-of-oz testing [70] can be used to get feedback from target users about abstraction and encoding designs without addressing the algorithm level at all. Considering these points, we considered three levels of fidelity: low, medium and high. They are explained in detail below.

Methods for Designing [1]:

1. *Low-fidelity Prototype:* We started the design process with low fidelity prototypes, that is paper-based sketching (also called "Paper Wireframing"). Low-fidelity pro-

[1]The technologies selected are discussed in the Appendix A.2

totyping can help in coming up with a lot of ideas [144]. This is used in expressing concepts, design alternatives and screen layouts [148]. The low-fidelity prototypes are used in the early part of the design and development cycle to develop ideas, conceptual approaches and helps in gathering requirements. The prototypes can be constructed with paper, pencils/sketches and/or simple storyboard tools to receive feedback/views from users/experts on the designs meeting their requirements. The feedback can be used to further iterate the low-fidelity prototype or as input criteria for the subsequent higher-fidelity prototype.

2. *Medium-fidelity Prototype:* Medium-fidelity prototyping is also used to get further feedback/views from the users/experts. These types of prototypes are usually software-based (on a computer screen instead of printed paper) to simulate some but not all features of the interface where the users can use a mouse and keyboard to interact. It is also called "Digital Mockups". Constructing medium-fidelity prototypes are inexpensive and fast to construct and change, providing advanced but restricted scenario for end users to attempt and test subtle design issues [42]. This sort of prototyping enables the design and helps in understanding the navigation, layout and functionalities. However, medium-fidelity prototypes do not fully communicate the look and feel of the final product but only gives a fair idea of what a final product could be [73]. We used R and Tableau for designing the medium fidelity prototypes.

3. *High-fidelity Prototype:* High-fidelity prototyping is a software-based too which is a very near version of the final product with the expected layout and functionalities. The feedback from the users can be used to further iterate the high-fidelity prototype or as a final version of the prototype, which can be called as "proof-of-concept (PoC)". The visualisation and interaction models were generated using Data-driven Documents JavaScript library (D3.js) by Mike Bostock [41]. After an initial evaluation of different charting libraries, we considered D3 to be the most suitable technology and the current best in class platform for building interactive data visualisation which uses JavaScript. The rationale for using D3 is mentioned in the Appendix A.2.

For fast interaction (incremental filtering, reducing and comparing), crossfilter techniques were implemented, as they are quite helpful for exploring large multivariate and/or multi-faceted datasets in the browser. The solutions implemented in the D3 are based in-line with the design tasks captured from the interviews.

Inward-facing Validation

Our work underwent several paper/prototype iterations before testing it on the real dataset, as visualisation solutions are best validated with real datasets. Various design solutions were discussed with the Red Sift analysts and E-discovery experts in the form of paper sketches, digital sketches (using Inkscape), medium-level (using R & Tableau) and high-level prototypes (using D3).

The visualisation literature contains a multitude of proposed methods for validating and evaluating visualisation tools in the wild [76, 196, 96, 115, 53]. The most common form of validation are use case scenarios with real users, real problems, and real data, as featured in many strong design studies by others [135, 155, 153, 154]. The steps we followed are listed below.

Step 1: Users, Tasks. We identified E-discovery experts which helped us collect challenges, analysis goals, tasks and data. Using taxonomies of E-discovery tasks, we expose uncertainty, determine interestingness and confirm hypotheses. This help in re-accessing the design thinking and its process in terms of better chart selections and develop better visualisation solutions (proof- of-concept) using an iterative approach by conducting several unstructured interviews and validations.

Step 2: Use-case Scenarios & Validation. We designed and documented use-case scenarios that demonstrates the effectiveness of the design solutions that are built and tested through personal validation. For improving the solutions, we merged "Personal Validation" [156] and "Inward-facing Validation" [156] based on the inputs from our collaborators (mentioned in the Appendix A.7). In software engineering, validation is about checking whether one has built the right solution based on the user requirements and ver-

ification is about checking whether the tool has met the requirements [131]. In our design validation, we carry out both validation and verification through use-case walkthroughs. Based on the discussions with the experts (mentioned in the Appendix A.7), one of the advantages of using this approach - it saves time in conducting empirical/user studies. So, personally validate a tool, report on findings, then quickly walkthrough a use case with experts and then deploy a solution immediately to check for engineering issues. This approach allowed us to continue with the next phase of the design process. In our design phase, the personal validation was conducted by walking through an analysis scenario with a real dataset, with a scenario and tasks, later with an expert to demonstrate how our solutions can support an analyst [164]. The tasks helped to determine the potential effectiveness of our techniques in email investigations. Our validations were discussed with the experts and they explored the tool to understand the same.

Step 3: Coding & Thematic Analysis. We conducted several meetings/interviews (unstructured) with the analysts to understand designs (improve) and we validated it regularly. We were making regular notes in our diary book and later it was transcribed straight after the interview to consider any clarification. This process was carried out on Google Documents (by sharing it with my supervisors). We focussed on the coding and thematic analysis to identify themes, which is described below.

Empirical Evaluation

As a final step, empirical study was conducted with the real users (experts - i.e. company partners, Red Sift experts) by providing the same real-world scenario and real tasks to demonstrate how our solutions can support an analyst [164]. The major goal in validating and evaluating a deployed system was to find out whether domain experts are indeed helped by our visualisations. This goal was confirmed by experts doing tasks faster, more correctly, or with less workload, or by experts doing things they were not able to do before. The sub-steps are discussed below [44] [45].

Step 1: Gathering Qualitative Data. we considered "*observational study*" as a form of empirical study [96], using real data and real problems, to collect the qualitative data.

Since, there are no specific guidelines on the minimum criteria or saturation point for the number of participants in the study [96], we considered three experts from the organisation. We gave a demo to the experts and helped them understand using the system (training) before we started observing them. This study observed the efficacy of finding/discovering/identifying interesting information in a selected set of E-mails and also to evaluate the visualisation design choices (characteristics and aesthetics) for some of the low-level tasks, such as aggregation, comparison, etc using Visual Data Reasoning (VDAR) [117]. The tasks were based on the design requirements and research questions. Each of the generalised tasks had low-level (specific) tasks. The interactive visual solutions were provided to the participants. Based on the tasks, the participants navigated and interacted with the observers (us), we were making "*constant notes*" [45] based on the design, encodings, interaction mechanisms, and the visual solution itself. The transcriptions were carried out straight after the interview to consider any clarification. This process was carried out on Google Documents (by sharing it with my supervisors).

Step 2: Analysing and Decoding of Data. To start with, we focussed on "*reading*" [45] and "*familiarisation*" [45] to take note of points of potential interest and that are relevant to the research question. We considered "*coding*" approach [45], which means we were keen on focussing on the complete data that were relevant to the literature and research question. This helps in understanding the views of the experts. The coding is the process of subdividing and labelling raw data captured in the notes, then reintegrating collected codes to form a theory. Later, we started searching for "*themes*" [45] that were sensed through review of the observations, coding and data. The goal of this process was to extract themes and to present a coherent, consistent picture of the tasks and situations under study [96][45].

Step 3: Validity. The data we analysed in qualitative research came in the form of artifacts (observations/notes). As explained above, qualitative data can be gathered and analysed systematically. In qualitative research, however, it is acknowledged that the participant's views, researchers views, research context, and interpretations are an essential part of the qualitative research method, often leading to a hypothesis about the situation

under study [96]. The type of study was conducted to get some more insights and for improving our solutions [45].

In addition to this, ethical issues were considered while carrying out the interviews and empirical study. We obeyed and followed the University's ethical approval process considering the latest GDPR rules and regulations. We considered all the points listed by Purchase [141] for the interviews/empirical study to be approved and the collected data are stored anonymously and securely. The ethical forms approved by the University are attached in the Appendix A.5.

Milestones: Our work underwent several paper/prototype iterations before testing it on the real dataset, as visualisation solutions are best validated with real datasets. Various design solutions were discussed with the Red Sift analysts in the form of paper sketches, digital sketches (low-level prototypes using Inkscape) and medium-level prototypes (using R & Tableau). Later, we created high-level prototypes, isolated D3 visualisations, with real datasets, which had some minimal interactions. The iterative feedback from the domain experts helped in improving the designs such that the tasks such as exploration, finding changes, similarities, anomalies, different communication patterns, and discovering interesting information were introduced and made accessible. We regularly reflected on findings and made changes to the prototype to develop an effective and robust visualisation tool to explore, understand patterns and find interesting underlying communication structures in E-mail communication which helps to improve E-discovery investigations. The complete stage is discussed in Chapters 5 & 6.

2.1.3 Phase 3: Post-condition/Analysis Phase

Deploying, Reflecting & Writing Stage

In the deploying stage (stage 7), we conduct outward-facing validation with users. The design process can be improved such that concepts, findings, principles, guidelines and/or

recommendations can be derived using better designs and strategies using real and specific examples. Using taxonomies of generic and specific E-discovery tasks, we can expose uncertainty, determine domain parameters and confirm hypotheses. In the reflecting stage (stage 8), we re-access the design thinking and its process in terms of design choices and develop visualisation solutions (prototypes) using an iterative approach by conducting several structured interviews and an empirical study. In the writing stage (stage 9) and in the complete Phase 3, we consider **outward-facing validation** [156] which focuses on justifying the results of a design study to the outside world, including the domain experts, stakeholders, readers and reviewers of reports/articles.

Milestones: In this phase, we took a step back to think in-depth about the design study, reflected on the the objectives, and further generalised those practical experiences, by providing theories, learnings, findings, implications and principles to the future design and study of email communication analysis. The complete phase is discussed in Chapters 6.

2.2 Summary

In this chapter, we discuss the Design Study Methodology (DSM) [156, 61] used in this project which provides a methodological framework and practical guidance for conducting a design study and research. We explain all the three phases (nine stages) of the DSM used in our work. Each stage in this nested form helps in analysing the problem and validate the solution independently. Figure. 6.1 represents how we make use of the three phases in this study and a detailed timeline is attached in the Appendix.

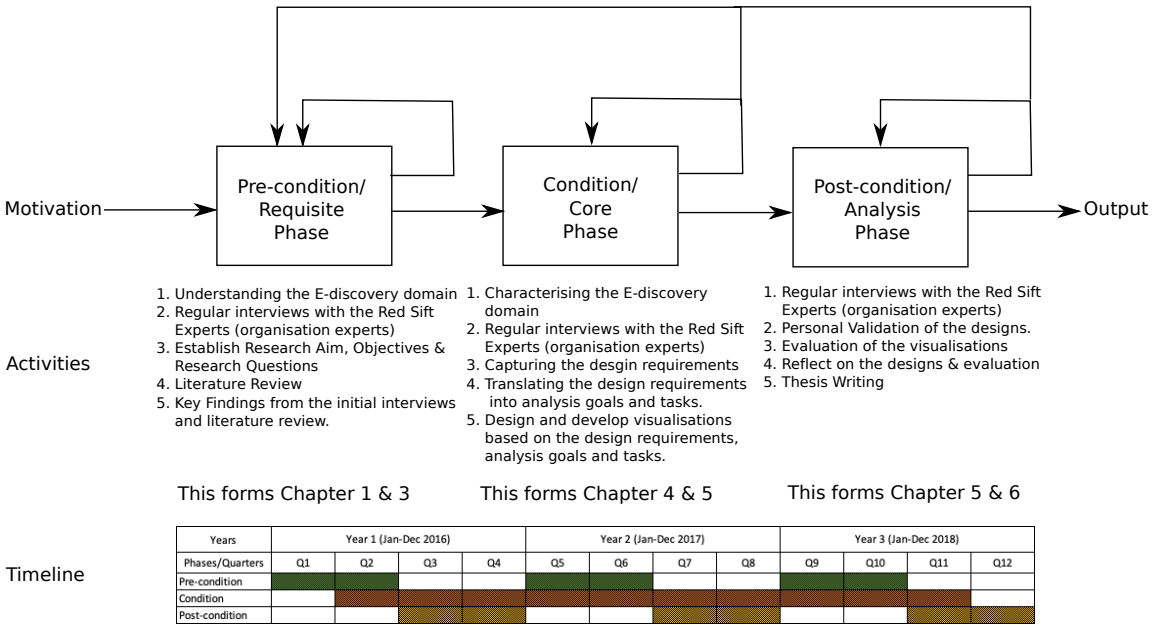


Figure 2.3: The three main phases of our design study in this project (adapted from Design Study Methodology (DSM) [156] are Pre-condition, Condition and Post-condition phase which includes collecting identifying users, capturing user requirements, design, development, evaluation and reflection. All the activities are mentioned in each of the phases along with a timeline.

Chapter 3

Related Work

In the Design Study Methodology (DSM) [156], the pre-condition phase is the first phase of the design study. This phase is also called as "Background Study", "Related Work", or "Literature" phase [156]. So, this chapter presents the background information (literature review) required to understand the concepts, ideas and reasons for using different approaches to address investigation challenges and identify the research gap. The chapter starts with the history of E-mail communication, the role of E-discovery & digital forensic investigation in E-mail communication and the importance of using visual analysis in digital data investigation. We also consider visual analysis that were developed by others specific to E-mail communication analysis.

3.1 Related Work Methodology

Data Collection and Analysis To conduct this survey, we first started reviewing papers without any particular models in mind. We used the following procedure to select papers for our review. As a starting point, we used the digital libraries of IEEE Xplore, ACM and Google Scholar. We also followed citations in both directions: we checked the list of references in the papers to find older works and investigated citations of the paper using Google Scholar. In this way, we could broaden our database and retrieve a comprehensive list of articles relevant to our scope. We also manually searched and scanned through many

journals and conferences.

Journals

- IEEE Transactions on Visualization and Computer Graphics (TVCG)
- Information Visualisation

Conferences

- IEEE Pacific Visualization Symposium (PacificVis)
- IEEE Symposium on Information Visualization (InfoVis)
- IEEE Conference on Visual Analytics Science and Technology (VAST)
- International Conference on Information Visualisation (IV)
- Joint EurographicsIEEE VGTC Symposium on Visualization (EuroVis)

Search Terms (Visual Analysis OR Visual Analytics OR Visualisation OR Visualization) AND
(Digital Forensics OR Electronic Forensics OR Computer Forensics OR E-forensics) AND
(Digital Discovery OR Electronic Discovery OR Computer Discovery OR E-discovery) AND
(Digital Investigation OR Electronic Investigation OR Computer Investigation)

Later, we included Feature Engineering AND Active Learning along with the above-mentioned searches.

This querying process resulted in more than 200 research papers. After the collection step, we refined our paper pool by filtering papers out based on the inclusion and exclusion criteria:

Inclusion criteria:

1. papers published from 2000 onwards;
2. papers that describe visualisation methods/approaches/techniques, and

3. peer-reviewed, high impact, full papers published in journals and conferences written in English.

Exclusion criteria:

1. papers that describe visualisations based on spam/viruses.

2. papers that describe visualisations with "toy" or "simulated" data.

The inclusion and exclusion criteria were considered based on a couple of considerations: there were very few design study papers related to E-mail communication data and the papers do not have distinguished solutions for Electronic mail communication data from the Electronic documents. Most of the email visualisation papers started publishing only from 2001after the Enron [110] emails were leaked. Through the identified papers, we deep tracked their cited papers to find works related to visualisation methods, approaches and/or techniques that were published in 1980s and 1990s. To deeply understand the methods, approaches and/or techniques, we considered the below three aspects, and this helped us to filter down the papers to focus on particular methods.

1. Visual Analysis of Digital / E-mail Communication Data

2. Visual Analysis in Feature Engineering

3. Visual Analysis in Active Learning

In the next section, we introduce "E-mail Communication & Investigation" followed by "Visual Analysis & Investigation" which will help readers understand the visualisation methods, tasks, approaches and/or techniques that are used for "information discovery" specific to E-mail communication.

3.2 E-mail Communication & Investigation

Electronic mail (E-mail)[1], was first invented by Raymond Tomlinson, an American scientist who sent the first E-mail in 1971 across a network to other user. The E-mail system

[1]E-mail can also be written as E-Mail, Email, EMail, eMail, or e-mail.

used the ARPANET (Advanced Research Projects Agency Network), the first system able to send mail between users on different hosts connected to the network. Later, Dr. Shiva Ayyadurai, an Indian-born American scientist, developed an electronic version of an interoffice mail system in 1979 and copyrighted in 1982. Now, E-mail is one of the popular means of communication at the organisational level (business world) and also to some extent at the personal level.

E-mail communication can be defined as an electronic-based communication system where messages can be exchanged between one or more individuals in an asynchronous form using internet. E-mail is a web-based technology (Web 2.0), considered as an ubiquitous and pervasive tool, has grown to become a dominant communication medium for both organisations and private individuals in recent years. The number of E-mails sent and received per day total is over 293 billion (as of 2019) [1]. This figure is expected to grow at an average annual rate of 4% over the next four years, reaching over 347 billion by the end of 2023 [1]. Worldwide E-mail use continues to grow at a healthy pace. In 2019, the number of worldwide E-mail users is nearly 3.9 billion. By the end of 2023, the number of worldwide E-mail users will increase to over 4.3 billion. More than half of the world population will be using E-mail by year-end 2023. The average user gets around 50 E-mail messages a day while high volume users (eg., person in an organisation) can get messages in hundreds [1]. This gives an indication how E-mail communication is ruling one's life in both personal and professional front (in organisation). The individuals use E-mail as a medium to send documents, share confidential information, copy/secretly copy (cc/bcc) other individuals while sharing some information.

In the last decade, E-mail has changed the way we communicate and work in organisations. However, one of the negative consequences is the email-related stress, or 'E-mail overload' [182][64], termed by Whittaker and Sidner, i.e. the lack of control while dealing with a growing number of emails. For a normal individual to manage their own account is a very big challenge [182][64], then let's imagine how difficult it must be for an investigation analyst to analyse several emails of several individual users (email inboxes). The task of identifying various connections between multiple inboxes is a very big challenge [63], we

can term it as "E-mail workload".

Many organisations and/or compliance teams (investigators) are constantly looking for effective tools that can effectively support in finding, discovering and identifying various information and relationships in E-mail communication. Investigating into E-mail communication has peaked up in the last decade especially in an organisation setting, due to E-mail data's multi-faceted nature, which means data contains multi features such as temporal, individuals, connections and content. As the E-mail systems have grown in complexity, organisations and E-discovery analysts are looking for effective interactive visualisations for analysing E-mail communication data in an organisation that can help organisation compliance team and E-discovery legal analysts to effectively explore, characterise key information and produce it as visual evidence.

Electronic Discovery (E-discovery)[2] is a process where electronic data is searched to find information and use it as an evidence in a legal case [54]. The revolution of the electronic and digital devices has changed the way information is stored and the electronic data presents a problem for discovery. E-discovery is also called as E-disclosure and one of the discovery areas in this domain is E-mail communication [59][110]. Organisations increasingly leverage the E-discovery process and technologies for various requests (i.e: criminal investigation, FCA Compliance, FOIA requests, fraud complaints and HR complaints etc).

In E-discovery, Electronically stored information (ESI) is a process where information is created, manipulated, communicated, stored, and best utilised in digital form [162]. ESI includes writings, drawings, graphs, charts, images, presentations, voice mails, audio files, video files, web links, social media, documents and other data compilations stored in an electronic medium. As show in the Figure. 3.1, Electronic/Digital Documents and Electronic/Digital Communication are a subset of ESI. The latter can be further classified into organisational communication and social media communication.

Currently, there are no effective E-discovery tools for E-mail investigation that have the ability to explore, display and identify interesting information in an effective way, that is to discover interestingness between multi-facets of the data (different granularity of

[2]E-discovery can also be written as E-Discovery, Ediscovery, EDiscovery, eDiscovery, or e-discovery.

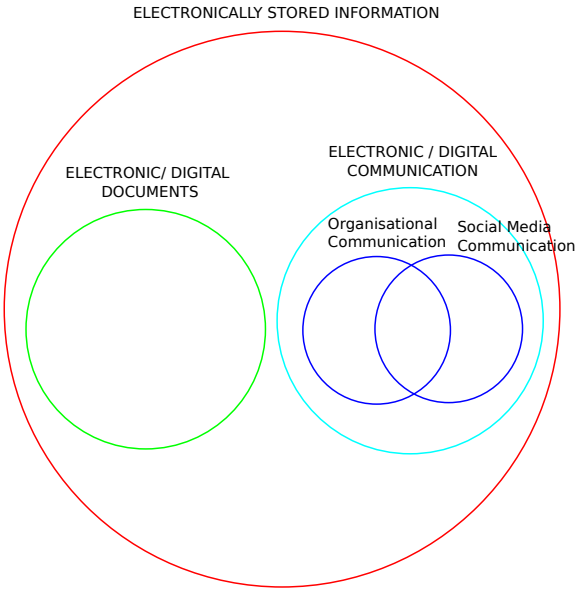


Figure 3.1: Classification of Electronically Stored Information (ESI) in E-discovery

time, individuals, connections and content) [30][120]. The tools currently available on the market are based on simple keyword search and legal firms charge companies based on the volume of information produced by the search, which is then manually reviewed intensely to find interesting information [30]. Investigators search through E-mails, seek answers to various questions in the reports-who? what? when?-to produce it as an evidence to the judiciary. Some of the tasks highlighted by the authors of Beagle [111] are **who**, are the individuals involved; **what**, the content of the messages were exchanged; **when**, the time and date of communication took place and for **how** long. E-mail data is a central resource in E-discovery processes [63] and the existing tools are not capable of handling this vast, dynamic, noisy, real-time, heterogeneous, unstructured and relational data. Addressing E-discovery requests that involve E-mail data, which nowadays can easily go up to millions, is becoming a task that is becoming unmanageably time consuming [30][59]. We need interactive visualisation empowered solutions that will help analysts in their E-discovery tasks that can lead to faster and effective processes to find interesting information [100].

Examples of the E-discovery domain related tasks were extracted from various pa-

pers [63][30][59][120] are as follows:

1. Who is talking to whom regularly and what is the context?
2. Can important/interesting individuals be zoomed in for evidentiary purposes?
3. Where are electronic messages coming from inside a corporation?
4. How many messages were recorded and between whom? How spikes in conversations about specific topics can be of evidentiary interest?
5. What was the first time that there was a spike in communications between individuals of interest?

The real tasks related to E-discovery are abstracted based on the interviews with the experts (included in the Appendix A.7). In a data context such as E-mail, identifying "interestingness" is ambiguous and the information obtained is multi-faceted which makes investigation process tedious and complex [30], which means there is a problem where visualisation could improve the process by incorporating the analysts expertise into the analysis. To help improve efficiency and reduce costs involved in an E-discovery process [63], visualisation techniques can be of great help, and they can change the way we present and understand time, individuals, threads, contacts, and contents exchanged.

Based on the literature survey - Remail [147], Mailview [77], EmailTime [99], Email Timestore [188], MatrixExplorer [90], Honeycomb [175], NetLens [101], Vizster [86], ZAME [71], MatLink [91], NodeTrix [92], SmallBlue [122], Perer and Shneiderman [137], SocialFlow [184], SocialHelix [51], EvoRiver [168], TargetVue [52], WordCloud [179], ThemeCrowds [28], OpinionFlow [183], HistoryFlow [178], OpinionBlocks [94], Pearl [190], Matisee [165], StanceVis [114], Thread Arcs [108], we harmonise the taxonomy of entities in digital communication analysis (specific to E-mail communication) based on the data aspects the authors handled in the papers. Hamonising the taxonomy of entities based on the visualisations related to email communication data aids in understanding the different entities, the association between them and their limitations. Based on the above visualisations related to digital communication data, we extract four main types of entities, including temporal,

individuals, threads and contextual information from the digital communication analysis in general. Each entity includes three subcategories and the taxonomy of entities in E-mail communication analysis is given in Figure. 3.2. We also discuss the corresponding digital and E-mail visualisation papers for each entity. This part is discussed in Sections 3.3.4 and 3.3.5.

ENTITIES IN EMAIL COMMUNICATION (for Information Discovery)			
TEMPORAL INFORMATION	INDIVIDUALS INFORMATION	THREADS INFORMATION	CONTEXTUAL INFORMATION
Changes in Volume of Email	People's Overall Communication Network	Thread Features	Keywords / Topics
Temporal Gaps	People's Sent Network	Single Thread	Sentiments
Multigranularity	People's Reply Network	Multiple Threads	Complete Message / Text

Figure 3.2: Taxonomy of entities in Email Communication Analysis. It includes four categories with three sub-categories each. For the temporal information, analysts need to understand the change in volume of emails, find the gaps, reason behind them and explore time in granular form (years to months to weeks to days to hours). For the individual's information, analysts need to understand the overall communication pattern, also focus on sent and received (independently and in combination). For the thread information, analysts need to understand the thread features such as pace of interaction, inclusion/exclusion of individuals in the threads, also analyse from a single thread and multi-threads perspective. For contextual information, analysts need to focus on keywords/topics/subjects, sentiments and complete message/text.

In this work, we aim to address the question "To what extent visualisation can support analysts in finding/discovering relevant/interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation?". In most of the investigation processes, the exact notion of what makes information "interesting" is not well defined. In

that case, it is good to consider visual analytics, as suggested by Tamara Muzner [132]. Having investigators in the analytic loop improves investigation process as the visual tool aids in identifying anomalies, changes, patterns, trends and the investigators can continuously refine the search process until the desired results are found. We design and develop interactive visualisations that will support our collaborators in an organisation specialising in E-discovery to unravel the multi-faceted information in the given communicated E-mails to discover interesting information and to develop evidence through which legal cases can be built.

The challenges identified through literature and interviews are discussed in detail in Chapter 4. The next section will discuss the use of the visual design principles and the role of visual analysis in the digital communication data.

3.3 Visual Analysis & Investigation

Visual Analytics (VA), a term coined by Jim Thomas [60], is the combination of visualisation with automated analysis and analytic reasoning, facilitated by interactive visual interfaces, to explore a large and complex datasets for an effective understanding, reasoning, sense-making and decision-making. Visual Analytics is also called as Visual Analysis, Visual Data Analysis, or Visual Data Mining [106], where visual analytics is more than just visualisation. It is a multidisciplinary field that combines various research areas including visualisation, human-computer interaction (human factors), data analysis, data management, geo-spatial and temporal data processing and statistics [106][107][109]. The ultimate goal of visual analytics is to create an effective and efficient tools and techniques to enable users to identify the expected and discover the unexpected and get insight from the large, complex and ambiguous datasets. The design of the visual analytics tools and techniques is based on the scope of visual analytics listed below and they are not limited to [109].

1. Automation

2. Analysation (statistical, data, information and/or text analysis)

3. Interaction

4. Cognition and Perception

5. Visualisation (design and visual representation)

6. Presentation and Dissemination

In the early 2000s, researchers invested their time and effort to replace the manual classification, filtering and analysis by providing visualisation support for one's own inbox. By 2010, visualisation had fundamentally changed the way we present and understand contacts in our inbox and spam. Visualisation started emerging, tracking the revolution in the business world. The impact of visualisation has been widespread and fundamental, leading to new insights and more efficient decision making. From the year 2010 to 2015, most of the visualisations developed were helpful in understanding one's timeline or frequency of messages sent/received (discussed in the next section).

Since, E-mail communication data are being generated at an incredible rate, exploring and finding interesting/relevant information is a challenge. Visual analysis can be a potential solution in solving the challenges. Visual analysis can be used for various driven approaches: user-driven, data-driven, task-driven or problem-driven – this has given the cutting-edge to transform analysts (decision-makers) task processes effective and efficient. Still, decision-makers background knowledge and other relevant skills are also required to examine and gain insight into complex and challenging problems. Automatic analysis, visual analysis and visualisation methods can be used independently or in combination to solve analytical problems. All the three approaches independently give good results for small datasets but they fail to solve when the data is big. Visual Analytics combining all the three approaches produce effective and efficient results by considering the user, the task, the visual representations, and the characteristics of the datasets.

- Automatic analysis methods are used for measuring and comparing
- Visual analysis methods are used for analytical reasoning facilitated by interactive visual interfaces

- Visualisation methods are used in generating visual representations of data.

As E-mails continue to grow in an organisation account, there is a pressing need to develop an efficient and effective tool that is supportive, explorative and interactive to explore E-mail messages and visualise them for identifying, finding and discovering various information and relationships such that the tool improves one's cognition and help in managing large amounts of E-mails to support E-discovery compliance.

3.3.1 Visual Design Principles

Visual Design Principles are centered on the elements of design (EoD), which are used as components (such as lines, dots, shapes, colours and texture) in the design process and also focuses on the principles of design (PoD) which defines how the components are put to use for a good user interface (UI) and user experience (UX). We have considered only well-known and commonly used visual design principles such as Gestalt, Tufte and Sneiderman's principles that can help in serving as a basis for effective visualisation design for a multi-faceted context such as E-mail communication data. Also, the design principles can help us in building a visual interface that can support analysts in making better, faster and improved decisions in the investigations.

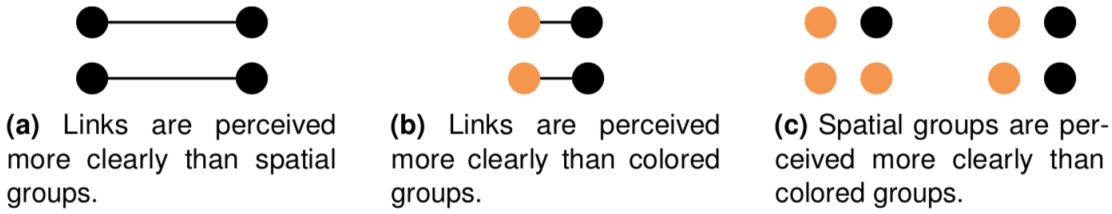
Gestalt Principles

Gestalt psychology founded by Max Wertheimer is a school of thought which deals with a theory of mind and brain: perception and learning. This theory describes the ability of the mind to recognise the patterns and shapes that are not present physically. Also, the theory explains how we use our brain while reasoning along with visual perception. There are a set of rules about human's perception towards visualising and interpreting objects (shapes), which are called as Gestalt Principles (also called as Gestalt Laws). Each of the laws are mentioned below. If readers are interested to know in detail, they can refer [160].

- **Continuity:** Elements arranged in a line or curve are perceived to be more related than elements not arranged in the line or curve.

- **Closure:** Elements arranged in a complex form makes one look for a single, recognizable pattern.
- **Common Fate:** Elements moving in the same direction are perceived to be more connected than elements moving in different directions or stationery.
- **Connectivity:** Elements connected by visual properties are perceived to be more connected than non-connected elements.
- **Similarity:** Elements that are similar tend to be perceived as a group.
- **Proximity:** Elements that are close to each other tend to be grouped together.

We consider connectivity, similarity and proximity for our review as these three principles are commonly used in the design and visualisation of graphics [160]. Connectivity has the strongest effect among these three principles, followed by proximity and then similarity. This comparison is illustrated in Figure 3.3. While spacing between dots in rows is shorter than spacing between dots in columns in Figure 3.3a, the linked lines make the vertical links have a stronger grouping effect than rows. The lines also make the horizontal links more notable in Figure 3.3b than colored circle groups. Two spatial groups are perceived more clearly than colored groups in Figure 3.3c.



(a) Links are perceived more clearly than spatial groups.

(b) Links are perceived more clearly than colored groups.

(c) Spatial groups are perceived more clearly than colored groups.

Figure 3.3: Comparison of Gestalt principles. Connectivity is stronger than proximity, and proximity is stronger than similarity. Image source: [160].

Tufte Principles

In his series of books, Tufte proposes a number of principles for the design of effective graphics, notably The Visual Display of Quantitative Information [171] and Envisioning Information [170]. This section reviews three principles commonly used in the design and visualisation of graphics.

- **Small Multiples:** small multiples can represent complex, multi-dimensional data into a simple comparison chart using plots, bars, lines, maps, heatmaps and other conventional charts. The role of small multiples can be used in comparative analysis which can reveal a range of potential patterns in the visualisations and this comparison by juxtaposition in small multiple displays was popularised by Tufte [171][170]. In general, small multiples are a set of similar design or graphs or charts that must have some logical order (for example, based on time) and must share the same measures, scales, axes, size, and shape which will help analysts/users to quickly find the charts that are interesting to them and be able to process information across many of these charts. Interestingly, several features can be superimposed, a good example of this principle is the visual analysis of road incidents data (Figure. 3.4).

- **Micro/Macro Composition:** this principle suggests that both complete details and an overall pattern can be contained in a visualisation. This allows user to look at the big picture from a distance to closely examine their individual pieces before sifting and drilling-down, which will convey both data and design ideas. A good example of this principle is the visual analysis of London Tube Map (Figure 3.5). Tufte refers to multi-layered graphs consisting of a large number of data points, but all combined to display a larger order of data. Through this greater visual representation, more meaning is given to the individual micro data by displaying all individual data items at a comprehensible level of detail and providing the distribution of data from an overview. This visual combination can be used to understand trends and relationships. In interactive visualisation, where zooming and panning are made possible, the micro / macro principle is widely applied.

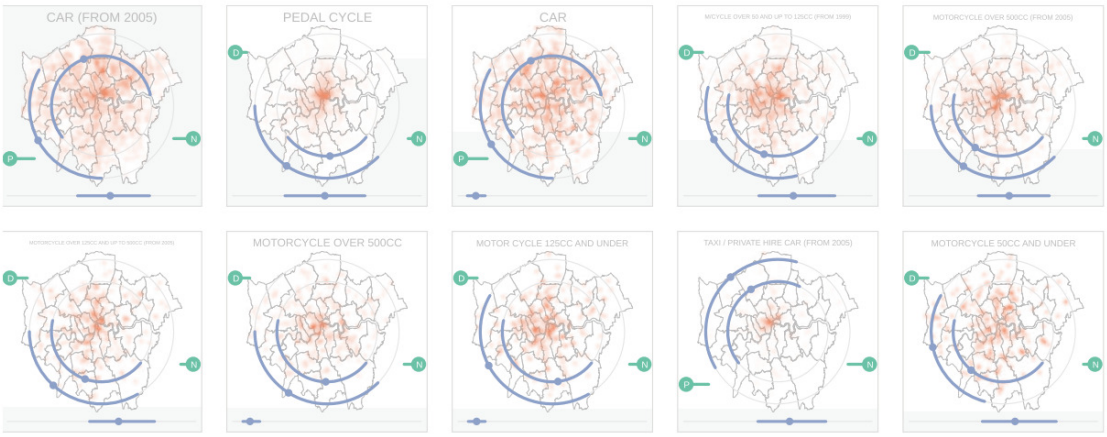


Figure 3.4: Small multiples with three perspectives summarising spatial (red), temporal (blue) and descriptive (green) are superimposed on each other to represent road incident data from London. Image source: [34].

- **Clutter:** Tufte suggests, if a design is too cluttered, do not remove data, change the design. Credibility comes from detail and a design can be clarified by adding detail in many cases. Also, one can remove unnecessary space, chart elements, overuse of dark colours, which will improve readability. Tufte also mentions that incomprehensible/cluttered visualisations can lead to confusion which leads to failure of design. Identifying and removing clutter from the chart reduces visual "noise", enabling users to focus on information or it is better to change the design itself.

Why Small Multiples?

Small multiples are profoundly used in investigation domains because they can represent complex, multi-dimensional data into a simple comparison chart which is generally a good solution as they are used with plots, bars, lines, maps, heatmaps and other simple charts. The role of small multiples within this comparative analysis reveal a range of potential patterns in the visualisations and this comparison by juxtaposition in small multiple displays was popularised by Tufte [171][170]. Small multiples are also referred as Trellis displays [33], collections [39], Polaris [167], dual-views [133], side-by-side views [116] and also juxtaposi-

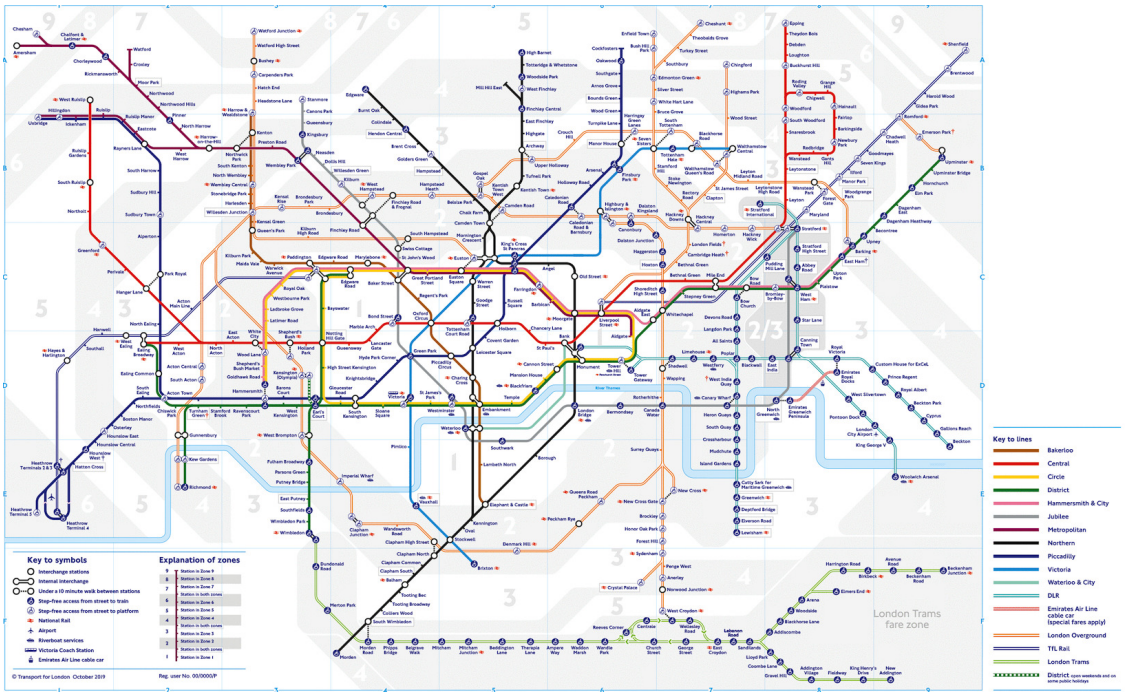


Figure 3.5: London Tube Map allows individuals to look at the complete map from a distance to closely examine their specific train routes/stations. Image source: TFL Gov UK.

tion views [81]. The commercial success of the small multiples approach credence from the fact that Tableau [27] demonstrated the usefulness in the early 2012. Small multiples are preferred due to their capabilities in representing large, complex, multi-dimensional, multi-granular and multi-faceted [104] data in the form of a compact comparative representation chart [173]. For a wide range of problems in data comparisons, small multiples are one of the good design solution [29] and they are used extensively in software development [149], finance [195], health care [139], social networks [43], sports [32], organisation analysis [187], crime detection [34], E-discovery investigations [63] and for many other purposes [191]. In general, small multiples are a set of similar design or graphs or charts that must have some logical order (for example, based on time) and must share the same measures, scales, axes, size, and shape which will help analysts/users to quickly find the charts that are interesting to them and be able to process information across many of these charts.

In the investigation domains, small multiples allow analysts to easily compare the differences in charts that are placed in a matrix form (horizontal and/or vertical, rows/columns) [75]. Small multiples are easily comprehensible for any non-visualisation experts and they have been widely used as a means to support various tasks such as identifying similarities, differences and anomalies.

Sneiderman's design & interaction Principles

In the work of Yi et al. [186], Heer and Shneiderman [87], and Brehmer and Munzner [47], high-level interaction taxonomies can be found. These classifications could help visualisation designers select the appropriate interaction techniques to serve their users' capabilities.

In order to explore the data or explain a story, interaction techniques are often combined. A classic visual information-seeking mantra proposed by Shneiderman [158] summarises many guidelines for information design and interaction techniques for effective visualisation of information: overview first, zoom and filter, then details-on-demand. The seven tasks suggested by Shneiderman [158] to develop a visualisation are:

1. **Overview:** We can gain an overview of the entire data.

2. **Zoom:** We can zoom in on the points of interest.

3. **Filter:** We can filter out uninteresting points.

4. **Details-on-demand:** We can select a point or a group to get details or information.

5. **Relate:** We can view relationships between entities.

6. **History:** We can keep a history of actions to support undo/redo, edit, replay for progressive refinement.

7. **Extract:** We can extract subsets of data by using query parameters.

Creating an overview and providing cues for further exploration is challenging with large datasets. Search, show context, expand on demand [174] is a more appropriate approach in this case. In order to improve existing interaction techniques, fluid interaction [72] can be applied. In addition to using direct manipulation as discussed above, the interaction should result in a smooth animated transition between the state before and the state after an interaction, helping users to maintain their mental maps. It must also provide immediate visual feedback, allowing users to know what's going on and/or what's going to happen next. The complete details of Shneidermans design and interaction principles (seven tasks suggested) [158] are discussed in Chapter 5 (design process & evaluation).

3.3.2 Visualisation Techniques/Methods

We derived different visualisation methods that are used in the digital communication data analysis. We consider different types of visualisation tasks/interaction techniques used in the papers. The next subsequent sections express how visualisations are used in the digital communication data analysis and more specifically to email communication data analysis. Some of the most commonly used visualisation techniques (conventional ones) are follows. **Basic Charts:** There are different types of basic charts. The most common ones are probably bar charts, line charts and pie charts. Bar charts are rectangular in shape that are used in representing different entities in the data, where the height or the width encodes

quantitative values (categorical variable), which could be either in horizontal (x-axis) or vertical direction (y-axis). The visualisation technique is well suited to represent relative differences in the communication data (Figure. 3.6(a)). Histograms are also rectangular in shape where each column encodes quantitative values (continuous variable), where skewness plays an important role; that is, the tendency of the observations to fall more on the low end or the high end of the X axis. Line charts are simple representation of lines (linear in shape) that are used in representing how different entities in the data changes over time. These can be called "timelines" as well. The visualisation technique is well suited to represent trends, analyse patterns, similarities and detect anomalies/outliers in communication data (Figure. 3.6(b)). Pie charts are circular in shape and are used in representing data that can be divided into different parts. In this way, the visualisation technique is well suited to represent relative proportions of multiple classes of communicated data (Figure. 3.6(c)).

Matrices: Matrix diagrams are either square or rectangular in shape representing the strength of relationship between pairs of items of two or more sets. The relationship can be indicated by a number or colour in each cell where the two items intersect in the matrix. The visualisation technique is well suited to visually query for patterns or outliers in a large amount of communicated data by individuals, where rows can represent senders and columns can represent receivers (this can be interchanged as well). The example is shown in Figure. 3.6(d).

Node-links: Node-link diagrams are circular dots (represents nodes) and lines (represents links) in shape representing relationship between entities in the data in the form of a network. The node-links can be represented using different layouts. The visualisation technique is well suited to represent relationship between senders and receivers in the communication data (Figure. 3.6(e)).

Scatterplots: Scatterplot diagrams are either points, dots or symbols in shape representing relationship between two variables in the data. Each position in the scatterplot is defined according to two dimensions produced. The visualisation technique is well suited to represent trends and correlations in the communication data (Figure. 3.6(e)).

Word Clouds: Word Clouds are in the form of text representing frequently appeared

words in a given data. The visualisation technique is well suited to represent frequency of keywords/topics used by the senders in the communication data (Figure. 3.6(f)).

The above-mentioned visualisation techniques could be used in building visualisations with interaction techniques (based on the tasks) such as overview, details on demand, searching, filtering, highlighting, and finding relationship between entities based on the Shneiderman [158] principles which summarises many guidelines for information design and interaction techniques for effective visualisation of information.

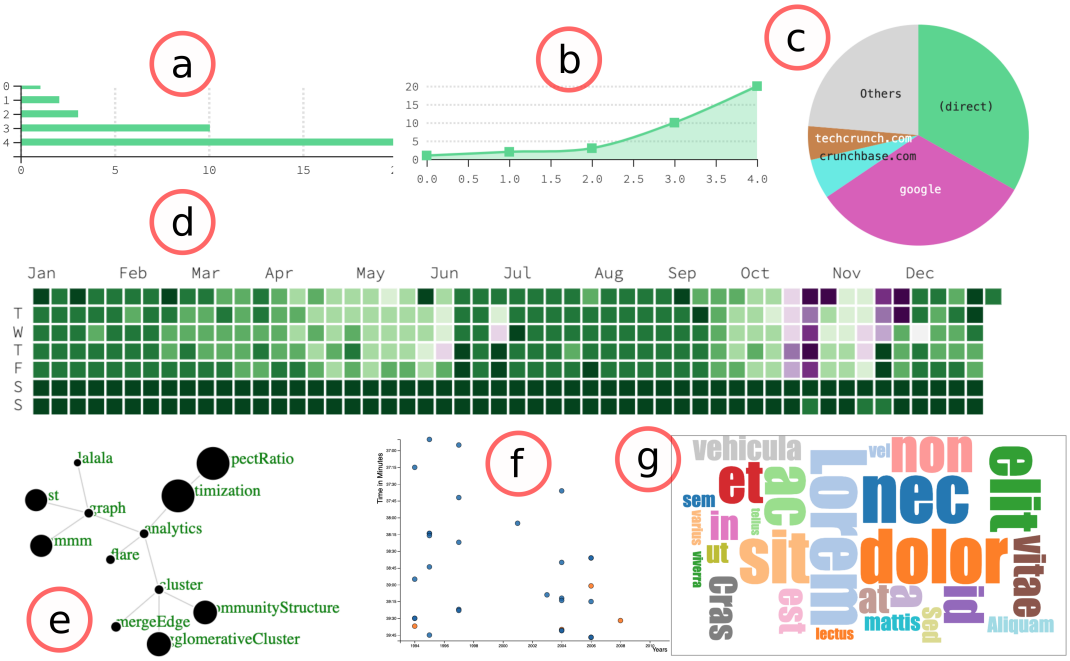


Figure 3.6: Some of the examples of visualisation techniques. (a) Bar chart (b) line chart (c) pie chart (d) Calendar matrix diagram (with a grid-based layout) (e) node-link diagram (f) scatterplot diagram (g) word cloud. All the above diagrams were generated as part of our design study using the Red Sift platform (organisation we collaborated with) which are discussed in Chapter 5.

3.3.3 Visual Analysis Techniques/Methods

Based on the literature survey, we identified different visual analysis methods that are used in the digital communication data analysis. We also considered different types of visualisation methods/interaction techniques used in the papers. Some of the most commonly used automated visual analysis techniques are clustering-based and classification-based, as they help in clustering/classifying various of group of people in social network based on temporal/textual activity/similarity.

Clustering-based Techniques: This is a commonly used technique to detect changes or observe patterns in digital communication data. Clusters are created for different timeframes based on different entities, which can be time frequency or any other metadata in the given communication data. In one of the works, FeatureForge [88] uses a visualisation to support hierarchical clustering (the example is shown in Figure. 3.7). In the clustering-based techniques, the mostly commonly ones used are k-means clustering and hierarchical clustering which are mostly represented using scatterplots and node-link diagrams.

Classification-based Techniques: This is also a commonly used technique to detect changes and the classifiers learn to detect the annotated events by themselves. Users don't need to create any specific rules for classifying a particular metadata in the communication data. In one of works, Heimerl et al. [89] explores the use of visualisation using a classification task (the example is shown in Figure. 3.7). In the classification-based techniques, the mostly commonly ones used are linear classifier and Support Vector Machine (SVM) classifier which are mostly represented using scatterplots and node-link diagrams.

The above-mentioned visual analysis techniques could be used in building interactive visual analysis with overview, details on demand, searching, filtering, highlighting, and finding relationship between entities based on the Shneiderman [158] principles which summarises many guidelines for information design and interaction techniques for effective visualisation of information. We further conducted research to understand there are several papers where visualisation can be used in feature engineering and active learning to improve data analysis.

Visual Analysis in Feature Engineering

The real data we want to analyse might have limited features and we might not be able to find/discover interesting patterns. In that real world case, the importance of feature engineering is to use domain specific knowledge and human insight to derive relevant features from the raw data to support analytical tasks. In this section, we discuss about visual analysis is used in supporting feature engineering.

Feature engineering can be described as the process of creating new features that help machine learning algorithms produce good results. Feature engineering is the key factor to the success of applying machine learning [69]. The process often requires domain knowledge; however, there have been attempts to automate it [102]. We discuss the use of visualisation in facilitating the manual feature engineering process. In our work, E-mail communication data has minimal features that does not aid in optimal analysis/results. The features can be engineered based on the requirements, domain knowledge or intuition to make models train faster and provide more accurate predictions. For example, an algorithm to predict the number of emails by a person with time. The person's emails peak during the evenings specially around weekends and bank holidays. Adding a feature that tells how many days we are away from weekends and bank holidays gives the algorithm a lot more intuition than the date itself. According to VIS4ML [151], in one of the goals, feature engineering provide visualisations of the feature in order to understand and improve the ML model with respect to feature selections.

Brooks et al. [48] conduct a controlled experiment to asses the benefit of using visual summary in creating new features compared to just using raw data. The participants add new keywords that they think can be used to classify web pages into cycling and non-cycling groups. The visual summary extracts keywords that could have largest impacts in the classification and display them as an interactive list. The results show that the visual summary condition outperforms the raw data condition in both producing more features and achieving better classification performance.

FeatureForge [88] supports the creation and exploration of new features using the out-

come of a classification algorithm. It includes a agglomerative hierarchical clustering with nodes colour-coded by the classes. This view allows examining potential problems with the model such as similar data records but having different classes. To provide additional insight, FeatureForge also visualises different attributes of features such as model weight and class distribution. In conjunction with creating new features, selecting relevant ones for model building is also vital. The example is shown in Figure. 3.7.

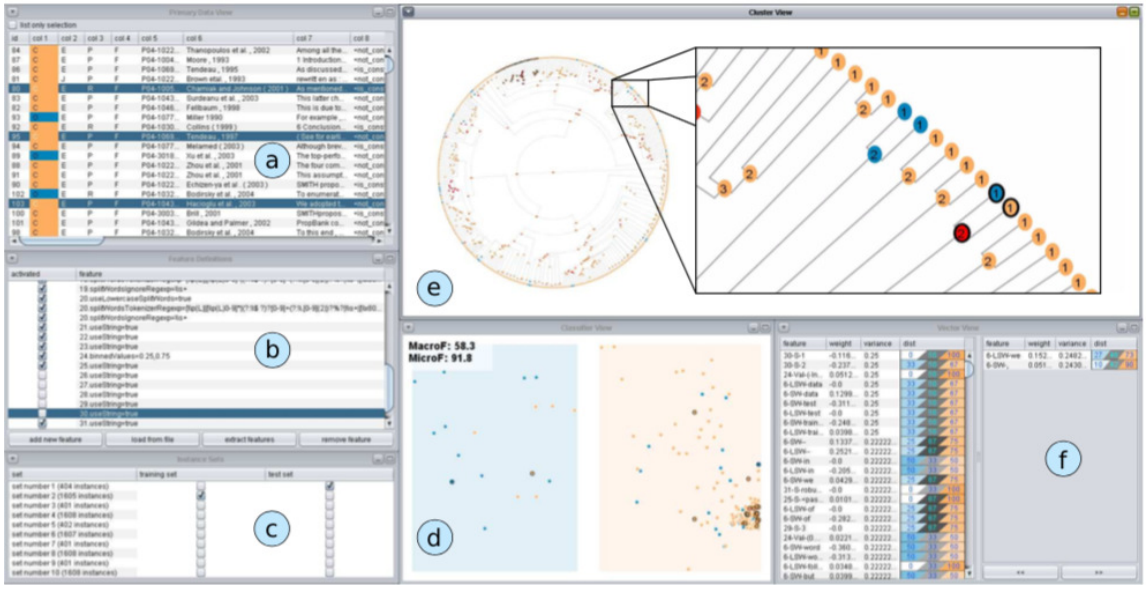


Figure 3.7: One of the examples of using clustering-based technique and classification-based technique with the support of visualisation (a) the Primary Data View helps in building a table of various instances, (b) the Feature Definition View helps in listing features for users to edit (add/remove), (c) the Instance Set View helps in selecting training set or test set of the data, (d) the Classifier View helps in visualising the classification model and understanding the classfier's state, (e) the Cluster View helps in understnading the hierarchical clustering of the test set computed, and (f) the Vector Set View helps in understanding the information about the attributes of selected instances. Image source: [88].

Janetzko et al. [152] present a visual analytics tool supporting different types of analysis of soccer data such as single-player analysis and event-based analysis. Depending on

the analysis type, the tool selects relevant set of features and allows users to customise the selected ones for model building. INFUSE [113] supports selecting features in high dimensional data by visualising the predictive power of features in machine learning models. Features are shown as glyphs revealing their ranks in different feature selection algorithms and facilitating feature comparison.

From the above-mentioned Visualisation-supported Feature Engineering papers, we can build an interactive visual analytic tool that can aid in building new dictionary features (semantically related groups of emails) for classifying various information of interest.

Visual Analysis in Active Learning

Active Learning (AL) in Visual Analysis (VA) is a special type of incremental supervised machine learning (ML) where the users are involved (user-in-the-loop) in the learning process to guide the training and analysis. In the AL, user will constantly query, annotate/label data and improve the quality of the learning model. In our work, Active Learning can be useful in cases where large portions of the email data are to be analysed that are unlabeled. For example, to understand the type of email communication between two individuals of interest, an analyst can label a particular set of data (based on their observation) as an instance and test for several data instances (which can be iterative). In the case of supervised machine learning algorithms often require a large number of labelled instances to perform well. However, the labelling process can be time-consuming and expensive. Active learning iteratively asks the user to label a small set of data with an aim to achieve high accuracy using as few labelled data as possible [157]. According to VIS4ML [151], in one of the goals, there is a training process involved with users to understand and assess the model learning process. Some of the papers are listed below.

Heimerl et al. [89] explore the use of visualisation with active learning for a document classification task. They compare three approaches in labelling the documents for active learning. First, documents are suggested by an active learner and labelling is based on the document text alone. Second, documents are also suggested by an active learner but an additional visualisation of classifier's result. Third, using a visualisation, an analyst has full

control over which documents to label. The results show that all three approaches lead to a good classification performance. However, Approach 3 produces less labelled instances, probably due to the sophistication of the interface and lack of training time. The example is shown in Figure. 3.8.

Figure 3.8: One of the examples of using active learning technique (visual classifier training) (a) the Search Bar for supporting the classifier, (b) the Main View showing the classifier's state (c) the Cluster View showing uncertain classification, (d) the Content View showing the selected or highlighted lists, (e) the Term Weight View showing the highest weights, (f) the Manual View shows what was used during evaluation, (g) the Classifier History to support undo/redo navigations, and (h) the Labeled Document View helps in listing labeled documents. Image source: [89].

Hoferlin [93], developed an inter-active learning approach, which extends active learning by integrating users' expertise for posing queries of data instances for labelling, annotating manually/automatically and adjusting complex classifier models. This helps in the detection and correction of inconsistencies between the classifier model trained by examples and

the user's mental model of the class definition. Moehrmann and Heideman [129], develop an advanced user interface and introduced active learning approach to facilitate the labelling of large image datasets, as the task of annotating large image data sets manually takes a lot of time and effort. The integration of overview+detail concepts allows the precise navigation inside large data sets.

Bernard et al. [38] propose "visual interactive labelling" process that address the differences and commonalities of the labelling process used in visual interactive analysis approaches and visual analytics. The iterative labelling process consists of four algorithmic steps (preprocessing & feature extraction, learning model, user suggestion and feedback interpretation) and two visual interfaces steps (labelling interface and result visualisation). Bernard et al. [36] also conduct an experiment to assess and compare the performance of different labelling strategies. The authors systematically compare the performance of visual-based active learning (interactive labelling) with that of basic method active learning, where the former outperformed the latter. During the experiment, participants use different strategies to select data for labelling. Those strategies are systematically analysed and formalised [37].

From the above-mentioned Visualisation-supported Active Learning papers, we can build an interactive visual analytic tool that can help in discovering important and interesting information in the E-mail communication based on classification and allow analysts to iteratively make changes to the observations. The feedback loop in the Visual Analysis pipeline can help to further improve the classification results.

3.3.4 Visual Analysis of Digital Communication Data

We identified visualisation of digital communication data papers closely related to E-mail communication data analysis. We specifically considered social media analysis. In these analysis, there are several relational components such as individuals, followers, re-postings, sharings, likes, emoticons etc. We restrict ourselves to time, people, networks, threads and text which is closely related to E-mail communication analysis which is of interest in this project.

Digital communication data contain features/facets such as temporal, individuals, and contents. It is often called multi-facetedness [104]. We consider each facet as an entity, which includes three subcategories. We discuss the corresponding visualization techniques for each entity. There are many visual analytic tools to sift through massive amounts of raw communication data (e.g., connections between individuals, contents exchanged, sharing behaviour, etc.) in order to discern patterns and trends. Each of the main approaches has three facets such as time, individuals and context. Each of these facets have three subcategories as shown in Figure. 3.2 (which can be considered in general for digital communication data as well).

Temporal-based approaches: understanding/analysis of temporal variations in digital communication data can be explored on many scales - daily, weekly, monthly and yearly to find temporal behaviour and/or any interesting patterns/trends. Visualisations help in identifying those variations in the communication data. The three subcategories in the temporal analysis based on our analysis are volume of communication, temporal gaps and multi-granularity.

- *Volume of Communication:* the total volume of messages sent and/or received by all the individuals involved over a complete period of time is extracted from the data to understand the overall communication. Whisper [50] helps in the overview of the temporal data and aids in filtering down the selected points. BCIT [79] is a bar chart over an interactive timeline visualisation tool for messages that can provide insights about the communication with others. This helps in understanding number of messages exchanged between the individual and their peers over time. Basic visualisations (e.g. line or bar charts) are mainly used to give an overview of the communication data by showing the time dependent relations. They are used to visualise the data volumes or frequencies over time.

- *Temporal Gaps:* the gaps in the temporal patterns help in understanding and identifying the reasons behind the patterns. EmailTime [99] helps in finding the temporal gaps (no data) such as holidays and/or trips.

- *Multi-granularity:* this offers multiple levels (higher dimension to lower dimension) of temporal granularity to understand the temporal sequences. Textflow [62] aids in drilling down from years to months to days to understand the text patterns.

A temporal-based visualisation is used in visualising the evolution of the communication data over time. Temporal-based visualisations are often combined with additional visualisations to show non-time dependent information. Some of the examples have all the three subcategories mentioned. They are Remail [147], Mailview [77], EmailTime [99], Email Timestore [188].

Individual-based approaches: understanding of individuals' communication variations in digital communication data can be explored to find anomalous behaviour and/or any interesting patterns/trends. Visualisations help in identifying those variations in the communication data.

- *Overall Communication Network:* the total volume of messages sent and/or received by all the individuals involved in the study is extracted from the data to understand the overall communication of the individuals.

- *People's Sent Network:* in the digital communication platforms, individuals contact each other by sending messages. This leads to visualising sent network of a particular individual. The analysis can also lead to visualising replying network.

- *People's Reply Network:* here the focus is on visualising reply network in the digital platforms.

Some of the examples have all the three subcategories mentioned. They are MatrixExplorer [90], Honeycomb [175], NetLens [101], Vizster [86], ZAME [71], MatLink [91], NodeTrix [92], SmallBlue [122], Perer and Shneiderman [137], SocialFlow [184], SocialHelix [51], EvoRiver [168], TargetVue [52].

Content-based approaches: understanding of contents exchanged between individuals in digital communication data can be explored to find anomalous behaviour and/or any

interesting patterns/trends. Again, visualisations help in identifying those variations in the communication data.

- *Keywords/topics:* keywords or topics are extracted from the texts exchanged between individuals which can basically represent the overall meaning of the content. One way to analyse the keywords or topics is to use WordCloud [179], where all the group of words are placed on a single plane, with the size indicating the frequency and importance. However, word clouds by itself cannot reveal deep insights of digital communication contents. However, ThemeCrowds [28] uses a hierarchical visualisation to examine keywords at different levels where the keywords are arranged chronologically. This facilitates thorough searches for keywords and allows users to discover and find keywords of interest. There are several other visualisations such as OpinionFlow [183] and HistoryFlow [178].

- *Sentiment:* Sentiment analysis helps in summarising people's sentiment towards various events or issues which helps in understanding the attitude of individuals for better understanding of the situation or circumstances. OpinionBlocks [94] use a matrix visualisation to encode supporting and opposite opinions with yellow and green color. Further, Pearl [190] is used to visualise a deeper classification of sentiment. There are several other visualisations such as Matisee [165] and StanceVis [114].

- *Complete Message:* Complete message exchanged helps in understanding the context and the situation better. Thread Arcs [108] has a message view which has an inbox list with a list of messages that can be selected to view the thread relationship visualisation in the preview.

To be specific: for digital communication data, as an application domain, most of the visualisation design studies on this type of domain focus on the abstract tasks of network analysis. Figure. 3.9 represents some of the visual analysis solutions for digital communication data. For the domain problem, many of these tasks are framed as identifying communities, key actors, and their roles. Based on the DSM [156], the tasks are abstracted

from the higher-level goals of identifying *relevant information* in the digital communication data. Some design studies in visualising communication data have indeed addressed a broader set of abstract tasks with respect to individual behaviour. However, in the background review study, several research papers were studied and following which, we classify visual analytics for digital communication data into time, individuals and contents. We further narrowed down to E-mail related papers (which will be based on threads) to understand the entities.

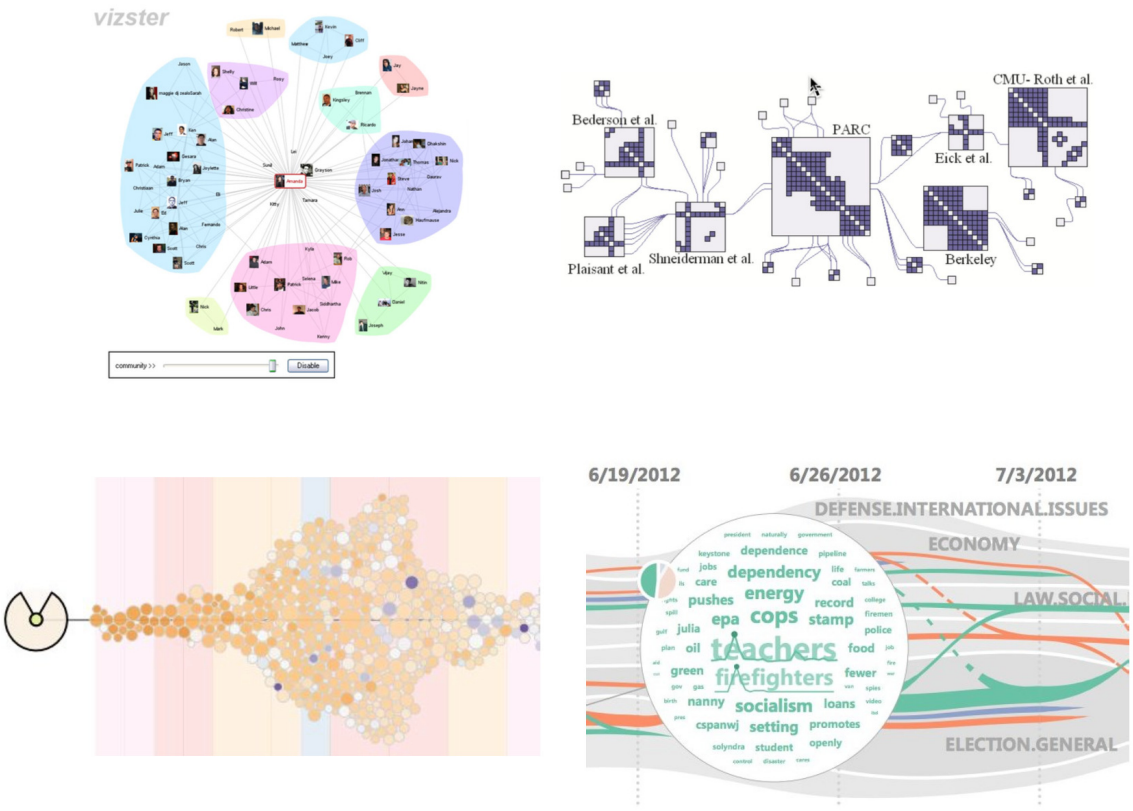


Figure 3.9: Examples of Digital Communication Data. L-R: (a) The Vizster [86] is used for visualising community network structures (the colored overlays represent communities identified within the network). (b) The NodeTrix [92] is used in visualising multi-level view of the underlying network. (c) FluxFlow [189] is used in visualising diffusion process of information on social media. (d) SocialFlow [184] is also used in visualising diffusion process of information.

3.3.5 Visual Analysis of E-mail Communication Data

Email communication data also contain multi-facetedness [104] such as temporal, individuals, and contents. There are many visual analytic tools to sift through massive amounts of raw communication data (e.g., connections between individuals, contents exchanged, sharing behaviour, etc.) in order to discern patterns and trends. To visualise E-mail communication data relevant to the E-discovery context, we start understanding the E-discovery tools used by analysts.

The experts/analysts use E-discovery tools such as Brainspace Discovery5^{TM} [4], Jigsaw [164], Concordance by LexisNexis [5], IN-SPIRE [6], Radiance [7], Zovy Advanced E-discovery (AeD) [8], DocuBurst [58], RingTail to investigate and analyse electronic documents. These tools support investigative process by mapping relationships (people, places, things etc.) found in datasets. Many investigative teams use these tools for identifying key information/relationships within data but sometimes the tool must be used in combination with other tools to carry out investigations. To the best of our knowledge, the above-mentioned tools do not focus or completely support investigating E-mail communication data.

We again exhaustively searched the literature to identify E-mail forensic tools available in the real-world. We again choose the following five open source E-mail tools which are popular and widely used and they are MailXaminer[3], Add4Mail[4], Digital Forensics[5], EMailTrackerPro[6], and Paraben E-Mail Examiner[7]. These tools aid in searching information and recovery capability. To the best of our knowledge, though the tools provide basic charts but none provide interactive visual exploration. As an observation, only Add4Mail tool can analyze emails stored both in hard disk (offline analysis) and on remote email servers (online analysis).

[3]http://www.mailxaminer.com/

[4]http://www.aid4mail.com/

[5]http://www.digital-forensic.org/

[6]http://www.emailtrackerpro.com/

[7]http://www.paraben.com/email-examiner.html

Point: from our analysis, we understand the above-mentioned E-discovery visual analysis tools are useful only for the analysis of *electronic documents* and not for *electronic data*, as the latter involves more complex information such as time-stamps, contacts, connections and contents.

As we explored different tools and techniques pertaining to E-mail visualisation, we recognised that some work has similar approaches. In these approaches, we identified two different perspectives at the high level. The first strand of papers focuses on email thread and response chain (thread-based category), while the second aspect focuses on just network and communication (non-thread-based category). We identified two main approaches in terms of visualisation techniques, which we grouped into categories based on visualisation methods, tasks and communication behaviour. Each of the main approaches has three facets such as temporal, individuals and context. Each of these facets have three subcategories as shown in Figure. 3.2.

From the background review study and analysis, several research papers were studied and following which, we classified visual analytics for E-mail communication data based on the three types of communication behaviour, that is temporal, individual and content behaviour. We derived six different visualisation methods (Basic Charts, Matrices, Node-links, Scatterplot, Treemaps, and Glyphs/Cloud) and six different visualisation tasks (Overview, Details on Demand, Searching, Filtering, Highlighting, and Relationship b/w Entities) that are used for "information discovery". As a result, we included 8 papers (email visual systems) in total for our review/learning to understand the visual methods/techniques, tasks/interactions, and features used; and any specific design studies conducted to build the systems. They are discussed below:

Thread-based Visual Analysis

E-mail communication data in specific also contain multi-facetedness features/facets (contains temporal, individuals, and contents) There are very few visual analysis tools/solutions to sift through massive amounts of raw communication data (e.g., connections between individuals, contents exchanged, sharing behaviour, etc.) in order to discern patterns and

trends for E-discovery investigations. There are different types of E-mail Visual Systems for thread-based data and they are as follows:

Re-Mail

Re-Mail [147] is a part of IBM's larger project on "reinventing email". The research group has been investigating how people use email and what solutions would help users. The authors developed Re-Mail which incorporates novel visualisations for the thread-based E-mail communication to aid in understanding and navigation. There are no specific user requirements in this work. Re-Mail [147] uses a real dataset Enron which focuses on the temporal aspect and individual involvement in email communication in thread structures.
Design and Visualisation System: The authors developed Re-Mail which incorporates novel visualisations such as Email Map, Thread Map, Correspondent Map and Message Map of the E-mail communication within mail databases to aid understanding and navigation. Though there are no specific user requirements in this work, the authors have considered several features in the Re-Mail system [147] which are mentioned below and represented in the Figure 3.10. The authors have developed a ThreadMap (A) which has a set of email threads, where email thread is a series of replies to a message and the replies to those replies. This map is used for navigating among related messages in a particular thread, where the Correspondent Map (B) is used for highlighting the senders of messages. In addition, there is a Time line (B) which is used to represent days with little or no activity. These visualisations are useful for showing the thread structure at a high-level, but they do not provide any clues about the individuals who were copied (cc), secretly copied (bcc), individuals who were active/passive in the complete thread.
Empirical Evaluation & Outcome: The authors have created design mockups, conducted surveys and interviews of email usage and gathered user experience data. The details are not mentioned in their publications.

Thread Arcs

Thread Arcs [108] is a thread-based E-mail communication visualisation that uses arcs to represent the messages relationship. By quickly scanning and interacting with the visualisation system, users can see various attributes of conversations and find relevant

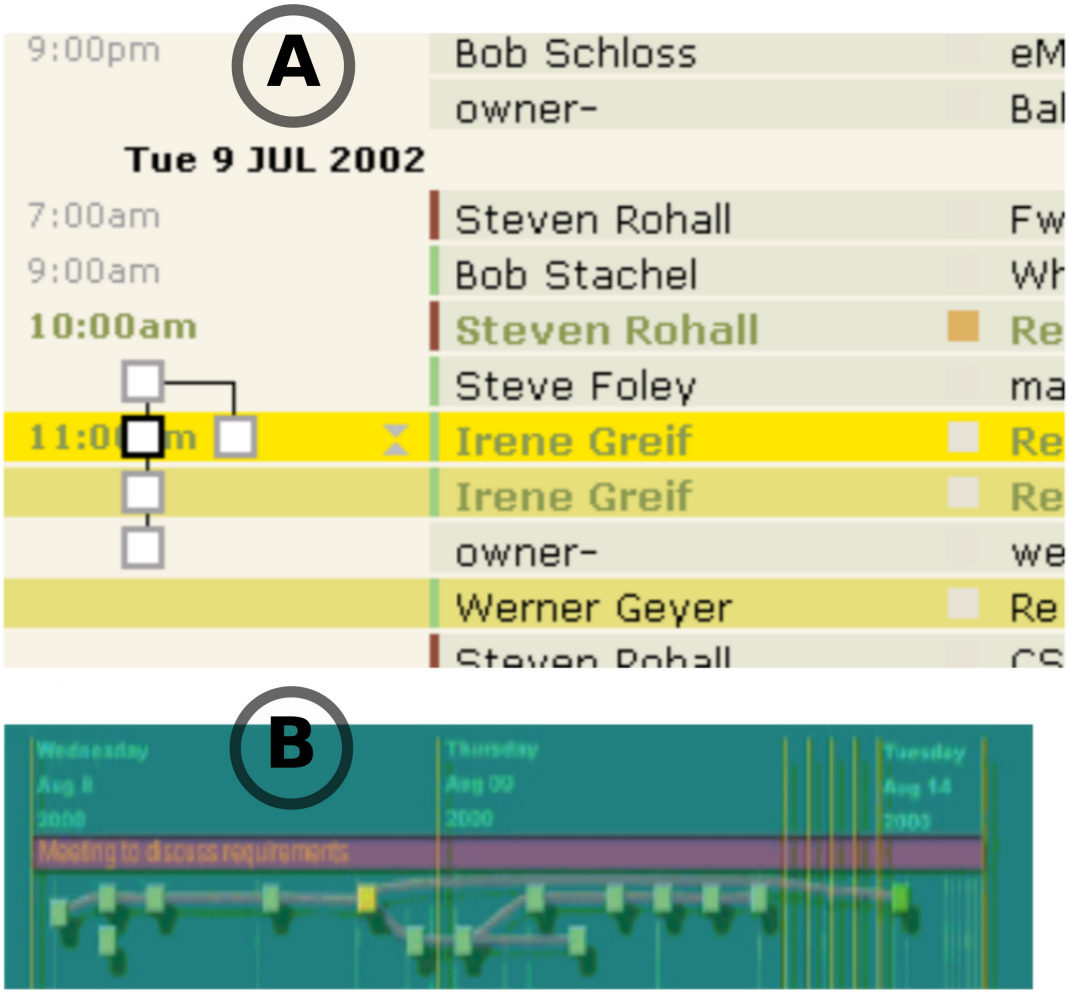


Figure 3.10: Re-mail: (A) visualisation represents the combination of thread map and the correspondent map. (B) represents timeline of the thread. Image Source: [147]

messages in them easily. Though there are no specific user requirements in this work, authors used a real dataset which focuses on the thread structures.

Design and Visualisation System:

The authors have taken inspiration from Tree Diagrams and Tree Tables visualisation, as they are common way to represent threads. Unfortunately, these two visualisations cannot represent chronology and are not stable. Thread Arcs [108] have a linear layout in the horizontal direction of message nodes connected by relationship arcs with a chronological order. Each circular node in the Thread Arc is a thread message making the complete design a stable compact visualisation. The users can see various information such as thread size and number of responses per message. Threads that have messages that receive two or more responses are defined as bushy, while threads that have messages that receive only one response per message are called narrow. The authors have considered several features/views in the Thread Arcs system [108], as shown in the Figure 3.11, where the arcs represent the relationship/connection between senders and receivers in an email thread. The work can be extended to visualising thread features (based on their engagement/activities/email behaviour), tracing the sender/receiver relationships based on the thread features (such as interaction pace) and can also integrate email subjects and contents to give a more picture to an investigation case.

Empirical Evaluation & Outcome: The authors conducted the study to learn about the usefulness and effectiveness of email thread visualisations in users' personal email inbox. In particular, the authors investigated chronology, message relationship, stability, scalability, compactness, attribute highlighting and interpretation. The participants have expressed their positive attitude towards all the above-mentioned thread qualities. Thread Arcs provide insight into the structure and evolution of email conversations in a concise manner, however it does not help in comparing multiple threads of interest.

Beyond Threads

The authors of the Beyond Threads [136] developed an interactive visualisation for users to explore, perceive a threaded discussion and navigate over time and individuals in the messages to gain the context they need. There are no specific user requirements in this

Figure 3.11: ThreadArcs: the complete visualisation to reprsent E-mail threads. (A) the visualisation in the preview pane shows the selection highlighting scheme based on the message selection (B). (C) Thread view represents the relationship between senders and receivers using arcs. (D) There are two drop-down menus which allows users to apply attribute highlighting schemes. (E) The senders (contributors) and recipients of a particular thread selected can be seen. (F) A list of all the messages in the thread with author and subject will be displayed. (G) The start point and finish point of the thread will be calculated. Image Source: [108]

work. The authors used a real dataset which focuses on the temporal aspect and individual involvement in email communication in thread structures.

Design and Visualisation System: The authors [136] present a novel interaction technique for exploring temporal data in the threaded email communication. In this interactive visualisation, users have quick access to all of the individual's messages during various time periods, as well as their conversation with other individuals. The authors have considered mainly temporal and individuals as features in their system [136], as shown in Figure 3.12. However, it would have been interesting to see if the messages/contents were integrated.

In the current system, we can identify the senders, direct receivers and copied (cc'd) individuals. However, understanding the volume of emails sent by the senders or received by the receivers is not possible. However, users also have the option to filter out nodes that are not of interest. It is possible to remove irrelevant individuals and messages from the visualisation (such as senders and/or recipients). This gives us an idea to introduce a classification approach to classify relevant/interesting threads of interest.

Empirical Evaluation & Outcome: The authors have not reported on any design mockups, surveys or interviews.

EmailMap

Email Map [124] is a thread-based visualisation tool that has two visualisation integrated components, contacts list and event flows, for users to understand contextual information and some interesting patterns between individuals. The contacts list shows the interaction between the email inbox owner and their contacts. The event flow shows the evolution of events in the email threads, helping the analysts to understand various events that took place in the communication. There are no specific user requirements in this work and the email dataset considered is from a personal account (real one).

Design and Visualisation System:

EmailMap [124] visualises the evolution of email and the contact relationship simultaneously (dual design focus) to find interesting patterns of how people and events are interrelated over time. It clusters email messages into a thread-based hierarchical email framework. The hierarchy is visualised as a stream that shows the evolution of associated

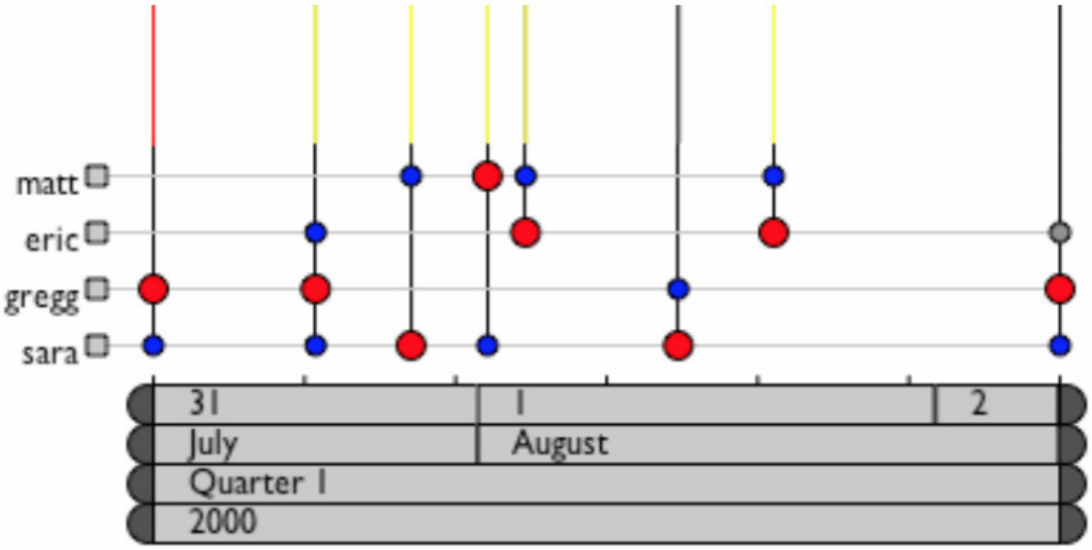


Figure 3.12: Beyond Threads: Visualisation of the discussion is represented. Each vertical line depicts an email oriented along the horizontal axis depending on the time it was sent. All the individuals involved in the conversation are listed on the left. For each message, an individual is addressed, a coloured circle is drawn aligned with their name and connected to the linear representation of the email. The senders are coloured in red, direct recipients are in blue, and copied recipients (cc's) are coloured in gray. The circles of the sender have a slightly bigger size emphasising the significance as the author of the message. Image Source: [136]

messages over time. The emails are categorised as a stream of events and the contacts are represented as curves, linking messages from the same senders, superimposed on the stream to highlight the interaction of contact. The horizontal axis represents time moving from left to right (in chronological order). A contact is shown as a curved track through the emails he / she has participated in. We can inform the frequency of email exchange between the contacts by tracing a contact track. The color is used to differentiate distinct contacts as shown in the Figure 3.13.

An email clustering method is considered for grouping email threads into significant chronological life occurrences based on email content and contacts. The branching technique helps in demonstrating how one event flow evolves into two or more sub-events. This gives us an idea of implementing a visualisation to show the inclusion/exclusion of individuals in an email thread (that is, when a particular individual in included or excluded from the discussion).

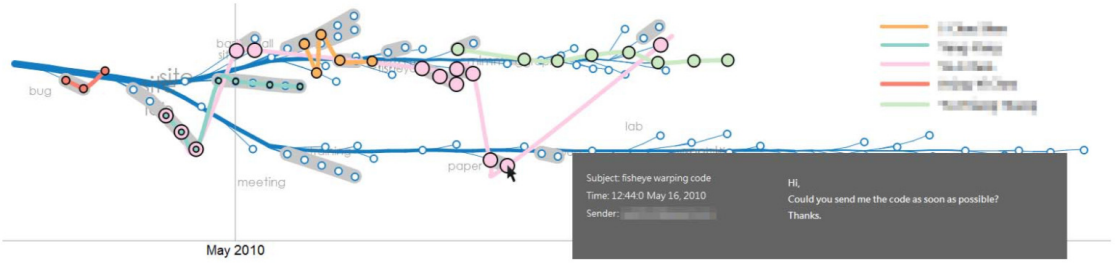


Figure 3.13: An example of EmailMap. Each email is represented as a circle. The blue color flow represents the event evolution in email communication over a period of time, and the color tracks reveal the interaction between the individuals. Image Source: [124]

Empirical Evaluation & Outcome: from the paper review, there are no empirical studies conducted. However, we know that the authors have integrated the two visualisation parts into one by incorporating event evolution and interaction between individuals over time. This novel strategy enables users to make sense of their own data with complementary context information and comprehend the overall pattern of email communication. The

visualisation works mainly for analysing one's email archive and it doesn't serve an investigation team. The tool is not efficient to match analyst's mental model of understanding and investigation.

Non Thread-based Visual Analysis

There are different types of E-mail Visual Systems for non-thread based data and they are as follows:

EmailTime

EmailTime [99] is an email explorer tool for visual analysis of email communication patterns over time that interactively portrays personal and interpersonal networks, which helps in exploring, interaction and visualising histograms. EmailTime [99] uses a real dataset Enron which focuses on the temporal aspect and individual involvement in email communication without thread structures.

Design and Visualisation System: though there are no specific user requirements in this work, the authors have considered three main features in the EmailTime system [99]:

1. Time Comparison - to compare different periods of time and identify time gaps (no activity), emails with large recipients etc.

2. Email Address Comparison - to compare different email address and identify duration, activity level and role (sender, receiver or both) of each email address or individuals.

3. Frequency Analysis - to find individuals who frequently correspond and type of communication (private or general messages) based on the sent emails.

EmailTime [99] uses a grid-based layout to organise emails: addresses are shown as rows and messages are shown as columns, positioned according to the timestamps (shown in Figure 3.14). The system uses interactions such as zooming, panning, searching, filtering, highlighting in both the visualisation and the control panel. This paper mainly explore techniques for the visualisation of temporal relationships of emails without threads. Considering only temporal relationships does not add value in the investigations, we need to combine more features to find interestingness in the data.

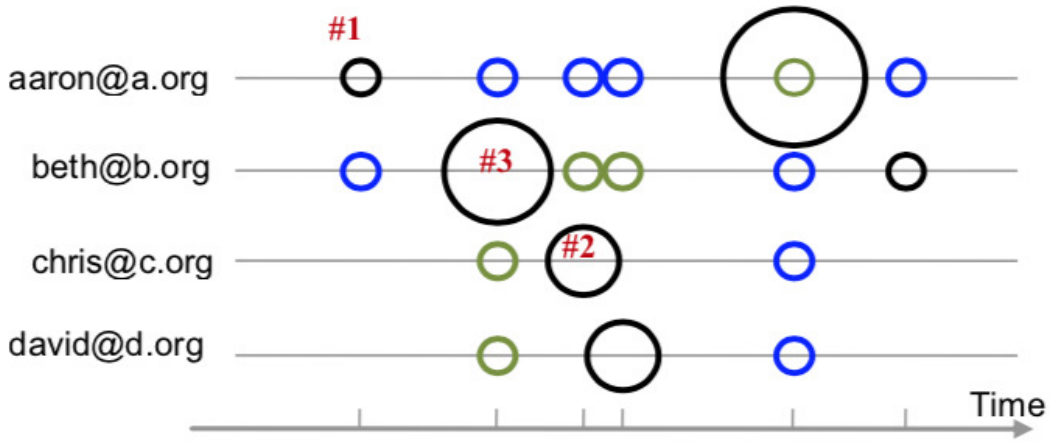


Figure 3.14: EmailTime: the visualisation represents email messages over time. A message has three different colors; black for Sent, blue for Received (To), and green for Received (Cc). The size of a Sent node represents the number of recipients. Image Source: [99]

Empirical Evaluation & Outcome: The authors conducted pilot studies followed by a user study with the graduate students. The study had five scenarios with multiple tasks for the participants to navigate, explore and explain/answer using Think Aloud Protocol. The tool helps an analyst visualise various patterns of communication over a period of time selected and to analyse their email behaviour. Interaction is an important aspect of the tool as it provides zooming, panning, filtering, and highlighting which will help analysts make more sense out of a collection of emails.

Themail

Themail [177] is a visualisation tool that uses communication histories to portray relationships, specifically used for non-threaded E-mail communication. Using the content of the exchanged messages, it demonstrates the phrases that characterise one's communication with another person and how they alter over the relationship span. There are no specific user requirements in this work and the email dataset considered is from a personal account (real one).

Design and Visualisation System: Themail [177] helps explore the content exchanged be-

tween two individuals through the words that characterise their conversation. Words are shown in monthly columns along a timeline and scaled by their frequency and distinctiveness (based on different colors and sizes), as shown in Figure 3.15. There are two interaction modes: exploration of overall trends and themes in email (high-level mode) and more detail-oriented exploration (low-level mode). The system uses interactions such as searching, filtering and highlighting for exploring contents exchanged between individuals.

The system helps in understanding the words exchanged between the two individuals but it does not help in understanding which other individuals were involved in the conversation (copied or secretly copied). Although the system has considered three main entities (time, individuals and content) it cannot support investigation cases. The work can be enhanced by considering threads.

Empirical Evaluation & Outcome: Participants were encouraged to upload their mailboxes (sent and received ones) to the Themail system. Based on the rating scale, 87% of the participants enjoyed looking using the tool to visualise their email communication. Approximately 80% of the participants used Themail in high-level mode, while 20% used low-level mode visualisation. This group of participants seemed more interested in general patterns than in picking apart individual words that emerged in the visualisation. They showed more interest in their personal communication rather work-related. The other 20% of the participants were more interested in discovering particular parts of information rather than concentrating on general visualisation patterns. Participants in this category showed little interest in seeing the visualisation of their personal communication, being more worried with visualising work-related communication.

SeeMail

SeeMail [65] is a web-based system for visualising email response patterns for non-threaded communication. SeeMail processes user email headers to produce visual summaries of response behavior. There are no specific user requirements in this work and the email dataset considered is from a personal account (real one).

Design and Visualisation System: SeeMail uses visualisation methods to display patterns

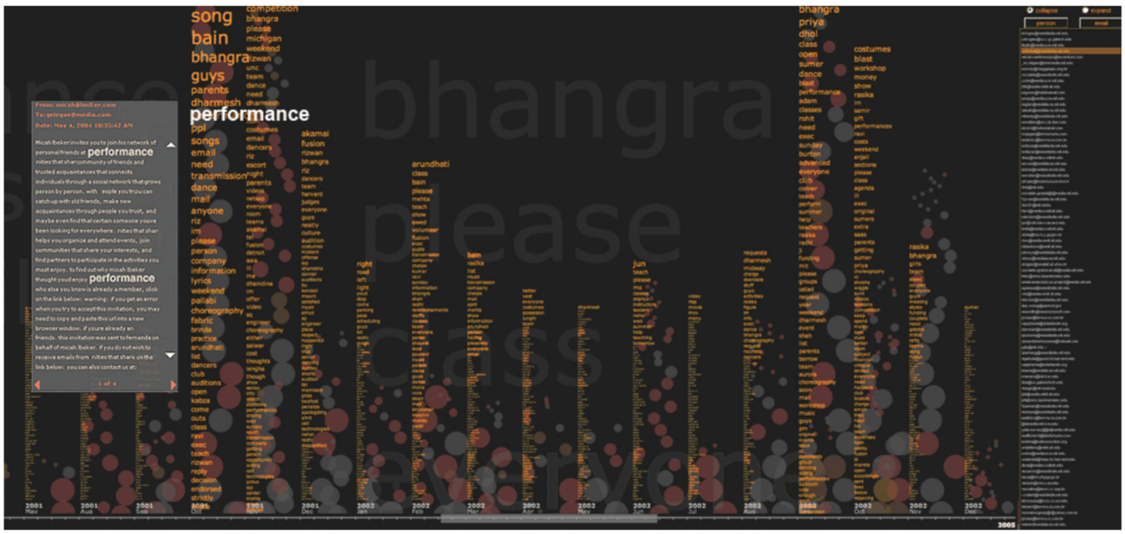


Figure 3.15: Themail represents an individual's email communication with a friend over a period of time. For yearly words, gray words in the background and monthly words yellow words in the foreground are arranged in columns. The information for the year and month is displayed at the bottom of each column. Each colored circle reflects an email message exchanged over a specified month. The circle size is the length of the message and the circle color is the direction of the message: incoming or outgoing. Image Source: [177]

over time and provide direct comparisons to make hidden communication patterns visible to users. The system is intended to assist users to understand their email behaviour. The user's response delay to an email and the inter-temporal interval between when an individual responded to a received email form the complete system. SeeMail involves six visualisations: overview, trends, comparisons, intervals, stripes, and summary table, as shown in Figure 3.16.

SeeMail helps in producing response time of individual's reply. This gives us an idea to generate email response feature and engagement feature using feature engineering techniques that can be supported using visualisations. SeeMail has a minimal interaction technique to select interested contacts and switch between the visualisations, which is a drawback for using in investigation cases.

Empirical Evaluation & Outcome: The authors conducted a think-aloud study with 20 participants who uploaded their email to the SeeMail system and viewed all six visualisations. Based on the study, the participants have felt easy to interact and interpret the information. They have also expressed concern about relationships between the contacts.

Beagle

Most recently, Beagle [111] supports investigative analysis in email data through linked visualisations of complex data query, correspondent connectivity and entity extraction from email content. The visualisation tool was formerly called as InVest. There are user requirements captured in this work in such as reduce duplicate, unimportant emails and find anomalies in the data. The email dataset considered is Enron and also from a personal account (real one).

Design and Visualisation System: Beagle [111] extracts different words from email contents (emails exchanged) such as names, organizations, locations, dates etc. This allows a clear separation between the words captured and individuals involved in the discussion, as shown in the Figure 3.17. Beagle uses Network Graph View where a node represents a keyword extracted and the thick edges represent the number of messages exchanged. The Bar Graph view (height of weekly bars) represents the number of messages sent during that week. The Bipartite Graph View represents the outcomes of the most popular senders and

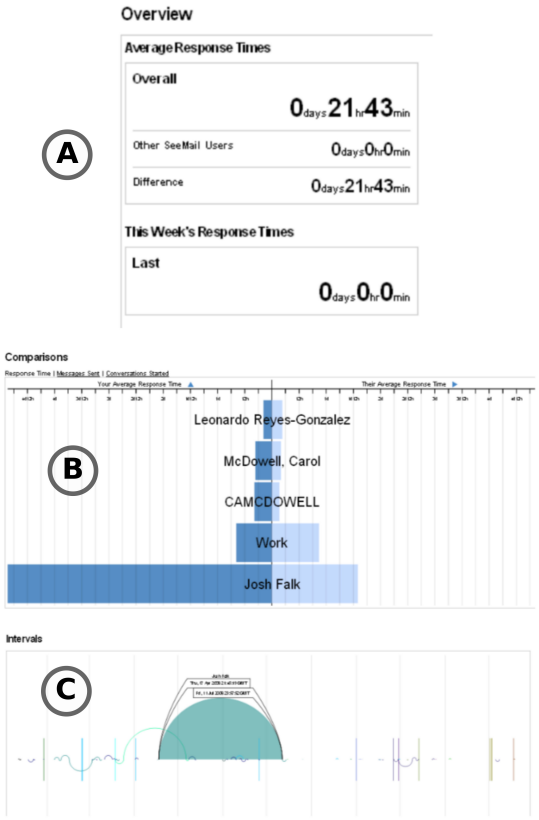


Figure 3.16: SeeMail for visualising email response patterns. (A) The Overview visualisation provides users with a summary of user's overall reply time across all email communication. (B) The Comparisons visualisation helps in understanding reply time of a contact to an incoming message. (C) The intervals visualisation show reply time on a per-email basis over time using an arc. Image Source: [65]

receivers. The thickness of edges between senders and receivers reflects the strong links between the two selected sets. Beagle has a good interaction mechanism to search, select interested contacts, keywords and switch between the visualisations, also filter and expand. General text analysis tools such as Jigsaw [164] can also be used for this type of investigation. The tool can be enhanced by introducing more features using feature engineering techniques, also aid in comparing multiple threads and a single thread of interest related to an investigation case.

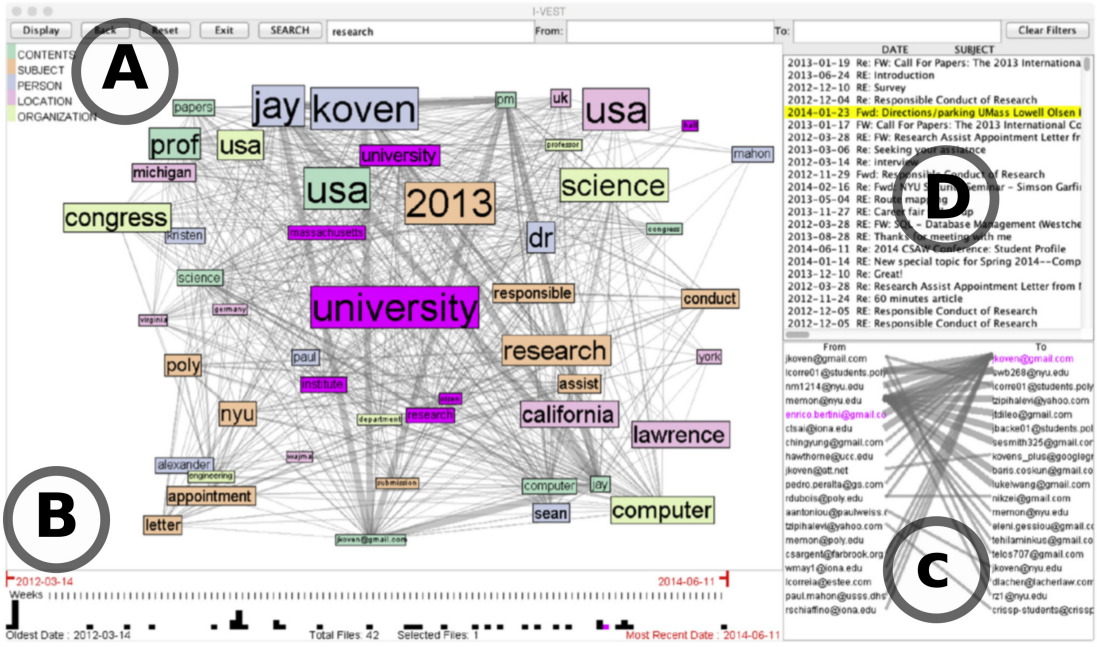


Figure 3.17: Beagle: represents network graph view (A) temporal bar graph view (B) bipartite graph (C) and subjects view (D). Image Source: [111]

Empirical Evaluation & Outcome: As part of the evaluation, the Enron case studies were conducted by authors after consultations with a professional investigation team. The visualisation revealed the problems encountered and listed by the investigation teams in the past.

In addition, we identified, a couple of more tools:

E-mail Personal Analysis We exhaustively searched the literature and various web page resources to identify E-mail personal analytic tools available in the real-world. We choose the following two open source E-mail tools which are popular and widely used. Again, none support all the three main criteria for E-discovery compliance within organisations. Also, there are many existing E-mail tools that offer metrics on one's E-mail history, that is E-mail behaviours that include: hourly/daily/weekly/monthly inbox volume; top senders and recipients; most active hours; average response time; word count; and attachments. This allows users to compare their behaviour within a community.

Gmail Meter[8] The tool was developed by ShuttleCloud Corp is the most complete form that includes all the said features but it is limited to a monthly report for Gmail accounts. However, Gmail Meter does not allow comparison between threads considering months/weeks/hours. Also, it is does not give you the flexibility to understand the contents exchanged.

Immersion[9] The tool was developed by MIT Labs provides an user-centric perspective of one's email history. It creates a social network analysis graphic representing one's E-mail communications and how they change over time. A tool like this helps users reflect on their E-mail behaviours and their high and low frequent contacts. However, it is not easy to analyse without considering threads and the contents exchanged.

The above-mentioned E-mail visualisations are quite good in exploration but they do not support in classifying email threads that are unlabelled and uncategorised. Based on the discussions with the experts (included in the Appendix A.7), the current visualisation tools do not support in the complete end-to-end investigations. Hence the need of visualisation assisted feature engineering and active learning, as there are no specific works on E-mail communication analysis.

[8]http://gmailmeter.com/:

[9]https://immersion.media.mit.edu/:

3.4 Key Findings

The review on email visualisations helped us to understand the visualisation methods, tasks/interaction techniques and features that can be used for investigating email communication between individuals. The tables will quickly help us understand how each visualisation tool/system considered the features, visualisation methods and interaction mechanisms. We derived 6 different visualisation methods that are used in the email communication analysis. There are number of tasks/interaction techniques used such as overview, details on demand, selecting, searching, filtering, exploring and relationship between entities.

Based on the visualisation techniques/methods: In the literature survey, we identified various visualisation methods (including digital communication data in general) such as basic charts (bars, lines and pies), matrices, node-links, scatterplots and wordclouds. More than half the surveyed papers use basic charts and matrices to investigate data. Many other papers use node-links and other complex visualisations to investigate. Bar charts are used in representing volume of communication and help in facilitating the task of visually comparing various groups in the chart. Line charts are used in representing one or more entities changing over time (number of messages of different users). Node-link diagrams are used for analysing various network relationships between senders and receivers.

Specific to email visualisation, Re-mail [147], SeeMail [65] and Beagle [111] use basic charts to visualise information in the E-mail data. Node-links are used by Email Maps [124], ThreadArcs [108], and Beagle [111]. It is important to highlight that some of the visualisation techniques and approaches are being considered in combination. For example, Beagle uses a combination of node-link diagrams, bar charts, and word cloud for visualising email communication. Clustering techniques are used to improve the node-link visualisations during interaction. The only drawback of using node-link diagrams is the scalability problems such as visibility, interpretability, usability, and high degree of nodes are likely to appear. Beagle also uses a combination of temporal-based, individuals and content-based approaches to find interesting information. It is also important to note, Be-

yond Threads [136], EmailTime [99] and few other visualisations related to communication data (based on our survey) use grid-based visualisation to support clarity, order, layout and consistency in visualising structures. The complete comparison table is shown in Table. 3.1.

Table 3.1: Comparison of visual analysis tools: *based on the visualisation methods*

VA Tools	Basic Charts	Matrices	Node-links	Scatterplots	WordClouds	Grids
Thread-based						
(1) Re-mail	*Yes*	*No*	*No*	*No*	*No*	*No*
(2) EmailMap	*No*	*No*	*Yes*	*No*	*No*	*No*
(3) Beyond Threads	*No*	*No*	*No*	*No*	*No*	*Yes*
(4) Thread Arcs	*No*	*No*	*Yes*	*No*	*No*	*No*
Non Thread-based						
(5) EmailTime	*No*	*No*	*No*	*No*	*No*	*Yes*
(6) SeeMail	*Yes*	*No*	*No*	*No*	*No*	*No*
(7) Themail	*No*	*No*	*No*	*No*	*Yes*	*No*
(8) Beagle	*Yes*	*No*	*Yes*	*No*	*Yes*	*No*

Based on the visualisation tasks/interaction system: In the literature survey, we identified various visualisation tasks (which are actually interaction techniques) such as Overview, details on demand, searching, filtering, selecting/highlighting and relationship between entities can be found in the most of the papers. Almost all surveyed papers (including digital communication data in general) provide highlighting of the selected data. From various papers surveyed in this project, the most popular interaction techniques used in digital communication data analysis are filtering and highlighting. One of the interesting observations is, papers in the past (early stages) related to digital communication data analysis dis not provide interactive features or techniques. Our interpretation is maybe during early times, interaction techniques were more cumbersome to be designed and implemented.

Specific to email visualisation, among all the visual analysis tools, Beagle [111] has considered all the visualisation tasks and the authors feel it is quite important to consider various visualisation tasks (interaction techniques), as it will help in filtering down the irrelevant data [111]. Overview visualisation give users a summary of the data analysis. Not

all summaries serve as navigation support but are often used as starting point for further analysis. Adding searching and filtering will in reducing irrelevant data especially when analysts are investigating to find some relevant information. Surprisingly, almost half of the surveyed papers only provide filtering and show their analytic or visual results but do not implement a details-on-demand and relationship between multiple entities functionality.

Table 3.2: Comparison of visual analysis tools: *based on the visualisation tasks (interaction techniques)*

VA Tools	Overview	Details on Demand	Searching	Filtering	Highlighting	Relationship b/w Entities
Thread-based						
(1) Re-mail	*Yes*	*No*	*No*	*No*	*Yes*	*No*
(2) EmailMap	*No*	*No*	*No*	*No*	*Yes*	*Yes*
(3) Beyond Threads	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*
(4) Thread Arcs	*No*	*No*	*No*	*No*	*Yes*	*Yes*
Non Thread-based						
(5) EmailTime	*No*	*No*	*Yes*	*Yes*	*No*	*No*
(6) SeeMail	*Yes*	*No*	*No*	*No*	*No*	*No*
(7) Themail	*No*	*No*	*Yes*	*Yes*	*No*	*No*
(8) Beagle	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*

Based on the visualisation features and taxonomic harmonisation: We proposed the taxonomy of entities based on the visualisations related to digital communication data. Hamonising the taxonomy of entities based on the visualisations related to email communication data aided in understanding the different entities, the association between them and their limitations. Based on the design requirements captured and the literature review, we focussed on all the main four entities mentioned in the taxonomy; such as temporal, individuals and contents, and threads. To understand the association, we grouped three entities/features (temporal, individuals and contents) into threaded and non-threaded based on the work carried out by the researchers in the past.

Almost all surveyed papers provide temporal features to analyse the data. Specific

to email visualisation papers, all the visual analysis tools provide temporal features to analyse the data. From a non-thread based point of view, only Beagle [111] considers all the three features (temporal, individuals and contents) as the authors argue this will provide investigators a better edge in finding the relevant information quickly [111]. However, from a thread-based point of view, EmailMap and ThreadArcs have considered all the three features but the lack interaction doesn't allow the system to have a smooth exploration or aids completely in the discovery process. Though all the papers have considered temporal, individuals and contents in various combination for analysing email communication but there is no smooth interaction between all the three.

The general limitations from the previous works are complex to use, cluttered and not aesthetically pleasing to identify interestingness in the email communication data. The specific ones from our understanding and interpretation are mainly related to exploration, interestingness, pattern-oriented interactive visualisation and discovery. In our work, we focus on building pattern-oriented interactive visualisation that could support analysts in exploring the data, observe the pattern changes in the data, discover relevant/interesting information and specify them with a set of labels.

Table 3.3: Comparison of visual analysis tools: *based on the features*

VA Tools	Temporal	Individuals	Contents
Thread-based			
(1) Re-mail	*Yes*	*No*	*Yes*
(2) EmailMap	*Yes*	*Yes*	*Yes*
(3) Beyond Threads	*Yes*	*Yes*	*No*
(4) Thread Arcs	*Yes*	*Yes*	*Yes*
Non Thread-based			
(5) EmailTime	*Yes*	*Yes*	*No*
(6) SeeMail	*Yes*	*Yes*	*No*
(7) Themail	*Yes*	*No*	*Yes*
(8) Beagle	*Yes*	*Yes*	*Yes*

Limitations and Benefits of Visual Analysis Techniques:

As briefly mentioned in the previous sections, analysts typically perform exploration tasks using their interaction functionalities to find interesting information. From our analysis, we identified relevant limitations. The general ones are complex to use, cluttered and not aesthetically pleasing to identify interestingness in the email communication data. The specific ones from our understanding and interpretation are mainly related to exploration, interestingness, visualisation and discovery. From our review and understanding, we identified two potential benefits of using a visual analysis approach: exploration and visualisation & discovery. All these observations are the basis of the inception of VIS4ML [151], FeatureForge [88], Feature Insight [48], INFUSE [113], VIAL [38], and Beagle [111].

Exploration. An important characteristic of investigative analysis is that it is almost always driven by some exploration to find some leads or interesting information which a could be a particular time, group of individuals, set of events, places, emails etc. For this, we argue that smooth interaction could support analysts in exploring the data, discover relevant/interesting information and specify them with a set of labels. Re-Mail [147], Thread Arcs [108], Beyond Threads [136], Email Map [124], EmailTime [99], Themail [177], SeeMail [65], and Beagle [111] are quite good in exploration but they do not support in classifying emails that are unlabelled and uncategorised. Visual analysis can support this process by making it easier to specify/classify and evaluate. In particular, interactive visual exploration mechanisms can help in making label specification more intuitive by making it easier to change labels based on various iterative explorations.

Pattern-oriented interactive visualisation & Discovery The biggest limitation of email investigation analysis is that these tools do not completely support multi-faceted data visualisation. To detect changes in the data/patterns between all the features is always been a challenge. Re-Mail [147], Thread Arcs [108], Beyond Threads [136], Email Map [124], EmailTime [99], Themail [177], SeeMail [65], and Beagle [111] uses various visualisation techniques but they do not support, especially when emails are considered with threads that are unlabelled and uncategorised. Hence the need of visual analysis. Visual analysis can overcome these issues by providing effective data visualisation methods for all the features (temporal, individuals, threads and contents) and for any other associated meta-data.

In email investigations, analysts need to identify important, interesting, required and/or relevant information, such as individuals, messages, discussions, to support a legal case [63]. Currently, analysts can only use keyword search to find emails of interest, which is limited and not scalable considering the huge volume and the variety of discussions taking place within the emails. One remedy to this is to be able to focus their attention to sub-groups of emails that are "likely" to contain the interesting interactions, e.g., a small group of staff making an executive decision through a long discussion, and "likely" to contain interactions that are of less relevance or interesting, e.g., an employee organising a social event. Also, to find threads of interest or any information related to threads, analysts need to understand the thread features such as pace of interaction, inclusion/exclusion of individuals in the threads, also analyse from a single thread and multi-threads perspective. Hence, the identification and classifying communication types in threads is a key step towards identifying threads of interest and, thereby, advancing in an investigation. We feel using visual analysis techniques can help in building multi-faceted understanding of emails, where can take a human-centric approach in designing novel visualisations and in engineering data features as the basis of our visual analysis environment.

The specific challenges identified through literature and interviews are discussed in detail in the Chapter 4. Our work on visual analysis (design, development & evaluation) will be described in detail in Chapter 5.

3.5 Summary

The study on related works helped in gaining a good knowledge and understanding of the domain, state-of-the-art, visual design principles, interaction techniques used, visualisation tasks, methods, techniques etc. and assisted in comparing findings that helped us in using for designing our visualisations for email communication. Since the DSM model is iterative, the research on related works helped us refine the papers that will be closely relevant to our research and also specific to design studies. The research work also helped us to understand the concepts, ideas and reasons for using different approaches to address

investigation challenges and identify the research gap. As a positive, the work also helped us to understand the visualisation approaches that can be used for discovering interesting information that can be helpful in an investigation. Some of the positives we identified are:

1. Based on the visualisation features and taxonomic harmonisation, we identified four main entities/features in visualising email communication, that is temporal, individuals and contents, including thread features. All the papers considered four main features in various combination for analysing email communication but there is no smooth interaction between all the four. However, hamonising the taxonomy of entities based on the visualisations related to email communication data aided in understanding the different entities, the association between them and their limitations.

2. Based on the visualisation techniques/methods, more than half the surveyed papers use conventional visualisations (basic charts and matrices) to investigate documents/data. Interestingly, many papers also use node-links and other complex visualisations to investigate. There is a trade-off between using complex visualisations and considering all the features that can support in the investigation. However, we argue that using conventional visualisations considering all the features will help in exploration and discovering interesting information that can be relevant to an investigation case. To find nuances in a communication, novel visualisations can be considered.

3. Based on the visualisation tasks/interaction system, almost all the surveyed papers provide an overview of the electronic documents/data. Overview visualisation give users a summary of the document/data collection. Not all summaries serve as navigation support but are often used as starting point for further analysis. Surprisingly, almost half of the surveyed papers only provide an overview and show their analytic or visual results but do not implement a details-on-demand and exploration functionality. Time-based visualisations are often used for exploring task, because they show the temporal development of events in data sources. Contact and content-based visualisations also require searching, filtering, and relationship between the entities

techniques to investigate data.

Chapter 4

Domain Characterisation

In this chapter, we discuss the discovery stage of our design study by employing DSM [156, 61], which is a part of the user-centered design (UCD) and addresses our first objective in the study **O1**. The discovery stage is also called requirements and task collection/analysis (user requirements) in software engineering [112], also called information collection approach, which is directly linked to talking with and observing domain experts.

Since our knowledge is rather limited about what constitutes an effective technique for visualising E-mail data and discovering interestingness, there is a need for understanding how users/analysts interact with data, how they perceive it visually and non-visually, how they investigate when searching for both known and unknown information.

Discovery Method. As suggested in the DSM [156, 61], the general practice in user-centered design is a combination of methods including interviews and observations [40]. However, *just talking* and *contextual inquiries* [103] along with deep literature study will provide interesting and relevant information where we observe users working in their real-world context and question them when clarification is needed, also clarify many points by referring/conducting literature review.

We first identified potential domain for our work, that is E-discovery. As discussed in Chapter 3, E-discovery plays an important role in investigating organisation's email communication. E-discovery requests are mostly conducted by Compliance Officer, Freedom of Information (FoI) Officer, Legal Counsel (E-discovery/legal officer), Human Resource offi-

cer, and/or IT Director/Manager. These officers might have many reasons to commence the E-discovery process in an organisation. In any case, organisation must produce data and/or relevant information in a timely and complete manner when necessary during legal proceedings (includes both pretrial and trial). Some of the cases could be [63][30][59][120]:

1. Freedom of Information (FoI) officer may request an E-discovery investigation by issuing "Freedom of Information request". Nearly all public sector organisations fall under Freedom of Information legislation requiring them have an accurate and speedy mechanism to locate all relevant data and identify key information.

2. Compliance officer many request Compliance Audit and Reporting.

3. An external or internal legal counsel (E-discovery/legal team) might need to see if emails sent/received in an organisation were linked to current legal case or any other legal situation(s) for an organisation.

4. Human Resource (HR) officer may wish to investigate if inappropriate or sensitive emails were circulated within an organisation or any other internal HR issue. Also, to check if critical business information/data has been sent/forwarded to individuals outside of the company.

5. Financial Regulator has the authority to arrive at a financial institution/office and demand access to data immediately, placing the organisation under immediate pressure to commence an E-discovery process. This is called as Financial Raid.

6. IT Director/Manager might also be interested to investigate on the E-mail communication. In many cases, Chief Executive Officer (CEO) or Chief Operating Officer (COO) may request Auditing and Analysis.

In our research, we closely worked with the domain experts working in the E-discovery technology (from Red Sift, a cyber security company).

4.1 Methodology

In this project, regular unstructured interviews were closely conducted with the domain experts/analysts of Red Sift London to understand the challenges, needs and requirements for the investigation of E-mail communication data. The engagement was monthly (once in a month) to discuss the designs, challenges, issues and/or progress. This iterative interview was considered to generate more ways of inferring/learning about domain specifics and help in developing designs and prototypes. In these types of series of interviews, *just talking* and *contextual inquiries* [103],[132] along with deep literature study provide interesting and relevant information where the researcher observes users working in their real-world context and interrupts to ask questions when clarification is needed, also clarifies many points by referring/conducting literature review.

Method: we conducted unstructured interviews, in-person monthly meetings, with the Red Sift experts (in total, three individuals) to understand requirements (general and design), how they work, what their tasks are and what kinds of issues they are facing. Each meeting would last up to one hour (maximum) at their location (Red Sift, London). Due to the nature of the iterative meetings, we ended up having three design & validation (DV) phases. However, some of the requirements and tasks considered in different phases were aligned to each specific phase.

- **DV Phase 1:** 25th January 2016 to 16th December 2016 (12 months)
- **DV Phase 2:** 09th January 2017 to 15th December 2017 (12 months)
- **DV Phase 3:** 10th January 2018 to 14th December 2018 (12 months)

We had interviews and regular meetings with the organisation, three domain experts (CEO, Co-founder & Engineer), to understand the challenges in the E-discovery and to understand domain experts' tasks, requirements and how they expect visualisation to support in the investigation. The experts are cyber security specialists who have a sound understanding of the Digital Forensics and E-discovery Investigation. They build in-house solutions for their own organisation and provide solutions to their clients. In all the sessions,

at a very high-level, we aimed at focusing "To what extent visualisation can support analysts in finding/discovering relevant/interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation?". The iterative interviews helped us to capture the following questions:

DV Phase 1: To what extent visualisation can support analysts in discovering interesting temporal information in the E-mail communication data?

DV Phase 2: To what extent visualisation can support analysts in discovering interesting individuals with their designations (organisation roles) in the E-mail communication data?

DV Phase 3: To what extent visualisation can support analysts in discovering interesting individual behaviour (conversations) in the E-mail communication data?

The analysts moved between these three questions throughout the three years of the design study in a regular interview session. The design study methodology (DSM) [156] iterates and we went through cycles of requirements, design and data sharing, and the data sets and scope developed (with divergence and convergence), which the DSM allows. For example, some of requirements had a mix of time, individuals, threads and contents during the three years. We came up with the three phases based on the data availability, design requirements and task complexity (mentioned in the Appendix A.7). The design phase dealt with an increasingly more complete picture of the data at each of the three phases. And we are doing this is to build increasing capability and confidence working with the data, and to make use of the earlier design phases as a means to learn more about the fuller picture. The initial data identified did not have organisation roles and it is was in a non-threaded form. Later, we identified a table of organisation roles for each of the employee in the Enron E-mail data. We manually parsed the data, merged two E-mail datasets into one to understand how designations / organisation roles can help in our analysis (based on the requirements captured). We also later managed to find a thread-based email data for further analysis (the Red Sift engineers implemented reverse engineering to group all the message IDs to form threads).

This made us come up with the three phases with the design, development and vali-

dation stage. Each phase builds on the other. There is a level of complexity relationship between the three of these (increasing) and we learn from one and move to the other. In the final phase, we deliver a final solution that address all of these in a much more integrated manner. We don't think of temporal information in isolation or threads in isolation. We are presenting a narrative that with the analysis of threads (in phase 3) we will be looking at tasks that require analysts to understand, time, people, content and conversations/communications concurrently.

In all the interviews, specific to capturing requirements, the discussions were informal without any structured approach. The discussions were open-ended and we imposed a little control over experts' responses. We jotted down the notes/points while the discussions were on (hand-written). Since informal discussions occur 'on the fly', it was difficult to tape-record this type of interview. We engaged in the discussion to develop an understanding of the current challenges, tasks and requirements. We were constantly making "*notes*" and the discussions/notes were transcribed immediately [45] (included in the Appendix A.7). The discussions helped to uncover new areas or topics of interest that may have been overlooked by previous research. In each of the phases, the initial discussions were highly informal because our understanding was still evolving, it was helpful as we had a good opportunity to speak with the experts on multiple occasions.

In an effort to better understand, we collected suggestions, comments and feedback from the Red Sift folks, by making notes in a diary book. Then, we elaborated on the points based on the conversation and observation with the experts (based on the hand-written notes) [45]. Finally, reflecting on the discussions, we decided to have further discussions. All the recorded notes were transcribed, encoded and elaborated in the report.

Encoding & Thematic Analysis:

In the interviews, at a high-level, our focus was to understand to what extent visualisations can support analysts in finding/discovering relevant/interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation. As a starting point to discuss about the email communication analysis, some of the open-ended questions (mentioned-above) were presented in all the interviews. The questions

re-occurred at times and sometimes not. The discussions in the series of interviews helped us understand the tasks and requirements in E-discovery that the analysts are interested in the email communication data investigations.

In an effort to better understand in the initial interviews (start of the project), we collected suggestions, comments and feedbacks from the experts (included in the Appendix A.7). The open-ended questions presented were:
Q1: What are the challenges of E-discovery (from digital communication investigations)?
Q2: What do you think about the role of visualisation in E-mail communication data within E-discovery? What are the visualisation requirements analysts are calling for?
Q3: What are the tasks carried out in E-discovery and investigations with respect to E-mail communication data? How will you investigate on the key time-frame, key words and key individuals/players involved?

The data collected through interviews with the experts were analysed on a regular basis (immediately after the interview) and by abstracting data into themes through a process of coding and representing the data. We did this in three steps following Braun et al. approach [44] [45] (1) data familarisation (2) data encoding (3) theme analysis.

First, familiarisation with data was internalised through understanding of the interviews. To do that, we focussed on "*reading*" [45] to take note of points of potential interest and that are relevant to the research question. The notes of the interviews for each of the phases were discussed a number of times for their accurate understanding of the information and meaning. This helped us to communication with our team during the process of coding and theme development. Most of the translated transcriptions were carried out straight after the interview to consider any clarification. This process was carried out on Google Documents (by sharing it with my supervisors).

Second, in the encoding process, we decided to identify the current challenges of the Email Communication within E-discovery. We considered "*coding*" approach [45], which means we were keen on focussing on the complete data that were relevant to the literature and research question. This actually helps in understanding the views of the experts.

The coding is the process of subdividing and labelling raw data captured in the notes, then reintegrating collected codes to form a theory. When we felt satisfied with the codes generated from each of the phases (initial interviews), we aligned the codes with the research questions to examine the Email communication. The data-driven coding helped us to focus on identifying analysis goals and tasks (based on the interviewees' perspectives).

Third, we developed "*themes*"that were sensed through review of the observations, coding and data. The goal of this process was to extract themes and to present a coherent, consistent picture of the tasks and situations under study [96][45]. We read and reread to identify significant broader patterns of requirements and tasks (potential themes). The preliminary analysis came up with a list of challenges, analysis goals, tasks, design requirements, features and case studies. As the interviews were conducted regularly, the thematic analysis used to be carried out regularly as well (after the interviews). The following labels were considered:

- Challenges: $C_1, C_2, ..C_n$
- Design Requirements: $R_1, R_2, ..R_n$
- Analysis Goals: $AG_1, AG_2, ..AG_n$
- Tasks: $T_1, T_2, ..T_n$

A table is created to map the coded artefacts: challenges (C) with the analysis goals (AG) and AG with the tasks (T). The details are provided in the Appendix A.7. The results of the discussion are mentioned below.

4.2 Results

4.2.1 Characterising the Domain

A number of insights were drawn from the initial discussions of the DD Phase 1. To start with, we tried to understand the current E-discovery problem related to analysing E-mail

communication. One of the traditional communication mode in organisations is E-mail system to exchange messages or documents. E-mail data are increasingly called upon for evidences in legal cases, either to protect organisations or even incriminate them. If organisations are unable to produce E-mail data or evidence when called upon by the courts or the authorities, they can face huge penalties. The simple current model of organisation compliance with E-discovery is represented in the Figure. 4.1, In E-discovery, Electronically stored information (ESI) is a process where information is created, manipulated, communicated, stored, and best utilised in digital form [162]. ESI includes writings, drawings, graphs, charts, images, presentations, voice mails, audio files, video files, web links, social media, documents and other data compilations stored in an electronic medium. As shown in the Figure 4.1, from the ESI, the next level in the E-discovery process is the "discovery model" which involves analysis and then the information is presented in a digital form, called "digital evidence" which needs to be presented in a court (legal system). The complete process (data gathering to legal actions) of the current organisation E-discovery model for organisation investigation is cumbersome (as discussed in Chapter 3).

Both the experts mentioned the manual string search for E-mail investigation/analysis is strenuous, time-consuming and huge costs are involved. One of the experts (E1) mentioned, "some of the challenges of E-discovery and investigations are: the process is very expensive, time-consuming, complex and tedious, difficult to compare and identify/detect unusual communication behaviour, difficult to explore freely to identify changes and find some interesting behaviours related to a particular case". One of the experts (E2) mentioned "Few legal experts and advanced users use E-discovery tools such as Jigsaw [164], Concordance by LexisNexis [5] and/or IN-SPIRE [6] to analyse electronic documents but for electronic mail data, we use only manual searching and excel for analysis. In many investigation cases, most of the E-discovery experts often do not know what they are looking for in their data which can be highly time-consuming to find relevance/interesting information and present it in legal proceedings". The analysis of the interviews, helped us identify three main challenges, three main analysis goals and ten main tasks, mapped to one another. The details are provided below.

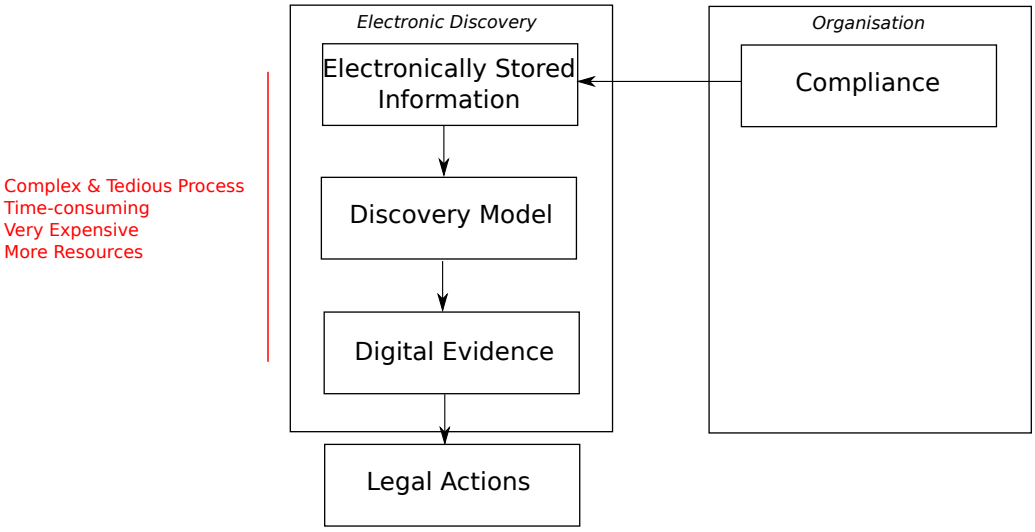


Figure 4.1: Current Model of Organisation Compliance with E-discovery. In this model, the organisation will allow E-discovery team to access the data and analyse for the case, where the whole process in manual.

4.2.2 E-Discovery/E-Disclosure Challenges

In E-discovery investigation and search analysis, "Information Discovery" includes finding, discovering and/or identifying information. So, *finding information* is when there is a definite thing to look in the data, *discovering information* is finding something accidentally, incidentally or exploring in the data and *identifying information* is finding something that is definite and establishing the identity in the data. These analysis with the help of visualisations will support to uncover various information. The challenges from the interviews and literature [161],[63] are expressed in detail below: (C1 to C3, provided in the Appendix A.7).

C1. *Finding interesting subsets within the large volume of data:* A manual sampling [161], searching strings from the data, is used for filtering that enables E-discovery analysts to work on the subsets of data manually and spot similarities and/or differences, where each of the features/attributes are stratified based on the reports/clues. The iterative process of sampling data and finding connections/information in the data consumes too much of time

and are disintegrated to identify an important/relevant change. Also, some of the current techniques/approaches does not aid in supporting various features in finding subsets of multi-faceted data that are of interest. So, representations must be effective for displaying multiple relationships and comparison when placed close together or side by side (in an integrated format), thus improve searching and filtering subsets of data to identify similarities and differences. With the ever increasing amounts and complexity of data, there is a need for simple and effective tools that can help non-tech-savvy lawyers or novice users to carry out investigation and analysis, and later present findings to judges for verdict and/or to share their investigated visuals with their colleagues to enable them to continue further investigation [63]. The effectiveness of finding connections within and between time periods, individuals, context and their connections in the E-mail communication can be improved if we consider an integrated interactive visual representation of all the features to search connections and understand information better. One of the example question is "How to focus on a particular time-frame to determine, visualise and identify interestingness in a subset of time periods considering individuals involved and based on their conversation".

C2. *Complex and dynamic nature of communication patterns:* The E-discovery analysts have difficulty in detecting changes in the E-mail communication due to its complex and dynamic nature [63]. As mentioned earlier, E-discovery analysts use manual sampling to work on the subsets of data and one of the problem is to identify and detect, whether or not, multi-faceted data changes over time. In fact, detecting whether several changes might have occurred, and identifying the times of any such changes is still not effective in the current tools. So, representations must be simple and efficient to identify and detect changes in data over time. Some of the studies on anomaly and change detections are not easy to fit into real-world application due to their cumbersome approach, especially when considered communication data over time. The challenge here is to develop a simple and efficient visualisation to detect, identify, and classify changes and anomalous behaviours in data over time. In a way the solutions will support analysts in taking proactive and preventive measures to win investigation cases [63]. One of the example question is " How to

identify what has really changed in the communication context in an individual or within a group of individuals in a given time-frame?".

C3. *Open-ended data exploration to find interesting communication patterns:* The E-discovery analysts have difficulty in exploring large E-mail datasets and this consumes more time due to navigation issue and finding interesting or relevant information [63]. The exploration of the email corpus must be beyond target search, i.e., supporting visual querying along temporal, connections, context and conceptual dimensions. So, interactive visualisation tool must be easy to explore that will help in navigating smoothly across various dimensions and also aid in suspension (pause and resume while exploring). This will also help analysts in demonstrating various communication patterns to a panel of judges (for proactive and preventive measures) [63]. One of the example question is "How strong are the ties in a particular network (i.e., what is the frequency of contact and how personal are those contacts?) over a period of time? Can you explore and understand the communication patterns of individuals?".

So, to summarise, we need to design and develop interactive visualisation solutions that can support in filtering/searching connections and information, detecting changes and exploration to find interesting subsets of data and interesting communication patterns that can support E-discovery analysts.

4.2.3 E-discovery Design Requirements

As a result of regular exchanges and iterating over design ideas and potential directions that our approach could be build on, we identify the following requirements to guide further design and development:

R1. Investigate high-level temporal characteristics. Having an overview and gaining a comprehensive understanding of the semantics of time periods is required for understanding the structure of the E-mails exchanged over time.

R2. Compare multiple time periods / temporal dimensions. Ability to compare and summarise the communication patterns across all the temporal dimensions is needed to infer

trends.

R3. Understand activities in a temporal selection. Exploratively navigate and investigate temporal patterns and gaps, such as who works outside of normal working hours, weekends, early in the morning or late in the evening).

R4. Investigate high-level individual characteristics. Having an overview and gaining a comprehensive understanding of the semantics of individual's communication behaviour is required for understanding the structure of the E-mails exchanged over time.

R5. Compare multiple individual connections. Ability to compare and summarise the communication patterns across a selected individual's connection is needed to infer commonalities and differences.

R6. Understand overall activities of individuals. Exploratively navigate and investigate the characteristics of each individual with a close inspection of the activities such as "sending emails" and "receiving emails" over a period of time.

R7. Investigate high-level thread characteristics. Gaining a comprehensive understanding of the semantics of threads is required for understanding the overall structures through multi-faceted overviews of threads.

R8. Compare multiple threads. Ability to compare and summarise the inherent patterns across several threads is needed to infer commonalities and differences within sets of threads for the purposes of specifying the categories.

R9. Understand activities in a thread. Due to their dynamic nature, many "events" take place within threads such as, new individuals being added, removed, series of quick replies, long gaps, to name a few. For a characterisation of the thread, oversight of these events, along with people involved and the dynamics of relations between them needs to be gained.

R10. Specify thread communication types. Exploratively investigate the high-level characteristics of the threads along with a close inspection of the activities that take place and externalise the common communication characteristics and types.

4.2.4 E-discovery Analysis Goals

Since the main aim of the research is to develop visualisation solutions to unravel the information in E-mail communication data and support investigative questions/tasks within E-discovery and investigation, we summarise our observations from the interviews within three high-level expectations in investigation cases that involve email communication data comprising temporal, individuals and conversations: exploration, comparison, anomaly detection and multi-faceted analysis. Considering the E-discovery requirements, we have come up with three analysis goals to motivate the need for the research work. The Analysis Goals (AG1 to AG3) are framed based on the challenges, requirements captured and the points identified in the Appendix A.7 by the encoding techniques, that is **R1-R3** to **AG1**; **R4-R6** to **AG2**; **R7-R10** to **AG3**.

AG1: Discovering and characterising time period(s) of interest (Pattern Discovery)

- **AG1a:** Exploring and understanding what makes a time period interesting.
- **AG1b:** Determining, visualising and identifying interestingness in a subset of time periods.

AG2: Discovering and characterising individual(s) of interest (Pattern Discovery)

- **AG2a.** Exploring and understanding what makes an individual interesting.
- **AG2b.** Determining, visualising and identifying interestingness in a subset/cohort of individuals.

AG3: Discovering and characterising thread(s)/conversation(s) of interest (Pattern Discovery)

- **AG3a.** Exploring and understanding what makes a conversation interesting.
- **AG3b.** Determining, visualising and identifying interestingness in a subset of threads.
- **AG3c.** Characterising and externalising patterns where meaningful categories are generated to serve as the basis for classification.

4.2.5 E-discovery Tasks

Nielsen's [9] article on "Goal Composition: Extending Task Analysis to Predict Things People May Want to Do" discuss the importance of extending a task analysis based on the principle of a goal composition. User goals (also called analytical goals or goal analysis) are a final state which user might strive to achieve at a high-level. To achieve the goals, users have to perform some tasks at a low-level (also called user tasks, analytical tasks or task analysis). In light of this, the discussions in the interviews helped us understand the tasks, based on the goals, in E-discovery that the analysts are interested in the email communication data investigations. The research questions along with the design requirements and the analysis goals, helped us abstract out some generalisable tasks. Basically, these tasks are the extended part of the analysis goals which helps in designing visual solutions. The tasks (T1 to T10), provided in the Appendix A.7 by the encoding techniques, are mapped to the analysis goals, that is **AG1** to **T1-T3**; **AG2** to **T4-T6**; **AG3** to **T7-T10**.

T1. Explore E-mail communication patterns (activities) of time periods of interest that differ between the following and find if it is interesting from different perspectives: based on years, months, days, hours, weekends and weekdays, mornings and nights.

T2. Identify interestingness in the temporal gaps using a subset of time periods. Therefore, if the analysts/users know some events they can easily relate those to them (e.g. weekends, holidays, trips, etc.).

T3. Understand and investigate the changes in the volume of emails over time and also assess whether the changes are indeed unusual.

T4. Explore E-mail communication patterns (activities) of individual(s) of interest with others over time and find if it is interesting from different perspectives: based on senders, receivers, senders and receivers in combination.

T5. Identify interestingness in the communication (contact/relationship) using a subset/cohort of individuals. Therefore, if the analysts/users know some events they can easily relate those to them (e.g. departmental meetings, sales, marketing, etc.).

T6. Understand and investigate the changes in the communication of individuals and also

assess whether the changes are indeed unusual.

T7. Explore E-mail communication patterns (activities) of threads/conversations of interest and find if it is interesting from different perspectives: based on time, based on individuals (senders/receivers, inclusion/exclusion, active/passive), based on thread types.

T8. Identify interestingness in the conversations using a subset of threads. Therefore, if the analysts/users know some events they can easily relate those to them (e.g. announcements, advertisements, etc.).

T9. Understand and investigate the changes in the conversation/thread characteristics and also assess whether the changes are indeed unusual.

T10. Compare multiple threads to understand individual behaviour.

4.2.6 Investigating Datasets and Case Studies

E-mail Datasets

Data Collection Methods According to Runeson and Höst [150] classification of data collection methods, we collected the data using indirect methods, where we collected raw data from the existing online sources. Due to Red Sift's privacy, we were not able to use their own data.

Although E-mails are very common, there are very limited options when it comes to availability of E-mail corpus. We had to find real email datasets to work within the design process. This is to facilitate the process and provide realistic prompts and design investigations. We coordinated with the Red Sift experts to identify and also confirm their suitability. Following are some available E-mail corpora for studies related to E-mail:

The Enron Corpus (EC) [10], British Columbia Conversation Corpus (BC3) [11], The World Wide Web Consortium Corpus (W3C) [12], The CSIRO Corpus [13], PW Calo Corpus, Enron Sent Corpus [14], Hillary Clinton Email Dataset [15], European Union E-mail Communication Network [16], Attachment Prediction Dataset [17], Person Name Annotations [18], Conversation Threads, Multi-Lingual Conversations, Communication Net-

work [19], Avocado E-mail Dataset [20], Jeb Bush Emails [21][22], Customer Interaction Data of German Emails and Online Requests [23], Spam email datasets [24], and ECUE Spam E-mail Datasets [25].

The technique of choosing what type of data to be collected, basing on the research goals or user requirements is defined as Goal Question Metric (GQM) method [150]. In this method, a set of criteria, goals or questions must be considered to decided on the type of data. In this study, for implementation purpose, certain criteria were considered: an E-mail corpus must have a rich collection of E-mails, must be a real data, publicly available to access, useful for investigation purpose, must contain features such as temporal, connections and conversations/context. In the survey, only two datasets with case studies, Enron and Hillary Clinton dataset, matched the criteria and hence the reason for using the datasets for implementing the framework and addressing the investigation tasks. However, Enron Email data is a thread-based organisational data and more relevant from building a solution required by the organisation. Based on the discussion with the experts, we decided to consider Enron data (mentioned in the Appendix A.7). More details on the Enron data can be found here [26].

E-mail Case Studies

Enron Case Study [110]: Enron Corporation was an American energy, commodities, and services company based in Houston, Texas, USA, which was one of the world's major energy companies, with claimed revenues of nearly $101 billion during 2000. Enron E-mail scam is a well-known case in the investigation field and it is a real-world benchmark as the nature of the organisation email data is private and unstructured. The Enron scandal, revealed in October 2001, where Enron produced fake profit reports and company's accounts which led to bankruptcy on December 2, 2001. 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 value the Enron data as a valuable test case as it is a large corpus that contains real-world distributions, challenges and tasks with respect to various features and noise.

For the purposes of demonstrating our approach, we are motivated by a use case where analysts are analysing a corpus of emails from an organisation, and interested in *identifying relevant threads* and in *classifying them based on their communication characteristics* to support an E-discovery investigation. In such investigations, analysts need to identify important, interesting, required and/or relevant information, such as individuals, messages, discussions, to support a legal case [59]. Currently, analysts can only use keyword search to find emails of interest, which is limited and not scalable considering the huge volume and the variety of discussions taking place within the threads. Hence, the identification and modelling of communication types in threads is a key step to identify threads of interest and advance in an investigation.

Data Characteristics. As a running example, we will analyse the email data emerging from the Enron scandal – a well-known legal case that involved high level executives of the firm in taking illegal practices to hide financial losses. During the case, several E-discovery analysts were involved to identify various important, interesting and pertinent information in the organisational email corpus and the corpus was made available publicly following the case [110]. The raw Enron corpus contained 619,446 messages with 158 users (sender & receivers) over a 4 year period (1999 to 2002). In the cleaned Enron corpus, there are 200,399 messages with 30,091 threads with the same number of users in the corpus with timestamp and the message type (to, cc, bcc) in which the messages were sent to recipients. There are many missing individuals and emails in the original dataset mostly for unknown or sensitive reasons. Enron data is represented in its raw format as shown in Figure. 4.2. Considering various analysis goals and tasks, we will design and test visualisations considering the three main features (temporal, individuals/connections and conversations/context) using the Enron data[1]. The features derived are expressed in Chapter 5.

As a preliminary method, the Enron E-mail archive [110] is the most popular E-mail corpus and widely used by researchers to understand the social relationships, message relationships, sentiment analysis, language analysis and crisis detection. This data set is

[1]http://news.bbc.co.uk/1/hi/business/1780075.stm

Attribute Name	Attribute Type	Description	Example	Derived?
message_id	categorical	The unique id assigned by the mail system to each email	e94f8397aaf63bf	No
subject	categorical	The subject line of the email	e.g. “California Energy Deal”	No
date	ordinal	The date/time the email was sent	e.g. Thu, 20 Dec 2001 06:05:30	No
to	categorical	The address(es) to whom this message was sent	tana.jones@enron.com	No
from	categorical	The address from whom this message was sent	sue.nord@enron.com,	No
cc	categorical	The address(es) to whom this email was “carbon-copied”	rika.imai@enron.com	No
bcc	categorical	The address(es) to whom this email was “blind carbon-copied”	paul.kaufman@enron.com	No
body	categorical	The content of the email	e.g. “FYI I received a call from Ryan...”, “Ok! Accept the changes and let's get”	No

Figure 4.2: Enron data is represented in its raw format.

a real-time and it is available publicly (free to access). The raw Enron corpus contains 619,446 messages sent by 158 employees of Enron that contains features such as temporal, connections and context. This E-mail data to be set-up in the Red Sift platform and the techniques/strategies to be implemented in this platform.

4.3 Summary

The interview conducted with the experts are expressed in this chapter, which has helped in understanding the challenges, visualisation requirements, tasks and questions that the experts conduct during E-discovery investigations. Also this research investigated on the E-mail datasets and case studies to work with.

Based on the literature review and based on the preliminary interviews conducted with the experts, we were able to understand important and immediate challenges and their requirements. They are discussed in detail along with the simple proposed model of organisation compliance with E-discovery (represented in the Figure. 4.3), where the complete process (data gathering to legal actions) aims to become simple, time-saving, significantly cheaper and consumes less resources.

Since the DSM model is iterative, the work conducting interviews with the experts helped us in improving and collecting information related to the challenges faced by the analysts, tasks required to address the problems and the design requirements that will help in solving in the investigation cases. Some of the positives we identified are:

1. We identified three main challenges (C1-C3) such as improving the comparison of subsets of data, anomaly detection and open-ended exploration based on the interviews conducted with the experts.

2. As a result of regular exchanges and iterating over design ideas and potential directions that our approach could be build on, we identified ten design requirements (R1-R10) to guide further design and development

3. Considering the E-discovery requirements, we came up with three analysis goals (AG1

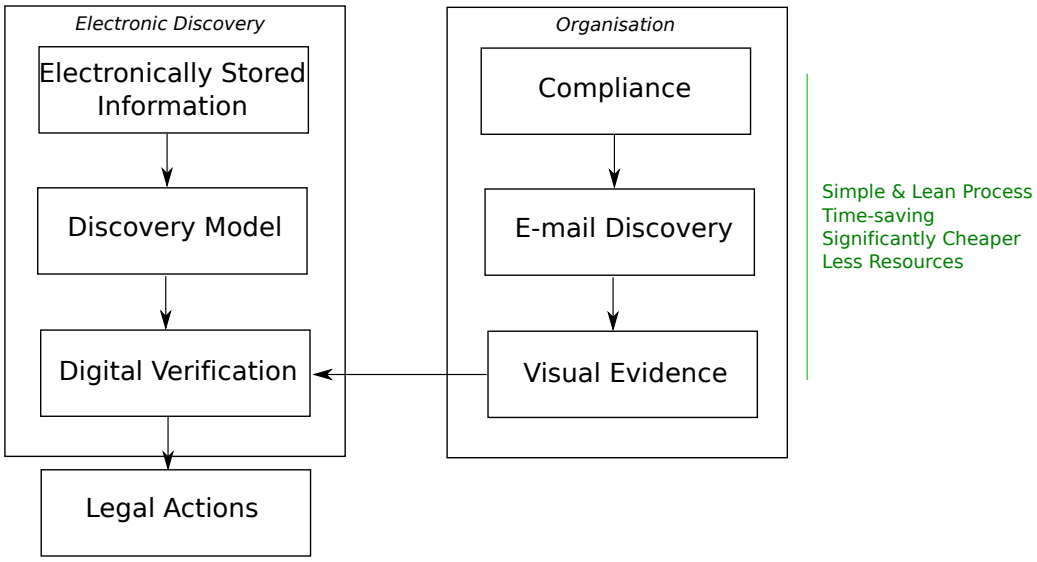


Figure 4.3: Our Proposed Model for Organisation Compliance with E-discovery. In this model, the organisation will have a email management system, which will have a visual analysis pipeline (for Email Discovery) to generate evidence that is needed for the case, where the whole process is automated. However, the focus of this research is to address the question *"To what extent visualisations can support analysts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation?"*.

to AG3) to motivate the need for the research work.

4. The research questions along with the design requirements, helped us abstract out some generalisable tasks (T1 to T10) which helped us build various designs and solutions.

Further details on the Enron data and abstraction are discussed in the next chapter.

Chapter 5

Design Process and Validation

In this chapter, we discuss the design process, which is a core part of the DSM [156, 61]. The goal of this stage is to facilitate data abstraction, visual encoding and interaction mechanisms [156] which address our second objective in the study (**O2**). The abstraction step is to map problems/requirements and data from the specific domain point of view into a more abstract and generic description in the visualisation. After reaching a shared understanding of problem with domain experts in the discovery phase, we started designing a visualisation solution (making visual encoding decisions) with interactions. After each design, we validated the solution with the experts from the organisation, which addresses our third objective in the study (**O3**). This stage is iterative as well.

As discussed in Chapter 2, we report on a three-year iterative collaborative design, where we had a use case walk-through and implementation in proprietary (Red Sift). In our work, the combined stages of design, development and validation has three phases based on the design requirements captured in the initial interviews (shown in Figure. 5.2), where each phase lead to the other phase (discussed in Chapter 4).

- **DV Phase 1:** Visual Exploration of Temporal Information
- **DV Phase 2:** Visual Exploration of Individual Information
- **DV Phase 3:** Visual Exploration of Thread Information

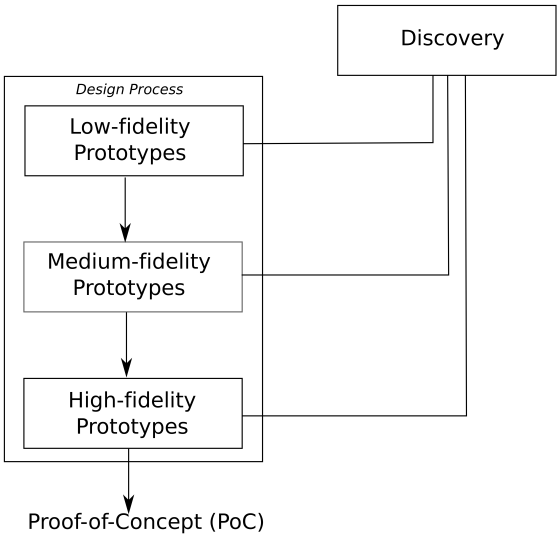


Figure 5.1: Our iterative design process has three main steps in prototyping (low, medium and high fidelity) for visualising patterns in the E-mail data (called as "Pattern Discovery").

In the final phase, we deliver a final solution that address all of these in a much more integrated manner. We don't consider temporal information, individuals or threads in isolation. We present a narrative that with the analysis of threads (in phase 3) we will be looking at tasks that require analysts to understand, time, people, content and conversations/communications concurrently.

All the phases have three levels of fidelity: low, medium and high (as discussed in Chapter 2), which leads to proof-of-concept (PoC) and real-time implementation (as shown in Figure. 5.1). Phase 1, 2 & 3 helps in building a multi-faceted understanding where different characteristics of patterns from various perspectives are discovered through visualisation representations, which we call it as "Pattern Discovery". The visualisation solutions that help in discovering various patterns interactively is called "Pattern-oriented Interactive Visualisation". The details are mentioned in each of the phases.

Every phase had design iterations and we validated the designs. The visualisation literature contains a multitude of proposed methods for validating and evaluating visualisation tools in the wild [76, 196, 96, 115, 53]. The most common form of validation are case

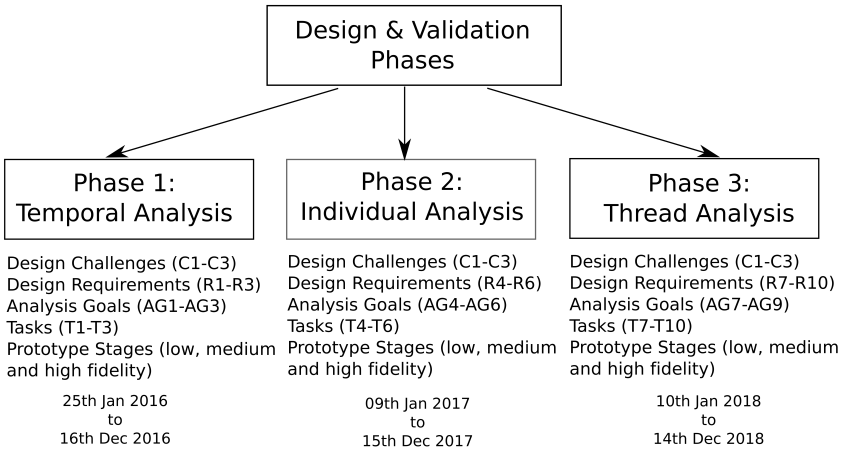


Figure 5.2: Our design & development stage has three main phases (temporal, individuals and threads) for visualising patterns in the E-mail data. Each of the phase is mapped to a particular set of design requirements, analysis goals, and tasks based on the interviews with the experts.

studies with real users, real problems, and real data, as featured in many strong design studies by others [135, 155, 153, 154].

For the final phase of the design, we empirically evaluated the visualisation. This was again conducted with the Red Sift experts by providing the same real-world scenario and tasks to demonstrate how our solutions can support an analyst [164]. The major goal in validating and evaluating a deployed system was to find out whether domain experts are indeed helped by our visualisations. This goal was confirmed by experts doing tasks faster, more correctly, or with less workload, or by experts doing things they were not able to do before.

The next subsequent sections will discuss the design process we adopted in the three phases of our study.

5.1 Design Process & Validation Phase 1: Visual Exploration of Temporal Information

In this section, we aim to address the first question, "To what extent visualisation can support analysts in discovering interesting temporal information in the E-mail communication data?".

Design Consideration:

Specific to visualising temporal information, we have three design requirements captured from the interviews. We are re-introducing this again for the benefit of readers (from Chapter 4).

Design Requirements:

R1. Investigate high-level temporal characteristics. Having an overview and gaining a comprehensive understanding of the semantics of time periods is required for understanding the structure of the E-mails exchanged over time.

R2. Compare multiple time periods / temporal dimensions. Ability to compare and summarise the communication patterns across all the temporal dimensions is needed to infer trends.

R3. Understand activities in a temporal selection. Exploratively navigate and investigate temporal gaps, who works outside of normal working hours, weekends, early in the morning or late in the evening.

Analysis Goals:

AG1: Discovering and characterising time period(s) of interest (Pattern Discovery)

- **AG1a:** Exploring and understanding what makes a time period interesting.
- **AG1b:** Determining, visualising and identifying interestingness in a subset of time periods.

General Tasks:

T1. Explore E-mail communication patterns (activities) of time periods of interest that differ between the following and find if it is interesting from different perspectives: years, months, days, hours, weekends and weekdays, mornings and nights.

T2. Identify interestingness in the temporal gaps using a subset of time periods to find relevance. Therefore, if the analysts/users know some events they can easily relate those to them (e.g. weekends, holidays, trips, etc.).

T3. Understand and investigate the changes in the volume of emails over time and also assess whether the changes are indeed unusual.

5.1.1 Pattern-oriented Interactive Visualisation Designs

Design Approach: The design process started with low fidelity prototypes, that is paper-based sketching, followed by medium fidelity (using Tableau & R) and high fidelity prototypes (using D3.js). The paper-based sketches aimed to design considering the following characteristics (based on the interviews, attached in the Appendix A.7):

- **Time.** When email messages were sent most? How are they distributed? - years, months, days, hours, weekends & weekdays, mornings & nights; Who works outside of normal working hours? weekends? early in the morning or late in the evening? Any unusual changes in the volume of emails? Any other temporal gaps such as holidays, annual leaves, trips etc.?
- **Individual.** Who are the senders and receivers of the selected time?
- **Engagement.** Who sent many messages to whom at a selected time?
- **Context.** What was the communication/message about?

Why grid-based square matrix diagram (heat matrices)?

During the paper sketches phase; based on the interviews with the experts (Appendix A.7), requirements captured (R1-R3), tasks abstracted (T1-T3) and literature review (Chapter

3, Section 3.3.2), we observed grid-based square matrix diagram (heat matrices) can be in either square or rectangular shape representing the strength of relationship between pairs of items of two or more sets. Since the grid will be in the form of horizontal and vertical (rows and columns), the experts suggested grid-based square matrix diagrams can provide a good visual structure; help in organising and realigning the time frames based on the selection. The relationship between time frames can be indicated by varying colours or saturation in each cell where the two items intersect in the matrix. The visualisation technique [90] is well suited to visually query temporal patterns or outliers in a large amount of communicated data by individuals. For example, MatrixExplorer [90] uses matrices to explore relationships of individuals, where rows can represent senders and columns can represent receivers (this can be interchanged as well). Discussing with the experts throughout the paper sketches phase (Appendix A.7) and design observation from the MatrixExplorer [90], we proceeded with the grid-based square matrix diagram using small multiples concept (discussed in Chapter 3) based on the requirements captured and the characteristics (specific tasks) of interest.

After iterations in the designs, the final version of the paper sketch was considered to be transformed into an interactive visualisation, as shown in the Figure. 5.3. We considered a multi-perspective/multi-faceted approach, which refers to modalities in the data (multi-modality), i.e., individuals in the form of a network, temporal changes, and the content exchanged (email text). Also, considered a multi-level granularity approach (characterised by several granularities/levels), i.e., each level will enable an expert/analyst in his/her search towards a representative sample or a reference point (which could be interesting in his/her view). For example, years at a first level will be broken into months in the second level and days in the third level. Let us say, an investigator has selected 28 May 2001 as a reference point for his further analysis.

For visualising temporal features, we considered two main views:

a) main view: this is a primary view that uses a grid-based square matrix (heat matrices) using small multiples concept to visualise relationship between multiple granularities (years, months, days, days of the week), shows the number of occurrences and help identify

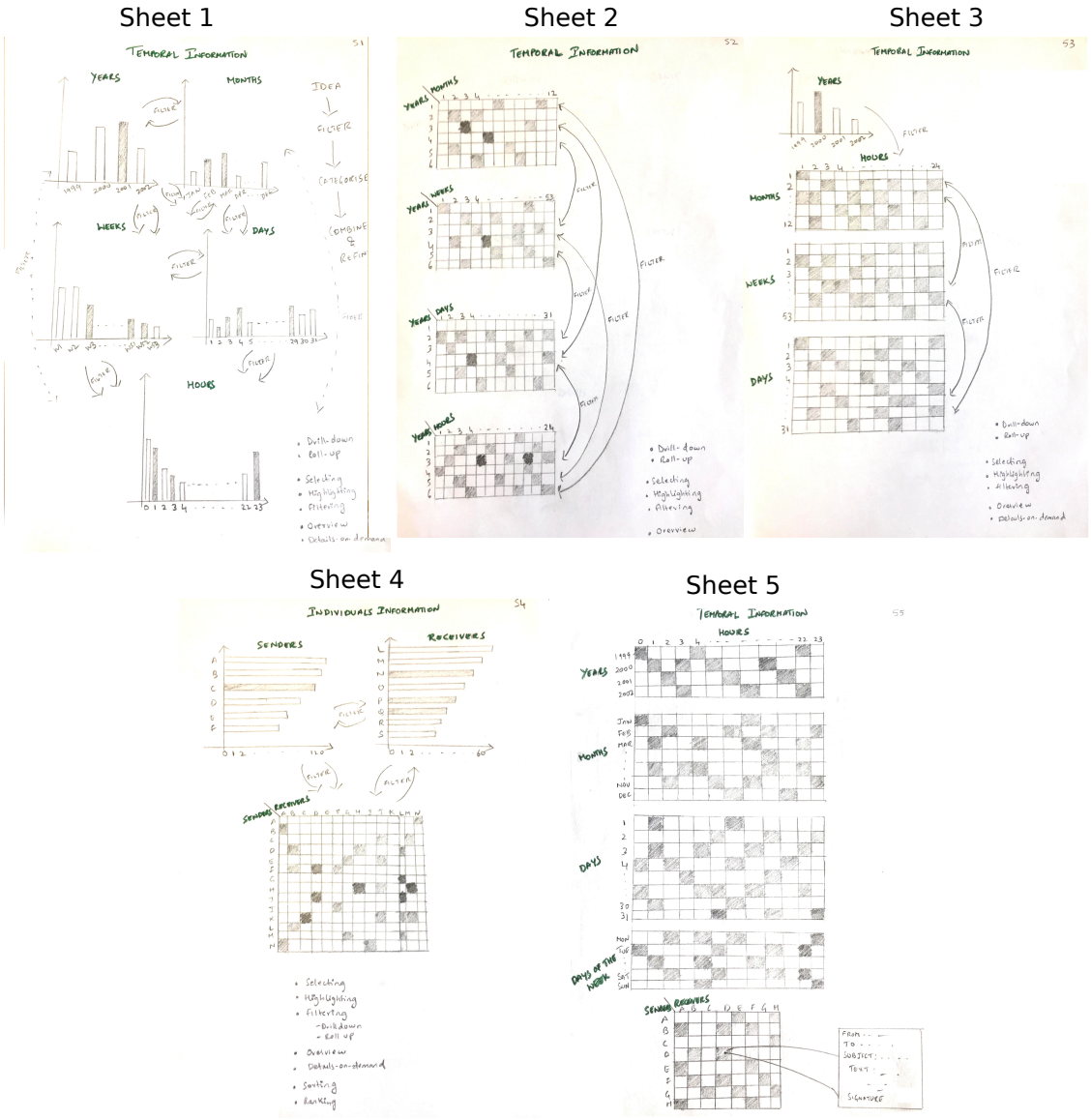


Figure 5.3: Low fidelity prototypes for designing visualisations had 5 design sheets (iterated) for analysing temporal patterns in the E-mail data.

areas for further analysis, such as peak periods of activity (patterns/trends).
b) auxiliary view: this is a support/linked view (along with main views) that uses bar charts to select components of interest (granularities such as year, months, days and days of the week) and find changes in the main views. This helps in filtering and comparing different subsets of data within the views.

For visualising individuals, we considered two views:
a) senders view: this is a view that uses a grid-based square matrix (heat matrices) to visualise individuals that sent emails during the selected time period.
b) receivers view: this is also a view that uses a grid-based square matrix (heat matrices) to visualise individuals that sent emails during the selected time period.

In the medium fidelity prototype, for temporal analysis, we considered only main view (small multiples) such that we have an overview and gain a comprehensive understanding of the semantics of time periods, which will help in understanding the structure of the E-mails exchanged over time (**R1**) (shown in Figure. 5.4). The small multiples are used such that it can help to compare and summarise the communication patterns across all the temporal dimensions that is needed to infer trends (**R2**). In the medium fidelity prototype itself, we introduced interactions such as clicking on the nodes and information being displayed (details-on demand) to understand how interactions could be used in the next version of the prototype.

In the medium fidelity prototype, for analysing individuals, we considered a grid-based matrix chart (heat matrixes) forming a communication network with individuals who sent and received emails (shown in Figure 5.4). In the informal feedback from the experts, they felt the design did not inform much about the temporal slices for each individual, that is when emails were sent or received by individuals. In the medium fidelity prototype, we introduced interactions such as clicking on the nodes and information being displayed (details-on demand).

In the high fidelity prototype, for temporal analysis, we considered auxiliary view (bars)

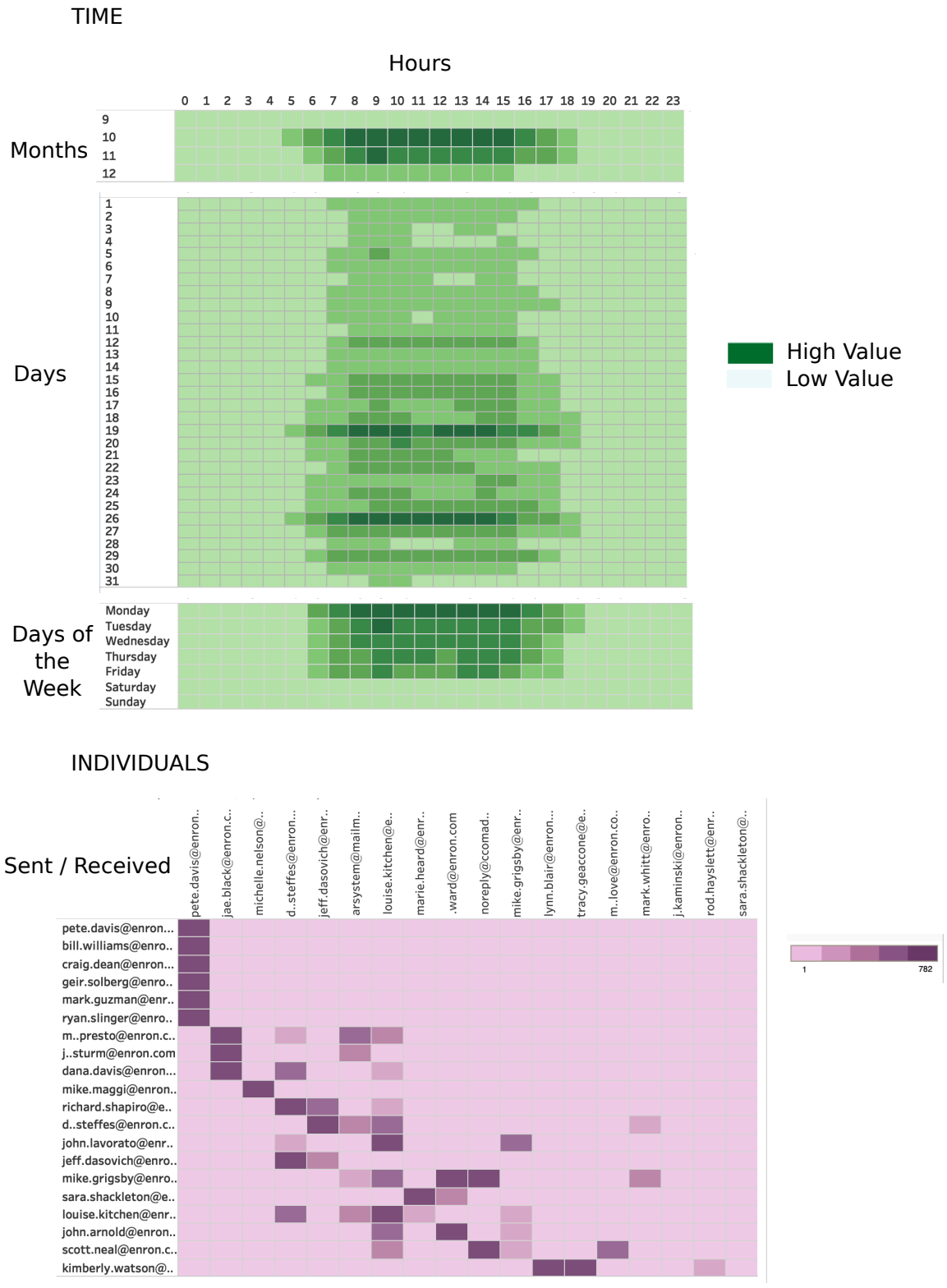


Figure 5.4: Final version of the Tableau design for visualising temporal patterns in the E-mail data.

to support filtering and exploring to navigate and investigate temporal gaps, who works outside of normal working hours, weekends, early in the morning or late in the evening (**R3**), as shown in Figure. 5.5. This was considered after the interview with the experts where they felt adding bar charts will help in filtering the granularities. The main view consists of small multiples with different granularities. Each small multiple represented a granularity such as years, months etc. in vertical direction (y-axis). All the small multiples had a common horizontal line representing hours. The multi-faceted data is color-coded enabling the facets to be distinguished and compared quickly to find various tasks and information. For each of the small multiples, the colour of each square is proportional to the number of emails sent. The time periods are sorted in sequential form.

In the high fidelity prototype, for analysing individuals, we considered two different views for sent and received with time slices. In both the visualisations, each vertical line (y-axis) represents a time-slice and each horizontal line (x-axis) represents an individual. We designed and developed prototype in a way that multiple individuals and time-slices are aligned in parallel, providing a good view of the communication pattern of the selected individual for different time slices, meaning the ability to discover communication patterns that differ from one individual to another. In each of the view, the colour of each square is proportional to the number of emails sent or received for that particular time. The individuals are sorted in increasing order of the messages exchanged. When one of the squares in time visualisation is clicked, messages exchanged opens up as shown in Figure. 5.6.

5.1.2 Validation & Findings

The visualisation is designed for analysts/investigators to explore, understand and find interestingness and extract valuable information out of communication patterns. Along with the experts, we were able to draw some conclusions about finding interesting time periods. For improving the solutions, we considered personal validation [156] merged with inward-facing validation [156] (discussed in Chapter 2). This was conducted by walking through an analysis scenario with a real dataset to demonstrate how our solutions can

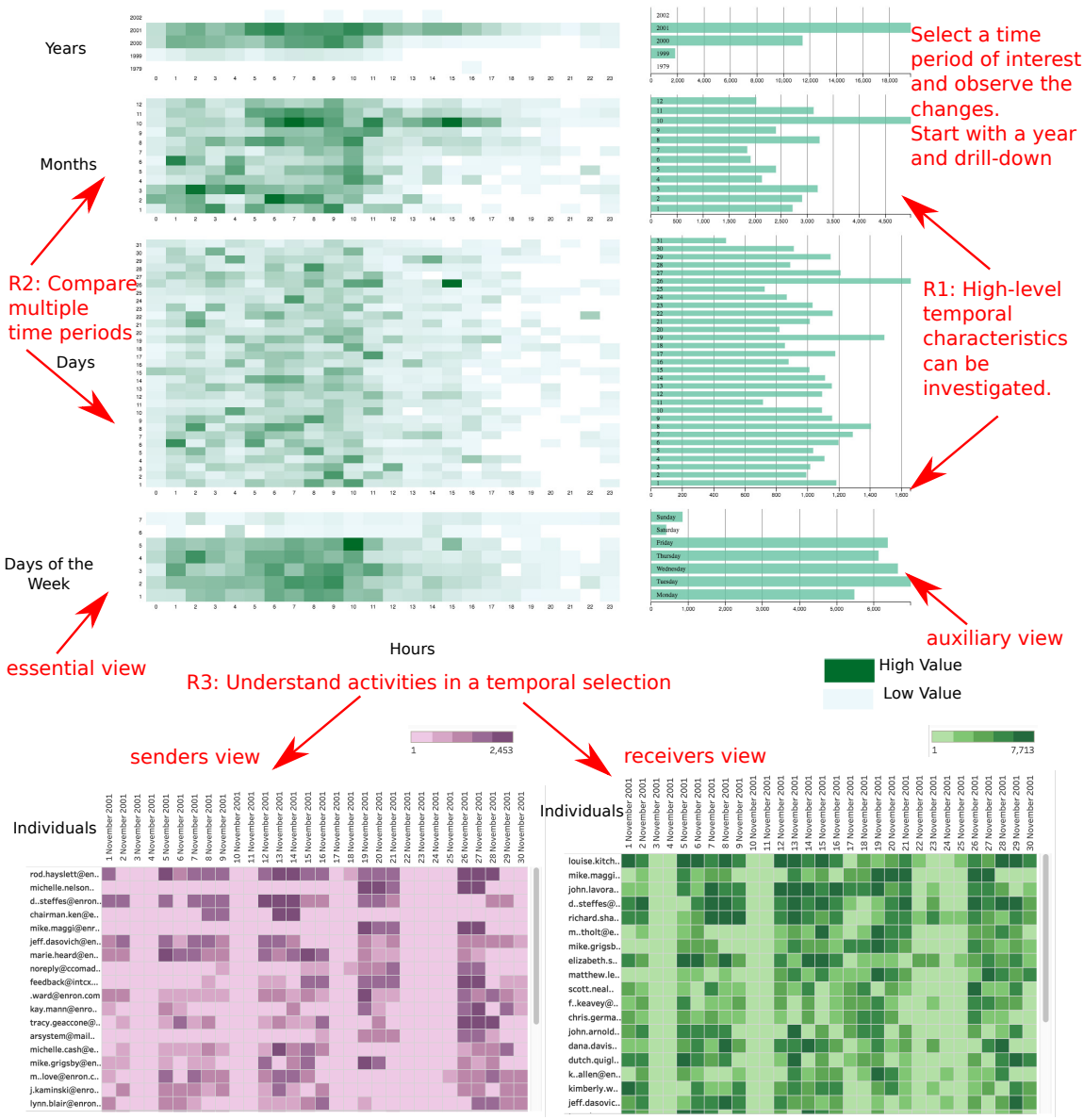


Figure 5.5: Final version of the High-fidelity prototype (D3) design for visualising time in the E-mail data. This is used in exploring the temporal communication patterns through the pattern-oriented interactive visualisation; to address specific domain problem, we consider design requirements (R1-R3).

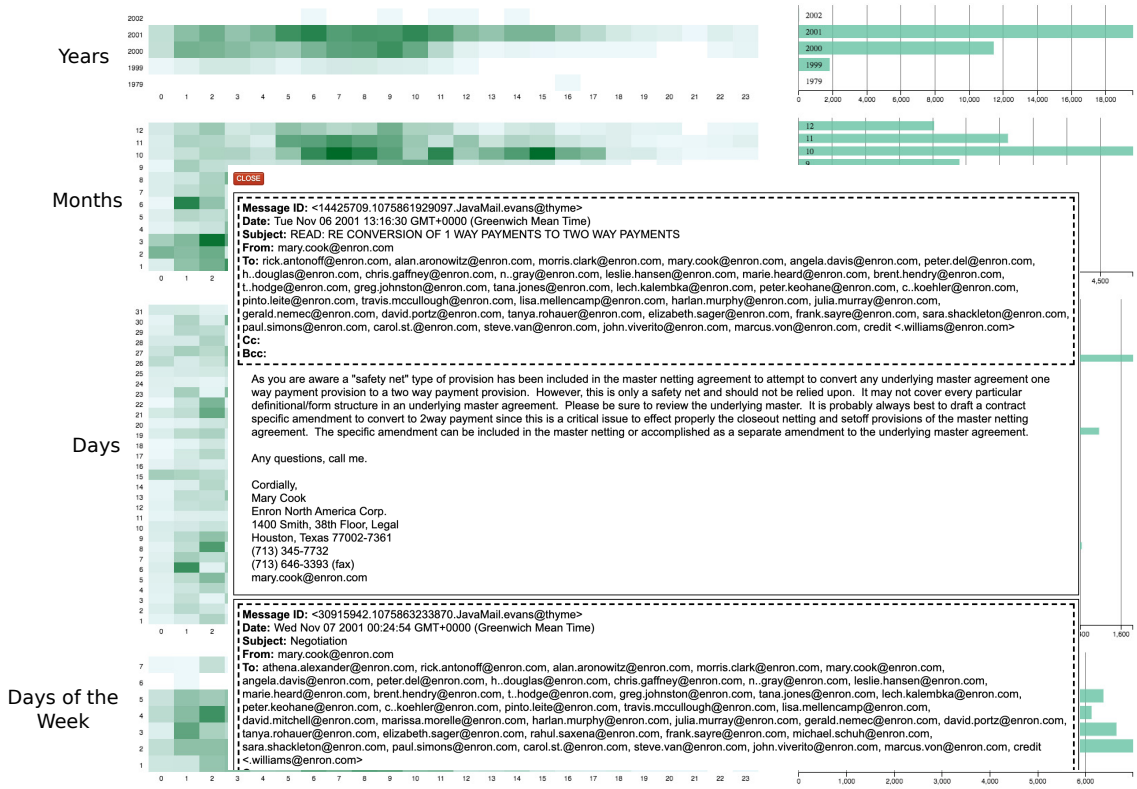


Figure 5.6: In the final version of the D3 Prototype, when one of the squares in the time visualisation is clicked, messages exchanged opens up. This gives analyst a good understanding of the selected points are of interest or not.

support an analyst [164]. The real tasks will help to determine the potential effectiveness of our techniques in actual ongoing email investigations. For validating our visualisation based on the analysis goals (AG1a & AG1b), general tasks (T1-T3), and the email characteristics, we considered a use case.

Use Case Scenario:

In many investigation cases, legal experts or lawyers or analysts do not get any clue about the time-frame or individuals part of the communication case. In that situation, experts are left with no option rather read all the emails manually and go through a strenuous iterative process to identify time and individuals involved. To make life easy for domain experts, we designed email temporal matrix visualisation linked with Individuals connection matrix to make an exploratory analysis such that to minimise the number of emails, to be read manually, by using our visualisation and maximise the interestingness/relevance.

As a starting point of analysis, we considered all the available data as shown in the Figure. 5.7. Based on the volume of emails being sent, we considered four years (1999-2002) of email data for the analysis as shown in the Figure. 5.7. We use temporal matrix charts to visualise emails sent in each of the year. The auxiliary views (bars) help in selecting one of the years. The year 2000 was considered as a reference year to start and later used for comparing with the year 2001 to understand the temporal pattern of email communication during office hours, out of office hours, holidays/trips etc. We use the temporal square matrix along with auxiliary views during the comparison process to understand the patterns and find some interestingness. We started exploring the year 2000 - the number of emails sent were 11,438 and October month was the maximum with 1445 emails. Specifically, on 24th October, 147 emails were sent. By drilling down, we found 77 emails were sent at 7AM. This is something interesting as emails were sent early morning when compared to other months/days (T1).

We continued to explore by investigating the year 2001 as shown in Figure. 5.7. The total number of emails exchanged in the year is 19,611. We found 1175 emails (Saturday-

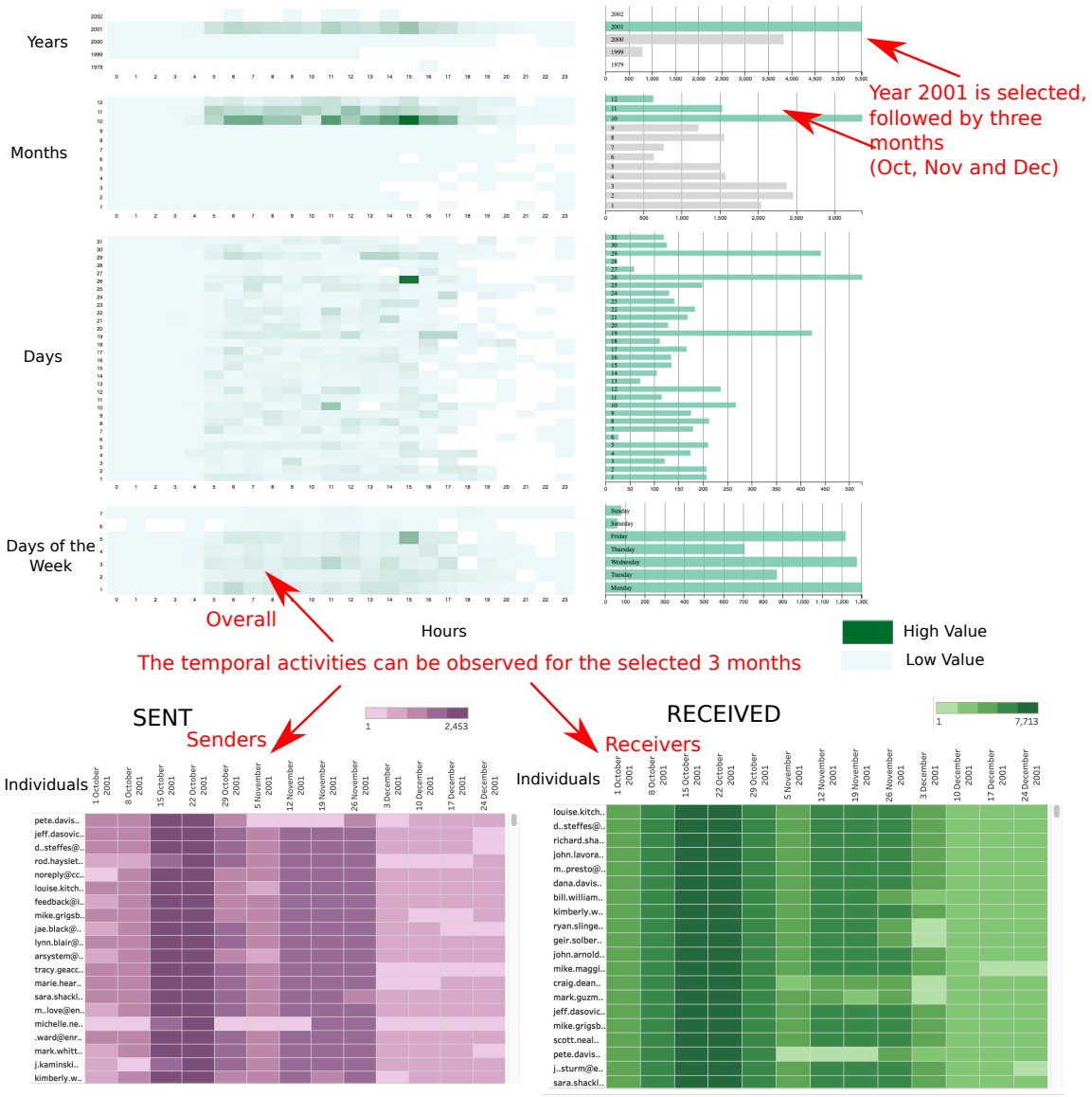


Figure 5.7: Use Case - Temporal activities of the overall communication can be seen by selecting a particular year of interest (in this case, it is 2001). Months, days and days of the weeks can be selected to further understand the points of interest (in this case, months such as Oct, Nov and Dec are selected). The temporal activities of specific individuals (senders and receivers) can be observed in the Sent and Received Views respectively.

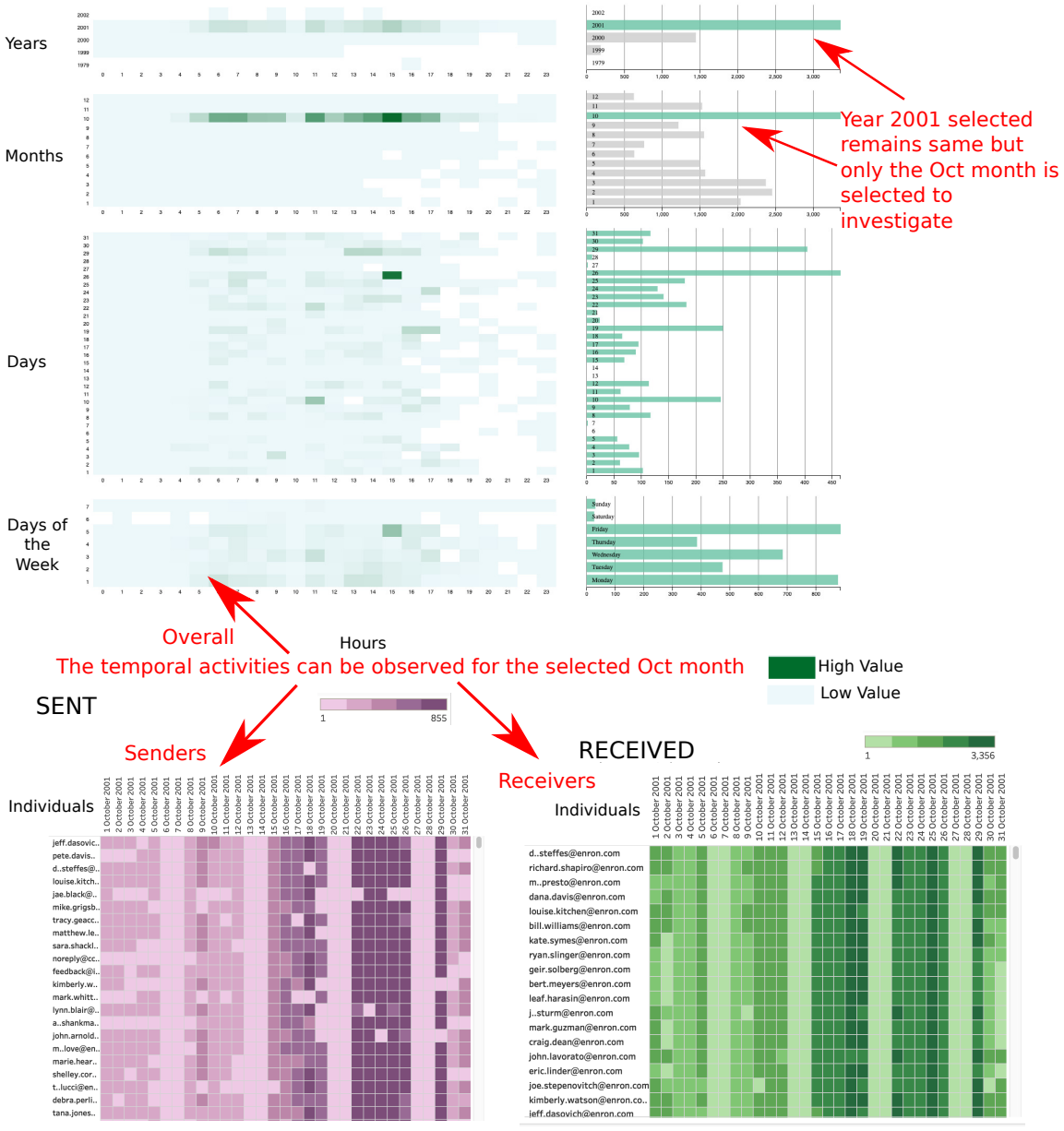


Figure 5.8: Use Case - Now only October 2001 is selected to observe the temporal activities of the overall communication and the temporal activities of specific individuals (senders and receivers).

393 and Sunday-782) were exchanged during the weekends in 2001 whereas no emails during the weekend in 2000 (T2). In our observation, we found this to be interesting. We decided to check for the last 3 months (experts were keen to look into 3 months based on the Enron's case). After observing the patterns, October 2001 was highest among all the months in the year 2001 where individuals exchanged 3362 emails as shown in Figure. 5.8. This is almost twice October 2000, where Mondays and Friday, combined had 800 emails exchanged and weekends 56 emails were exchanged. Interestingly, 26th Friday of Oct 2001 had 466 emails (T3). We further drilled down to find the patterns for an hourly basis as shown in Figure. 5.9. Specifically, at 3 PM, 330 emails were exchanged, and we felt there was something suspicious, which seems to be "abnormal", and motivated us to investigate further. Tana Jones with email ID tana.jones@enron.com had sent those emails to various people in the management, legal team and other employees. The email was more to do with the Enron case, that is credit borrowed and legal issues as shown in Figure. 5.10.

In this exploration, we compare changes not only with years but within the years (months, days, hours, weeks, weekdays and weekends). This helped us to find interesting patterns, interesting time frames and individuals. Our visualisation helps in selecting various granularities of time to filter down based on one's interest. The analysis results using our visualisation is consistent with the literature on the Enron case. The result helped us narrow down the year, months, days, hours and individuals, which also helped us understand changes and compare various subsets of data. With further exploration, we found several other temporal behaviours such as emails sent during office hours, out of office hours, holidays/trips, low volume of emails, intermittent, consistent and high volume of emails. Thus, visualisations have helped us to discover and characterise time period(s) of interest.

Discussion with the Expert. After the personal validation, we discussed our findings with the Red Sift experts, we walked them through the findings and they explored the tool by themselves. The experts were satisfied with our visualisation to analyse temporal behaviour and discover interesting patterns. The expert (E1) mentioned "The tool seems to be useful and this can be connected with a GMAIL account so we can test for our

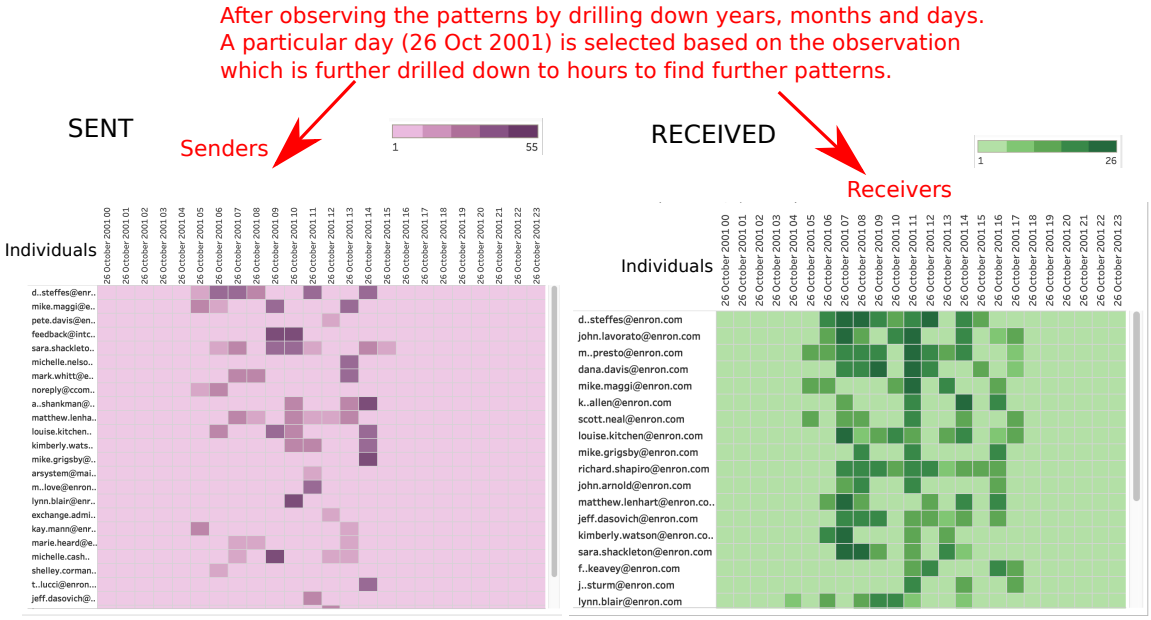
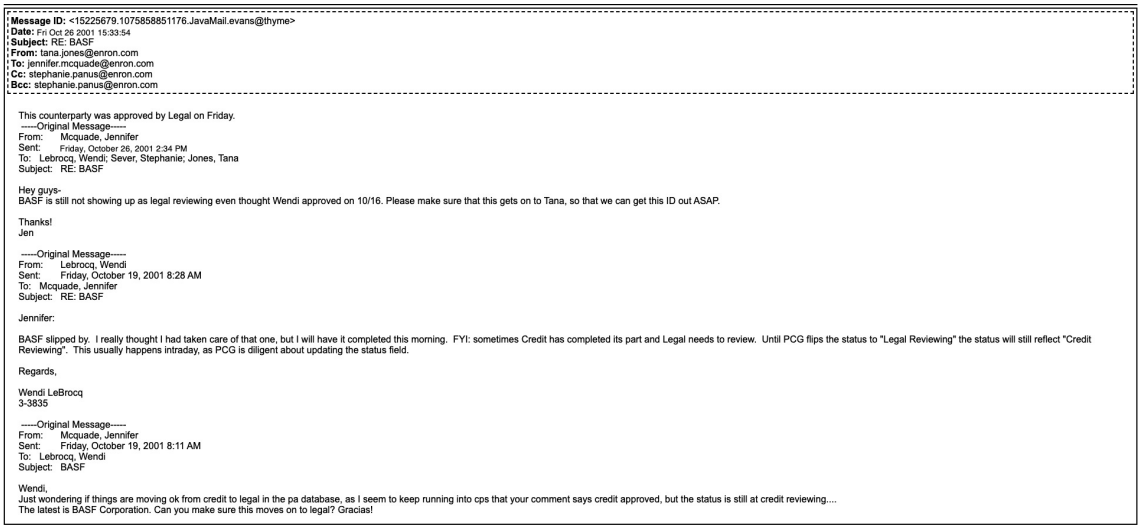


Figure 5.9: Use Case - Now a particular day is selected based on the analyst interest (26 October 2001) is selected to observe the temporal activities of the overall communication and the temporal activities of specific individuals (senders and receivers) - this gives the breakdown of hourly communication of individuals on the selected day.



Message ID: <15225679.1075858851176.JavaMail.evans@thyme>
Date: Fri Oct 26 2001 15:33:54
Subject: RE: BASF
From: tana.jones@enron.com
To: jennifer.mcquade@enron.com
Cc: stephanie.panus@enron.com
Bcc: stephanie.panus@enron.com

This counterparty was approved by Legal on Friday.
-----Original Message-----
From: Mcquade, Jennifer
Sent: Friday, October 26, 2001 2:34 PM
To: Lebrocq, Wendi; Sever, Stephanie; Jones, Tana
Subject: RE: BASF

Hey guys-
BASF is still not showing up as legal reviewing even thought Wendi approved on 10/16. Please make sure that this gets on to Tana, so that we can get this ID out ASAP.

Thanks!
Jen

-----Original Message-----
From: Lebrocq, Wendi
Sent: Friday, October 19, 2001 8:28 AM
To: Mcquade, Jennifer
Subject: RE: BASF

Jennifer:

BASF slipped by. I really thought I had taken care of that one, but I will have it completed this morning. FYI: sometimes Credit has completed its part and Legal needs to review. Until PCG flips the status to "Legal Reviewing" the status will still reflect "Credit Reviewing". This usually happens intraday, as PCG is diligent about updating the status field.

Regards,

Wendi LeBrocq
3-3835

-----Original Message-----
From: Mcquade, Jennifer
Sent: Friday, October 19, 2001 8:11 AM
To: Lebrocq, Wendi
Subject: BASF

Wendi,
Just wondering if things are moving ok from credit to legal in the pa database, as I seem to keep running into cps that your comment says credit approved, but the status is still at credit reviewing....
The latest is BASF Corporation. Can you make sure this moves on to legal? Gracias!

Figure 5.10: Use Case - Based on the time selected and patterns observed, we identified Tana Jones might be of interest and we read the emails exchanged.

business". The learnings are discussed in the next section. The solution was deployed as a proof-of-concept (PoC) in the Google Suite (Google Platform) to analyse their organisation emails to discover interesting temporal patterns related to their business collaborations. Due to commercial sensitivity and confidentiality, we are not able to include the organisation's email visualisation. However, the experts wanted to focus more on the individual's communication behaviour and the details are mentioned in the next phase of the design.

5.1.3 Learnings

L_{dv1}**: Understanding the dynamics of individual's communication should be given relevance.** In the current visualisation, analysts have some basic understanding of how individuals communicate but much more is left to be understood - especially searching and selecting a particular individual of interest and finding/investigating their connections. For example, while working out our task (T2), we found 1175 emails (Saturday-393 and Sunday-782) were exchanged during the weekends in 2001 whereas no emails during the weekend in 2000 (T2). In our observation, we found this to be interesting. We decided to check for the last 3 months (experts were keen to look into 3 months based on the Enron's case) and found individuals who sent and received those emails which were related to the Enron scam [110]. During the design study process, we learnt introducing small multiples will help in representing large, complex, multi-dimensional, multi-granular and multi-faceted data in the form of a compact comparative representation chart. This helped in representing time in multi-granular form (that is, years was broken into months, days, weeks, days of the week and hours to analyse patterns and gaps). During the design process, the three key challenges (C1-C3) considering the design requirements (R1-R3), analysis goal (AG1) and tasks (T1-T3) were addressed and the visualisation was able to reveal things that were interesting. This confirmed our solutions were working and specifically addressed the question *"To what extent visualisation can support analysts in discovering interesting temporal information in the E-mail communication data?"*. When we demonstrated the functionality of the tool that supports pattern discovery (for discovering interestingness in the patterns), the expert (E1) liked the idea of the multi-faceted

exploration and multi-granularity approach, where the high-level view of all the temporal patterns represented using small multiples to visualise relationship between multiple granularities (years, months, days, days of the week) that shows the number of occurrences and help identify areas for further analysis, such as peak periods of activity (patterns/trends) and temporal gaps. Also, the supporting view (bar charts) that helps in filtering and comparing different subsets of data was acknowledged by the expert (mentioned in the Appendix A.7). Though the Enron use cases and tasks fit in with what Red Sift wanted to do, the expert (E1) came up with questions such as "Can you select Richard Shapiro and find his connections?", "Can you randomly search for an individual in the organisation and find to whom most messages were sent and received?" which shifted our focus to more individual-centric analysis. We were not able to demonstrate these questions as we were struggling with too many data items and our visuals were not coping with all the variations in the data. The main lesson from this phase we learnt is to focus more on finding various individuals and understanding their connections in terms of emails being sent or received over a period of time (through exploration). Also, understanding individuals and their connections with/without their designations (organisation roles), comparing and finding who sent/received most/least. This helped us understand the current email dataset (in csv format) must be merged with the organisation roles. The Enron organisation roles were mentioned in the literature and they were manually typed in the csv. Based on our personal validation and expert's view, we understood more research/design process on exploring how individuals communication changes may help design more effective visualisations relevant to investigations. This helped us move into the next phase of building Individuals-based Email Visualisation by extending the set of requirements, analysis goals and tasks.

5.2 Design Process & Validation Phase 2: Visual Exploration of Individuals Information

In this section, we aim to address the question, "To what extent visualisation can support analysts in discovering interesting individuals with their designations (organisation roles)

in the E-mail communication data?".

Design Consideration:

Specific to visualising individual(s) information, we have three design requirements captured from the iterative interviews. We are re-introducing this again for the benefit of readers (from Chapter 4).

Design Requirements:

R4. Investigate high-level individual characteristics. Having an overview and gaining a comprehensive understanding of the semantics of individual's communication behaviour is required for understanding the structure of the E-mails exchanged over time.

R5. Compare multiple individual connections. Ability to compare and summarise the communication patterns across a selected individual's connection is needed to infer commonalities and differences.

R6. Understand overall activities of individuals. Exploratively navigate and investigate the characteristics of each individual with a close inspection of the activities such as "sending emails" and "receiving emails" over a period of time.

Analysis Goals:

AG2: Discovering and characterising individual(s) of interest (Pattern Discovery)

- **AG2a:** Exploring and understanding what makes an individual interesting.
- **AG2b:** Determining, visualising and identifying interestingness in a subset/cohort of individuals and finding relevance.

General Tasks:

T4. Explore E-mail communication patterns (activities) of individual(s) of interest with others over time and find if it is interesting from different perspectives: based on senders, receivers, senders and receivers in combination.

T5. Identify interestingness in the communication (contact/relationship) using a subset/cohort of individuals. Therefore, if the analysts/users know some events they can easily relate those to them (e.g. departmental meetings, sales, marketing, etc.).

T6. Understand and investigate the changes in the communication of individuals and also assess whether the changes are indeed unusual.

5.2.1 Pattern-oriented Interactive Visualisation Designs

Design Approach: The design process again started with low fidelity prototypes, followed by medium fidelity (using Tableau & R) and high fidelity prototypes (using D3.js). We list the following characteristics that are helpful in understanding activities of a particular individual in an email communication (based on the interviews, attached in the Appendix A.7):

- **Individual.** Who are the receivers of the selected individual/sender?
- **Engagement.** Who sent many messages to the selected individual? Who received many messages from the selected individual? Who only sent? Who just received?
- **Time.** When email messages were sent most? When email messages were received most? How are they distributed?
- **Context.** What was the communication/message about?

After iterations in the designs, the final version of the paper sketch was finalised as shown in the Figure. 5.11. We again considered a multi-perspective/multi-faceted approach, which refers to modalities in the data (multi-modality), i.e., individuals in the form of a network, temporal changes, and the content exchanged (email text).

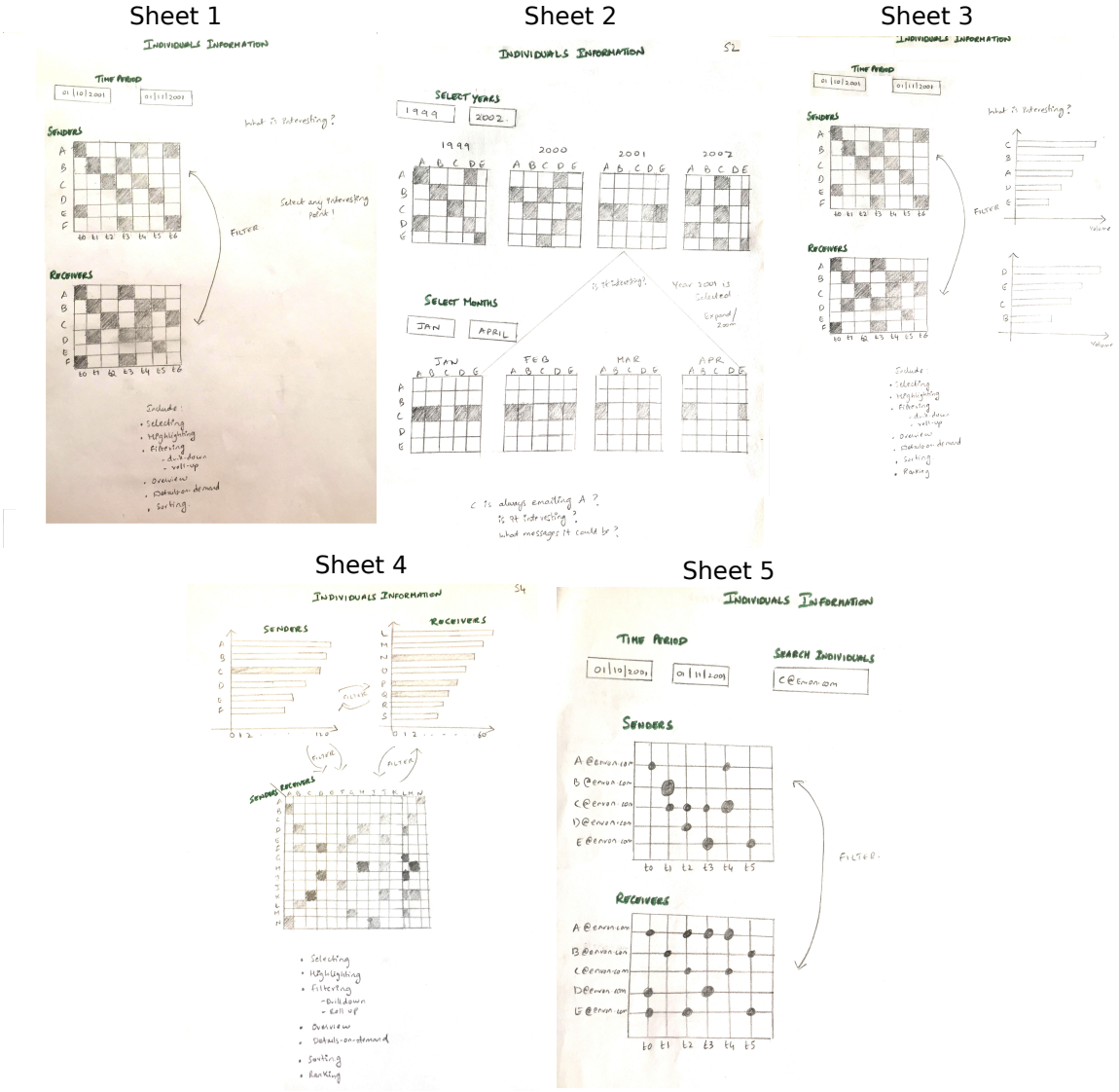


Figure 5.11: Low fidelity prototypes for designing visualisations had 5 design sheets (iterated) for analysing individuals and their connections in the E-mail data.

For visualising individuals' connection, we considered two main views:

a) main view: this is a main view that uses a grid-based bubble matrix to visualise relationship between the selected individual and their connections, shows the number of occurrences and help identify areas for further analysis, such as peak periods of activity (patterns/trends).

b) supporting view: this additional view along with main view displays aggregate statistics such as total number of emails for the selected individual (both sent and received), as well as the breakdown for each of their connection (both sent and received) across the whole time period.

In the medium fidelity prototype, we considered only main view (bubble matrix) such that the experts will be able to investigate characteristics of all individuals at a high-level (R4), compare and summarise the communication patterns across a selected individual's connection to find commonalities and differences (R5), and understand activities of each individual (R6). We designed and developed a prototype in a way that multiple individual connects are aligned in parallel, providing a good view of the communication pattern of the selected individual, meaning the ability to discover communication patterns that differ from one individual to another. In the medium fidelity prototype, we introduced interactions such as clicking on the nodes and information being displayed (details-on demand) to understand how interactions could be used in the next version of the prototype.

Basically, the medium fidelity prototype aims to helps analysts in understanding the communication network of the selected individual for a selected time interval. We designed two visualisations, one for sent and another for received. In both the visualisations, each vertical line represents a time-slice. Multiple time-slice are aligned in parallel which gives the ability to detect changes, and consequently patterns, over time for each visualisation (sent and received separately). Each horizontal line represents an individual to whom he/she has communicated with the selected individual, which are represented using bubbles in different colours (denoting the volume of communication). The colour of the bubble is proportional to the number of emails exchanged between the selected individual and the

Figure 5.12: Final version of the Tableau design for visualising individuals and their connections in the E-mail data.

corresponding individual. The individuals are sorted in increasing order of the messages exchanged.

We developed medium fidelity prototype to visualise a single individual as shown in Figure 5.12. The individual of interest can be searched in the search box. As an example, in the Figure 5.12, we had an informal feedback with the industry experts at their office. We searched pete.davis@enron.com to understand the activities of his communication (R6). We were able to understand his email behaviour such as when he sent and received more emails to which all individuals. We were not able to discover any interesting communication pattern or information based on the Enron case. The experts felt the visualisation had to be improved to find some interestingness.

Some of the suggestions from the expert (E2) during the informal feedback on medium fidelity prototypes are,

- Merging sent and received messages in one visualisation, in the high fidelity prototype, will help in understanding the activities and communication patterns better.
- Considering the departments/groups in an organisation will help in understanding the activities and communication patterns better.
- Building an overview with aggregated statistics will help in refining and finding interesting individuals.
- Considering a supporting view will help in understanding the volume of emails sent and/or received.

Visualising a Group of Individuals (based on the organisation roles)

In the high fidelity prototype, again based on the requirements (R4-R6), analysis goals (AG2a & AG2b), general tasks (T4-T6), and suggestions from the experts, we list the following characteristics (specific tasks) that are helpful in understanding activities of a particular individual that belong a particular department/group of interest. Similar to

visualising a single individual, we again focus on the three email characteristics: time, individual and engagement.

- **Individual.** Which department/group sends more emails in the organisation and who are the receivers of that group selected individual/sender?
- **Engagement.** Who sent many messages from a particular department and from a particular individual from that department? Who received many messages from the selected department and from a selected individual from that department?
- **Time.** For a given time slice, with which other individuals the selected individual is communicating and to whom the most messages were sent/received? For a given time slice, what is the total volume of communication (sent + received) with other individuals? For a given time slice, what is the volume of communication (sent + received) with each other individual separately? How do the selected individual and engagement characteristics (communication / activity / contact / relationship) above change over time? How are they distributed?
- **Context.** What was the communication/message about?

Why grid-based bubble matrix diagram?

During the paper sketches phase; based on the interviews with the experts (Appendix A.7), requirements captured (R4-R6), tasks abstracted (T4-T6) and literature review (Chapter 3, Section 3.3.2), we observed grid-based square matrix diagram does not support in representing both sent and received messages within a particular timeslot/timeframe. The grid-based square matrix worked well for analysing temporal behaviour (design & validation phase 1) but to analyse individuals' email communication behaviour (both sent and received emails) there was a need for other visualisation technique. Following Email-Time [99], we retained the grid-based from the design & validation phase 1 but changed square matrix to bubble matrix, where each bubble will be in circular in shape and several bubbles can be merged within a particular timeslot/timeframe representing the number of messages sent or received, which can be indicated by different colours in each bubble.

The visualisation technique [99] can aid in visually quering patterns or outliers in a large amount of communicated data by individuals. For example, EmailTime [99] uses bubbles to explore messages sent/received by individuals, where rows represent timeline and columns represent individuals. Discussing with the experts throughout the paper sketches phase (Appendix A.7) and design observations from the EmailTime [99], we proceeded with the grid-based bubble matrix diagram based on the requirements captured and the characteristics (specific tasks) of interest.

In the high fidelity prototype, we replicated the medium fidelity basic designs such as grid-based layout, retaining the x-axis (individuals) and y-axis (time slices) but merged sent and received individuals in one visualisation (shown in Figure. 5.14) . Once the individual of interest is searched and selected, the visualisation will represent both sent and received messages using bubbles in different colours and different sizes (denoting the volume of communication). The size of the bubble is proportional to the number of emails exchanged between the selected individual and the corresponding individual. The individuals are sorted in increasing order of size. When one of the nodes is clicked, messages exchanged opens up as shown in Figure. 5.14.

We identified a table of organisation roles for each of the employee in the Enron E-mail data. We manually parsed the data, merged two E-mail datasets into one to understand how designations / organisation roles can help in finding interestingness.

We included two additional views (bar charts) that are used in representing the organisation roles and the email domains of the email data.

a) additional view 1: Bar chart for visualising which domain sent more emails. Each bar represents the number of messages sent. The different domains we considered are corporate/private (in this case enron), personal (aol, Hotmail, yahoo, etc) and others. The individuals are grouped based on the email domains used in communicating.

a) additional view 2: Bar chart for visualising which department/group sent most number of emails. Each bar represents the number of messages sent. The different organisation roles we considered are CEO, directors, managers, managing directors, president, employee and others. The individuals are grouped based on their roles in the organisation.

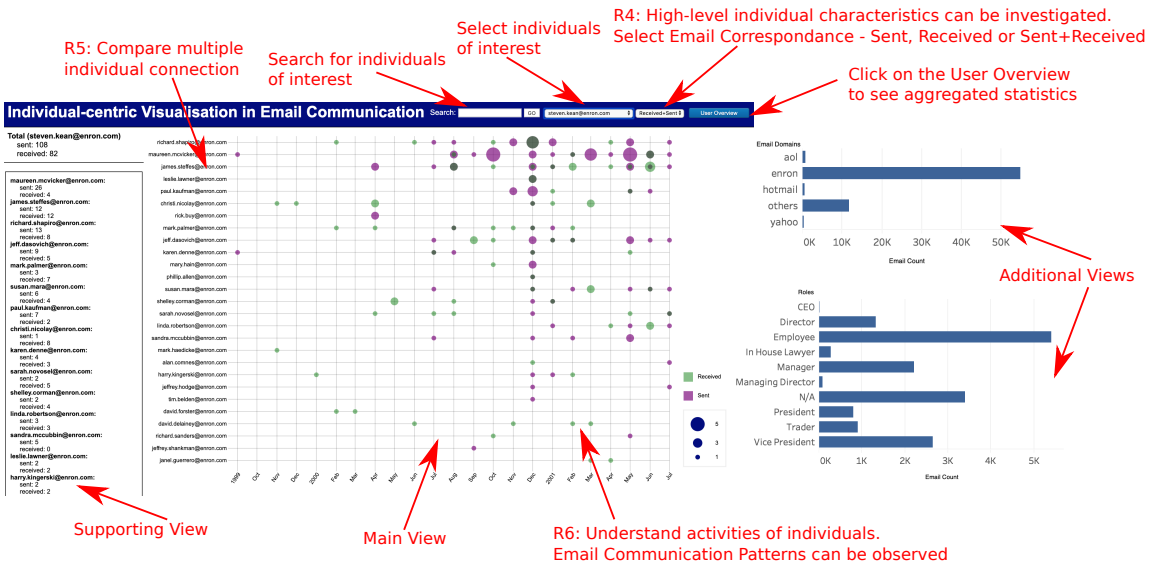


Figure 5.13: Final version of the High-fidelity prototype (D3) design for visualising individuals in the E-mail data. This is used in exploring the communication patterns between individuals through the pattern-oriented interactive visualisation; to address specific domain problem, we consider design requirements (R4-R6)

.

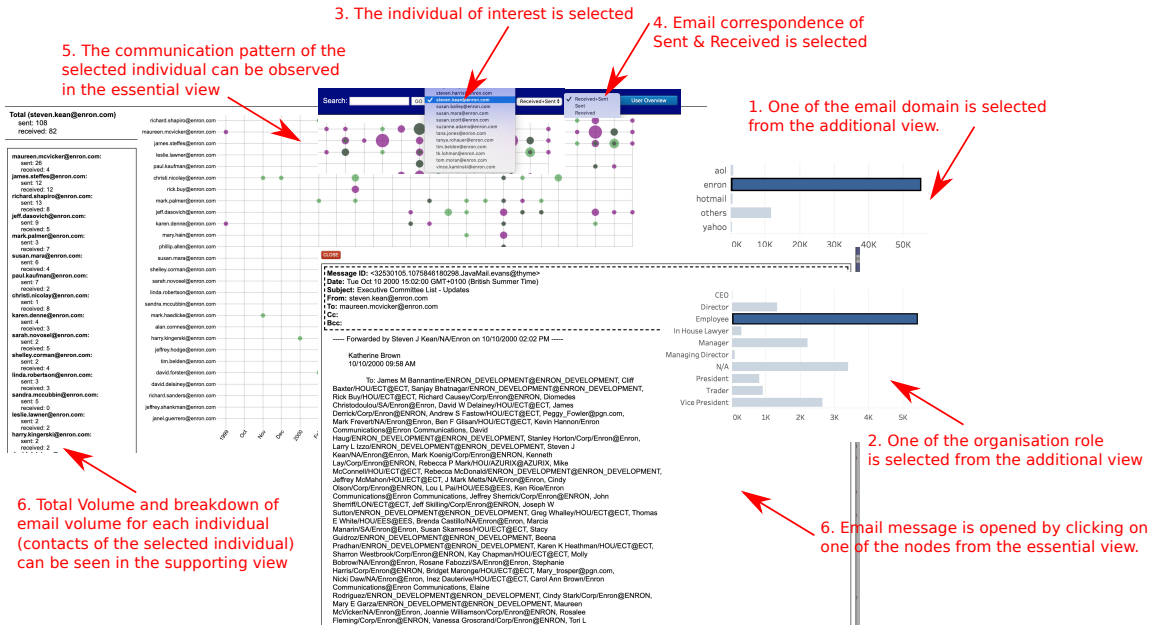


Figure 5.14: Prototype for visualising individuals and their connections in the E-mail data.

5.2.2 Validation & Findings

The research questions along with the E-discovery requirements, helped us abstract out some generalisable tasks, as discussed in the Chapter 3. We again considered personal validation [156] merged with inward-facing validation [156] using real users, real problems, and real data, as featured in many strong design studies by others [135, 155, 153, 154]. The visualisation is designed for analysts/investigators to explore, understand and find interestingness and extract valuable information out of communication patterns. Along with the Red Sift company experts (at their office), we were able to draw some conclusions about finding interesting individuals. The Personal Validation was conducted by walking through an analysis scenario with a real dataset to demonstrate how our solutions can support an analyst [164]. The tasks will help to determine the potential effectiveness of our techniques in actual ongoing email investigations. For validating our visualisation based on the analysis goals (AG2a & AG2b), general tasks (T4-T6), and the characteristics (specific

CLOSE

User	Total Sent	Total Received	Total Communication Volume
pete.davis@enron.com	747	1	748
jeff.dasovich@enron.com	626	109	735
tana.jones@enron.com	260	90	350
james.steffes@enron.com	129	148	277
susan.mara@enron.com	136	120	256
sara.shackleton@enron.com	125	89	214
richard.shapiro@enron.com	28	168	196
steven.kean@enron.com	108	82	190
mark.taylor@enron.com	45	121	166
mary.hain@enron.com	116	32	148
carol.clair@enron.com	88	53	141
geir.solberg@enron.com	1	140	141
ryan.slinger@enron.com	0	140	140
karen.denne@enron.com	57	82	139
mark.guzman@enron.com	1	138	139
paul.kaufman@enron.com	19	116	135
craig.dean@enron.com	0	132	132
mary.cook@enron.com	92	38	130
christi.nicolay@enron.com	105	22	127

Figure 5.15: Prototype for visualising individuals and their connections in the E-mail data.

tasks), we considered 4 use case scenarios.

Case 1: According to the experts, an interesting user based on the medium fidelity prototype and from the user overview (highest number of emails sent) is Pete Davies with email ID pete.davis@enron.com. In our visualisation, we can see that the emails sent by pete.davis@enron.com is very consistent, with almost all emails sent and a high volume as shown in Figure. 5.16. Based on the characteristics identified, we found people who received emails from Pete, the total volume of messages sent to all individuals/connection, also total volume of messages sent to each individual is observed. The distribution of emails over time can be observed. In all, Pete has sent 747 emails and received only one. Through reviewing the actual emails, we notice that pete.davis@enron.com sends several log files and error warnings which makes his sending behaviour uniform. It indicates that Pete Davies may be some kind of system administrator for reporting technical errors. From the inspection, based on the organisation roles, we understand Pete Davies is just an employee and based on the communication pattern (consistent engagement) and content exchanged, this individual is not of interest.

Case 2: After further exploration (T4), we considered Kay Mann with email ID `kay.mann@enron.com`. From our observations, we see that Kay Mann has consistent communication with `suzanne.adams@enron.com` as shown in Figure. 5.17. It indicates some kind of special working and personal relationship between them, as the pattern between them is very distinct in contrast to Kay and other users. Suzanne would also be of interest to an analyst interested in Kay, so we can now delve into the actual emails by picking one of the nodes. Selecting the node from the left, we see two messages, one describing some misunderstanding about the plane, and the other about what seems to be Kay's brother, Michael. Kay depends on Suzanne for issues with Michael's ADD, supporting the similarity the node pattern indicates. From the inspection, we infer that Suzanne is Kay's assistant or secretary, as we can see in a few cases where Kay asks Suzanne to place an event on her schedule, or to plan for the meeting, etc. So, based on the communication pattern and

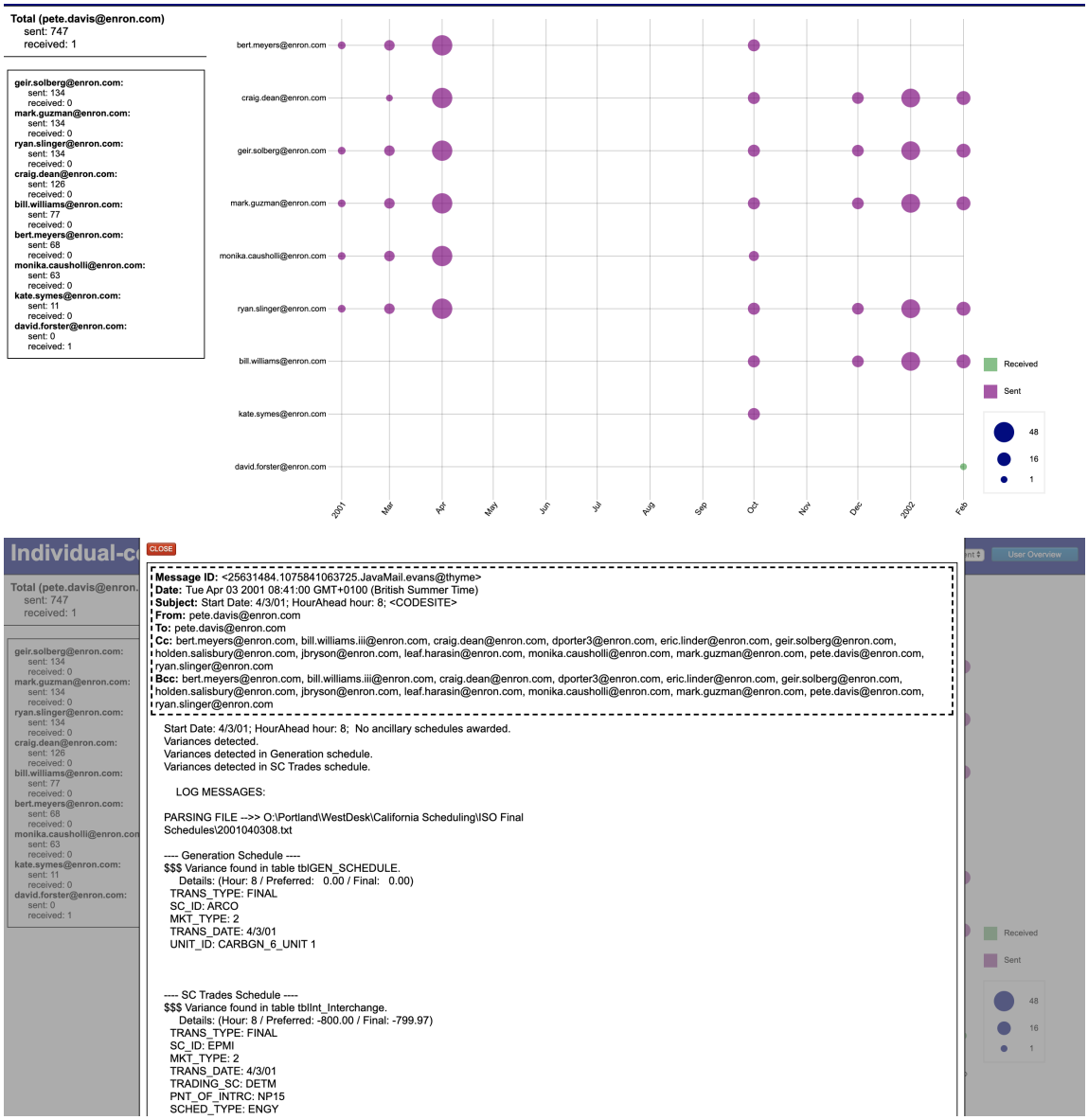


Figure 5.16: Use Case 1 - In the visualisation we developed, based on the individual characteristics, we can see that the emails sent by "pete.davis@enron.com" is very consistent, with almost all emails sent and a high volume. This is something interesting to us.

content exchanged, Kay Mann is of interest (T5).

Case 3: Continuing with further exploration (T4), we considered Richard Shapiro with email ID richard.shapiro@enron.com. From our visualisation, we can see Richard has received several emails from Jeff Dasovich (jeff.dasovich@enron.com) between Sep 2000 and Sep 2001 (a high engagement than the normal) as shown in Figure. 5.18. A close look at the emails, help us understand a lot of the emails were related to Government affairs, getting favourable laws passed by connecting with politicians, and Enron database tracking. A majority of the emails sent by Jeff were related to the California power and energy. Detailed reading of some of the emails revealed that Enron had a comprehensive approach to lobbying government agencies on the common themes of opening markets to Enron's companies and deregulating energy markets that Enron traded in. So, again based on the communication pattern and content exchanged, Richard, the Vice President of Enron, is of interest to us and this gave a lead to investigate Jeff Dasovich (T6).

Case 4: We directly searched for Jeff Dasovich (jeff.dasovich@enron.com) in the search bar to see if any node is highlighted on the time slice that we are interested in (based on the insights from the previous case). Jeff has sent several emails to a group of individuals between Sep 2000 and Sep 2001 (high engagement than the normal) as shown in Figure. 5.19. The individuals are `james.steffes@enron.com`, `karen.denne@enron.com`, `harry.kingerski@enron.com`, `janel.guerrero@enron.com`, `paul.kaufman@enron.com`, `susan.mara@enron.com`, `alan.comnes@enron.com`, `richard.shapiro@enron.com`, `sandra.mccubbin@enron.com`, `tim.belden@enron.com`, `linda.robertson@enron.com`, `edward.sacks@enron.com`, `richard.sanders@enron.com`, and `sarah.novosel@enron.com`. All the emails were related to the California power and energy deals, which are related to the Government affairs, getting favourable laws passed by connecting with politicians, and Enron database tracking. Interestingly, Tim Belden and Edward Sacks were removed from the emails, which means they did not receive any emails between Sep 2000 and Jan 2001 from Jeff on any topics (T6). We note this unusual change in their communication over

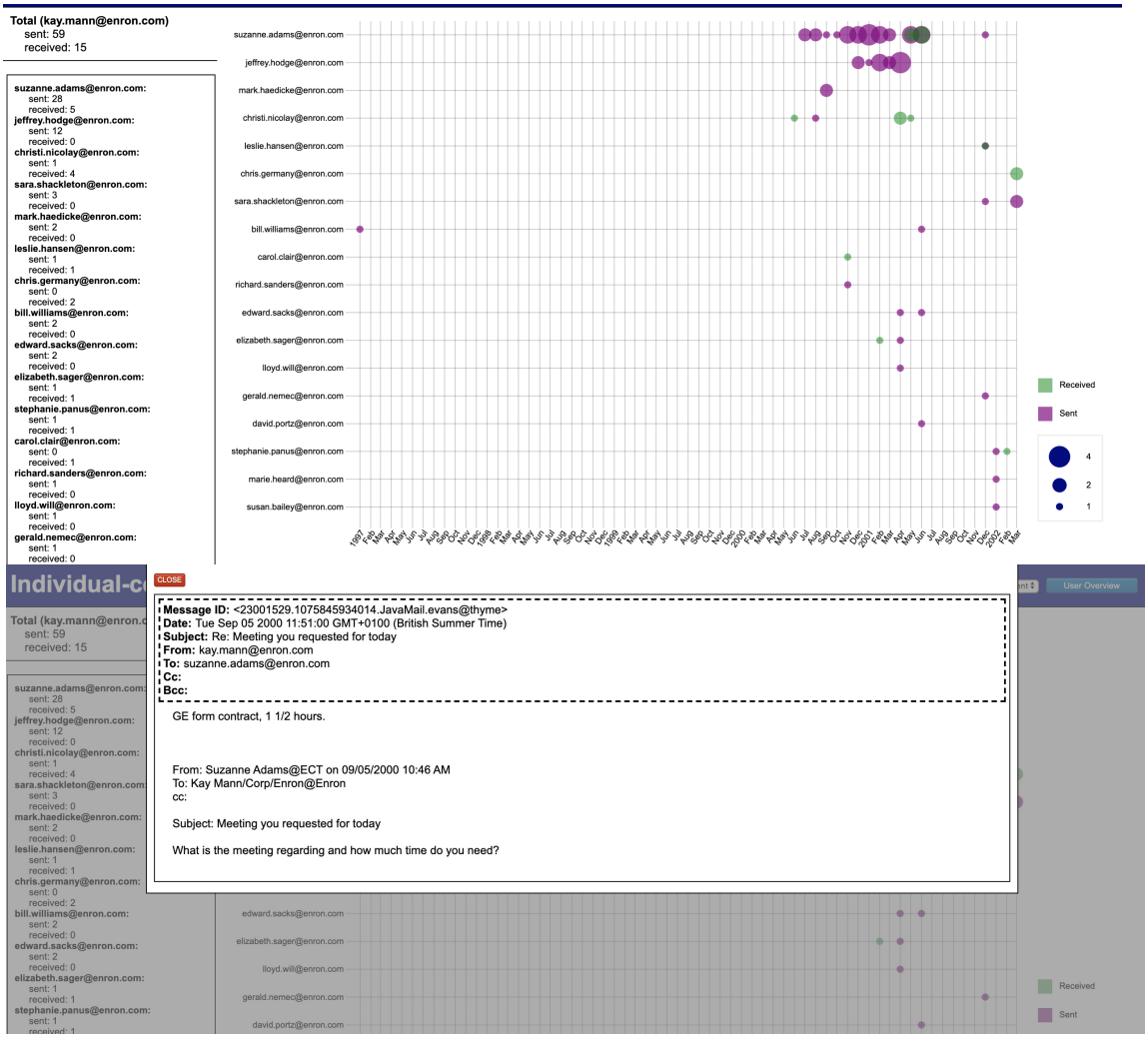


Figure 5.17: Use Case 2 - In the visualisation we developed, we explored communication patterns of individuals from different perspectives such as senders, receivers and/or both (T4). From the exploration, we can see that "kay.mann@enron.com" has consistent communication with "suzanne.adams@enron.com". This is something interesting to us (T5).

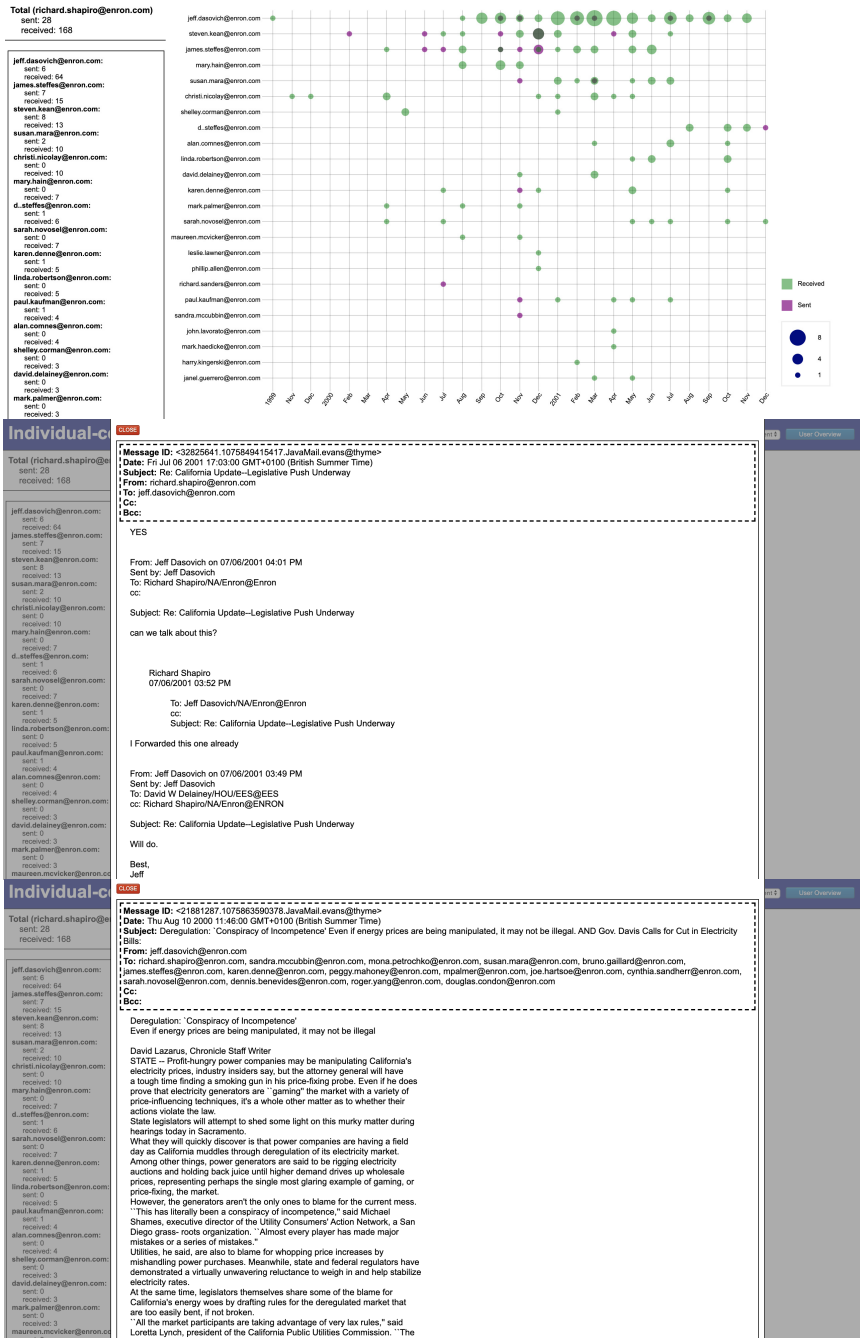


Figure 5.18: Use Case 3 - From our visualisation, after further exploration (T4), we can see "richard.shapiro@enron.com" has received several emails from "jeff.dasovich@enron.com" between Sep 2000 and Sep 2001 (a high engagement than the normal). This is something interesting to us (T6).

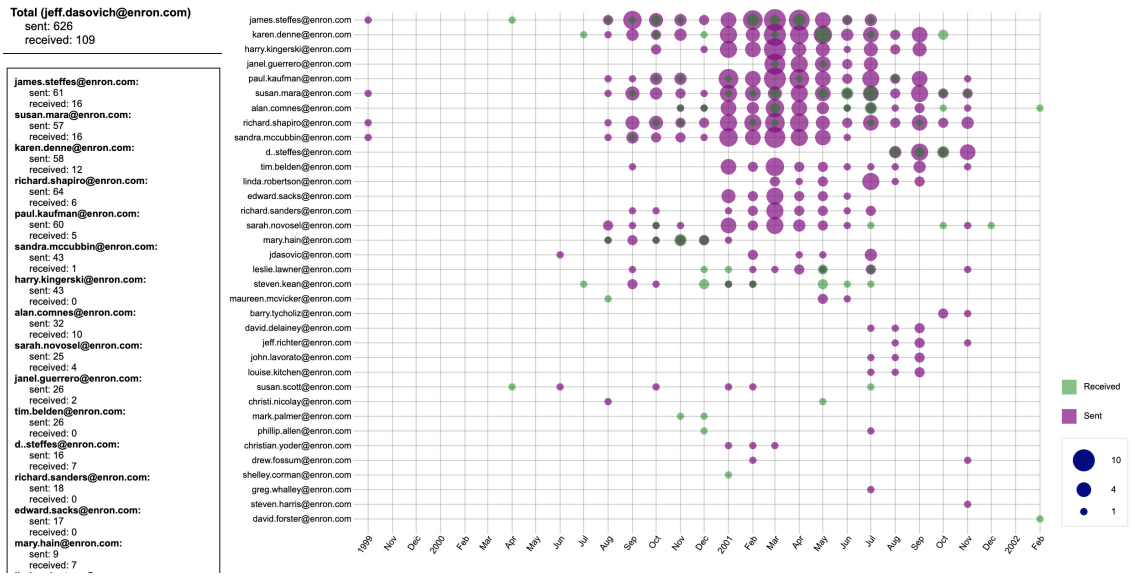


Figure 5.19: Use Case 4 - The individual "jeff.dasovich@enron.com" has sent several emails to a group of individuals between Sep 2000 and Sep 2001 (high engagement than the normal). This is something interesting to us.

time. Based on the communication pattern and content exchanged, we can confirm Jeff Dasovich as a director of Enron has played a key role in the Enron case.

In this exploration, the selected individual is not only compared with the communication changes with his/her group but with temporal changes. This helped us to find interesting patterns, interesting time frames and individuals. Our visualisation helps in selecting various individuals to filter down based on one's interest. The analysis results using our visualisation is consistent with the literature on the Enron case [110]. The result helped us identify some interesting individuals. With further exploration, we found several other connections with low engagement and intermittent communication between individuals. Thus, visualisations have helped us to discover and characterise individual(s) of interest as shown in Figure. 5.16, Figure. 5.17, Figure. 5.18 and Figure. 5.19.

Discussion with the Expert. After the personal validation, we discussed our findings with the Red Sift experts, we walked them through the findings and they explored the tool by themselves. The experts were satisfied with our visualisation to analyse behaviour

of individuals and discover interesting patterns/information. The expert (E1) mentioned "The tool can help in visualising individual connections and this can be tested to see our business connections over a period of time". Further discussions are mentioned in the learnings (next section). The solution was deployed in GMAIL to analyse their organisation emails to discover interesting individual communication patterns related to their business collaborations. Again, due to commercial sensitivity and confidentiality, we are not able to include the organisation's email visualisation. However, the experts wanted to focus more on the thread patterns/structures and the details are mentioned in the next phase of the design.

5.2.3 Learnings

L_{dv2}**: Understanding the dynamics of thread communication structures should be given relevance.** In the current visualisation, analysts have some basic understanding of how to randomly search for an individual to investigate their emailing behaviour and also search for an individual of interest. For example, while working out our task (T4), we discovered Richard had received several emails from Jeff between Sep 2000 and Sep 2001 (a high engagement than the normal), we found this to be interesting. A close look at the emails, helped us understand a lot of the emails were related to the California power and energy, which is also a part of the Enron scam [110]. During the design process, we learnt using bubble-based matrix visualisation will help in visualising relationship between the selected individual and their connections in an aestically pleasing way when compared to a graph-network structure. The bubble-based will also represent the number of occurrences and help identify areas for further analysis, such as peak periods of activity (patterns/trends). To support this, we learnt introducing an additional view to display aggregate statistics will help analysts in better understanding the communication, such as total number of emails for the selected individual (both sent and received), as well as the breakdown for each of their connection (both sent and received) across the whole time period along with their organisation roles. During the design process, the three key challenges (C1-C3) considering the design requirements (R4-R6), analysis goal (AG2) and tasks (T4-

T6) were addressed and the visualisation was able to reveal things that were interesting. This confirmed our solutions were working specifically addressed the question *"To what extent visualisation can support analysts in discovering interesting individuals information in the E-mail communication data?"*. When we demonstrated the functionality of the tool that supports pattern discovery (for discovering interestingness in the patterns), the expert (E1) liked the idea of the multi-faceted exploration and multi-granularity approach, where the high-level view of all the individuals connected for a selected individual is represented using a bubble matrix to visualise relationships between the selected individual and their connections to the low-level view of each individual connected (sent/received emails) and the content exchanged. This can help analysts to identify interesting points and seamlessly switch between the different levels of main view and the supporting views. This also helps in carrying out further analysis to find peak periods of activity (patterns/trends). Also, the supporting view that displays aggregate statistics supports further. The representation of organisation roles adds value to the individual analysis was acknowledged by the expert (mentioned in the Appendix A.7). Though the Enron use cases and tasks fit in with what Red Sift wanted to do, the expert (E1) came up with questions such as "Can you find who are the senders and receivers in a thread?", "Can you find which individuals have been excluded and included back?" which shifted our focus to more thread-centric analysis. Again, we were not able to demonstrate these questions as we did not have email data with threads. The main lesson we learnt is to focus more on finding behaviour of individuals within a thread to understand the connections better. That is, in terms of emails being sent/received, being passive/active in the complete conversation and individuals being included/excluded in a thread over a period of time. This helped us understand the current email dataset must be engineered to get thread ids and the company experts supported us to get the threaded email data. In the current two visualisations (two phases), analysts have some basic understanding of how to compare and summarise the communication patterns across all the temporal dimensions and how to compare and summarise connections of a selected individual (may based on interest or based on a random search). However, much more is left to be understood - especially to visualise communication of individuals within

a thread. Analysts will not only be interested to understand the communication structure for a particular thread but also for a group of threads. It will be interesting to from an investigation point of view to understand why some of the individuals were included or excluded in a particular thread. Analysts must also be able to infer commonalities and differences within sets of threads for the purposes of specifying the communication types (such as announcements, one-way communication, two-way communication, etc.). Generally, investigating these kind of questions can typically give rise to several follow-up questions. Based on our personal validation, we understood more research/design process on exploring how thread pattern (including individuals) changes may help design more effective visualisations relevant to investigations. This helped us move into the next phase of building Thread-based Email Visualisation by extending the set of requirements, analysis goals and tasks.

5.3 Design Process & Validation Phase 3: Visual Exploration of Threads Information

In this section, we aim to address the next follow-up question, "To what extent visualisations can support in discovering interesting individual behaviour (conversations) in the E-mail communication data?". To address this question, we need to consider email threads in the existing dataset.

Design Consideration:

Specific to visualising thread(s) information, we have three design requirements captured from the iterative interviews (mentioned in the Appendix A.7). We are re-introducing this again for the benefit of readers (from Chapter 4).

Design Requirements:

R7. Investigate high-level thread characteristics. Gaining a comprehensive understanding of the semantics of threads is required for understanding the overall structures through multi-faceted overviews of threads.

R8. Compare multiple threads. Ability to compare and summarise the inherent patterns across several threads is needed to infer commonalities and differences within sets of threads for the purposes of specifying the categories.

R9. Understand activities in a thread. Due to their dynamic nature, many "events" take place within threads such as, new individuals being added, removed, series of quick replies, long gaps, to name a few. For a characterisation of the thread, oversight of these events, along with people involved and the dynamics of relations between them needs to be gained.

R10. Specify thread communication types. Exploratively investigate the high-level characteristics of the threads along with a close inspection of the activities that take place and externalise the common communication characteristics and types.

Analysis Goals:

AG3: Discovering and characterising thread(s)/conversation(s) of interest

- **AG3a.** Exploring and understanding what makes a conversation interesting.
- **AG3b.** Determining, visualising and identifying interestingness in a subset of threads and finding relevance.
- **AG3c.** Characterising and externalising patterns where meaningful categories are generated to serve as the basis for classification.

General Tasks:

T7. Explore E-mail communication patterns (activities) of threads/conversations of interest and find if it is interesting from different perspectives: based on time, based on individuals (senders/receivers, inclusion/exclusion, active/passive), based on thread types.

T8. Identify interestingness in the conversations using a subset of threads. Therefore, if the analysts/users know some events they can easily relate those to them (e.g. announcements, advertisements, etc.).

T9. Understand and investigate the changes in the conversation/thread characteristics and also assess whether the changes are indeed unusual.

T10. Compare multiple threads to understand individual behaviour.

5.3.1 Pattern-oriented Interactive Visualisation Designs

Design Approach: The design process again started with prototypes (low to high). We list the following characteristics that are helpful in understanding activities of individuals email communication in a thread (based on the interviews, attached in the Appendix A.7).

- **Individual.** Who are the senders and receivers in a thread? Who are secretly added as BCC? Have they been excluded and included back?
- **Engagement.** Who sent many messages in a thread? Who received many messages in a thread? Who only sent (active)? Who just received (passive)?
- **Time.** When email messages were sent most? How are they distributed? How do the individual and engagement characteristics above change over time?
- **Context.** What was the communication/message about?

Building a multi-faceted understanding where the different characteristics of the patterns from various perspectives are discovered through visualisation representations including thread features (R7, R8, R9). After iterations in the designs (stage 2 to stage

4, which also had medium fidelity prototypes), the final version of the paper sketch was finalised (realisation sheet 5) as shown in the Figure. 5.20. We again considered a multi-perspective/multi-faceted approach, which refers to modalities in the data (multi-modality), i.e., thread structures, individuals (based on sent/received, included/excluded, active/passive), temporal changes, and the content exchanged (email text).

In the medium fidelity prototype, for thread analysis, we considered a series of scatter plots using the small multiples concept, each for a feature, stacked on top of each other. These scatter plots share the same horizontal time axis, allowing to see the temporal trend within and between features more effectively.

Sender analysis - The number of recipients in proportion to the number of discussions in a thread.

Types of communication - We manually labelled different communication types such as announcement, bursty discussion, discussion, information, ping-pong (one to one).

Types of correspondence - The different types of correspondence are TO, CC and BCC.

Types of engagement - The number of active individuals in proportion to the number of all individuals involved in a thread. We termed it as low, medium and high engagement based on the thread length.

We were able to understand different thread structures based on our selection in the features (R7) as shown in Figure. 5.21 (left). The threads are aligned in a parallel form such that it can help to compare and summarise the communication patterns across all the threads that is needed to infer thread types or communication types (R8) as shown in Figure. 5.21 (right). In the medium fidelity prototype itself, we introduced interactions such as clicking on the nodes and information being displayed (details-on demand) to understand how interactions could be used in the next version of the prototype. In the medium fidelity prototype, we introduced interactions such as clicking on the nodes and information being displayed (details-on demand).

In the high-fidelity prototype, for visualising threads (conversations), we considered four main views - single thread, multiple threads, thread features and content view:

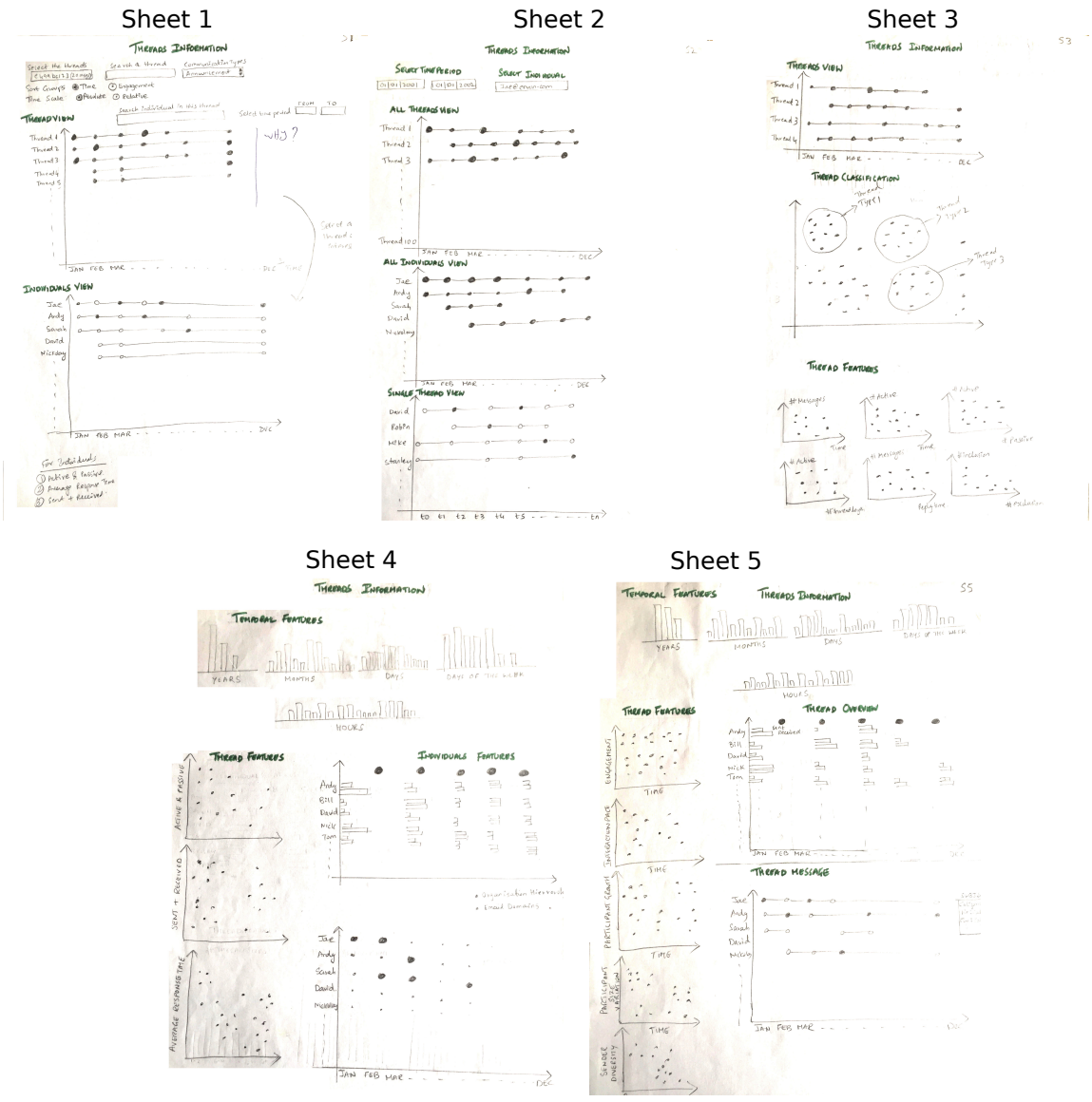


Figure 5.20: Low fidelity prototypes for designing visualisations had 5 design sheets (iterated) for analysing thread patterns in the E-mail data.

Figure 5.21: Final version of the Tableau design for visualising threads in the E-mail data.

Single Thread View

Why grid-based node diagram?

During the paper sketches phase; again based on the interviews with the experts (Appendix A.7), requirements captured (R7-R10), tasks abstracted (T7-T10) and literature review (Chapter 3, Section 3.3.2), we observed grid-based bubble matrix diagram (from the design & validation phase 2) does not support in representing sent and received messages by considering all the three types of communication (TO, CC and BCC) within a particular timeslot/timeframe for a particular thread of interest. The grid-based bubble matrix worked well for analysing individuals' behaviour (design & validation phase 2) but to analyse individuals' email communication behaviour in a thread, there was a need for other visualisation technique. After several rounds of paper designs and following Email-Time [99], we retained the grid-based from the design & validation phase 2 but modified the bubble diagram to adapt to the requirements. We retained the bubble (in this design phase, we call it as "node" as the bubble sizes don't vary) by colour coding to represent if

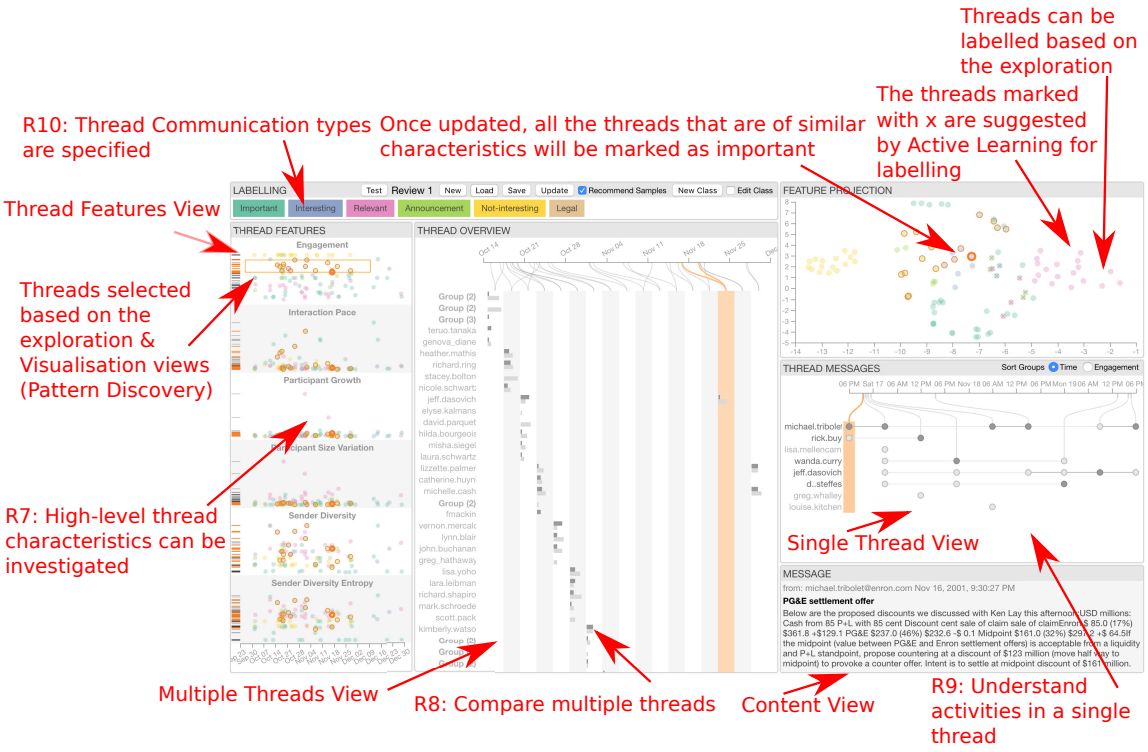


Figure 5.22: Final version of the High-fidelity prototype (D3) design for visualising threads in the E-mail data. This is used in exploring the communication patterns in threads through the pattern-oriented interactive visualisation; to address specific domain problem, we consider design requirements (R7-R10).

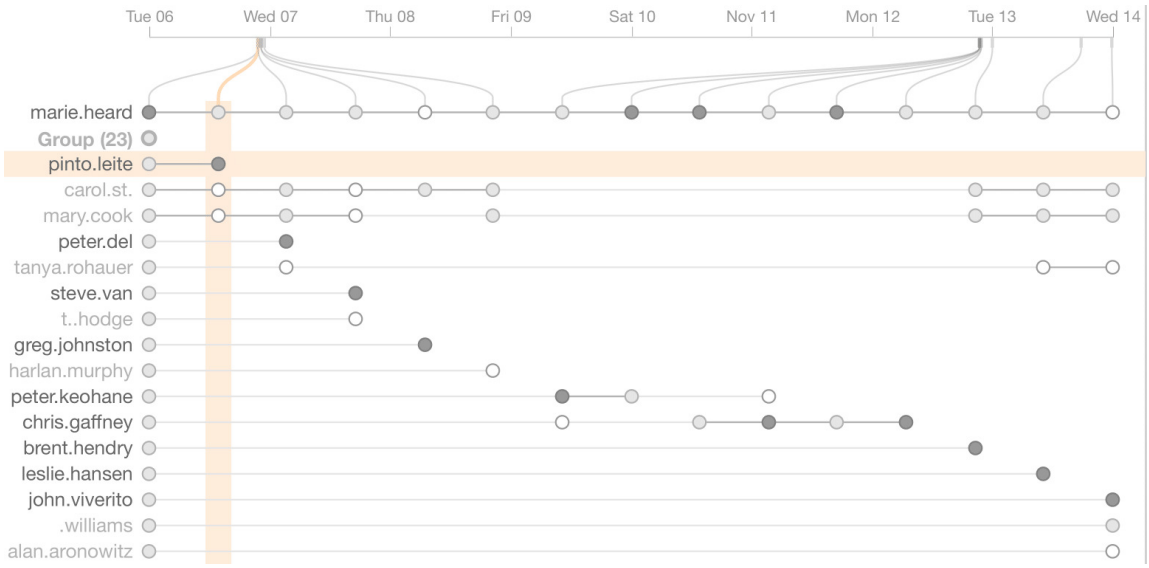


Figure 5.23: A single thread visualisation. Individual and message are highlighted according to mouse position.

the communication was TO, CC or BCC. The visualisation technique [99] is well suited to visually query for patterns or outliers within a thread communicated by individuals. Discussing with the experts throughout the paper sketches phase (Appendix A.7) and design observations from the EmailTime [99], we proceeded with the grid-based node diagram based on the requirements captured and the characteristics (specific tasks) of interest.

A grid cell is marked with a circle glyph if the corresponding individual involved in the corresponding message. To indicate the involvement type, we shape the glyph: dark circle for senders, light circle for TO/CC receivers, and white circle for BCC receivers. We do not distinguish TO and CC because they change constantly whenever email messages are replied. Column messages are sorted by sent time and equally spaced for a clean design. A horizontal time axis is shown at the top and helps messages connect to absolute time. Messages in the first two groups in Figure. 5.23 are sent within a short amount of time.

In each row, all circles are connected to distinguish messages from other individuals and to see the number of messages an individual is involved more effectively. Void space indicates that individuals are not included in the messages. A long line segment with no

circles in-between reveals an individual was excluded and included back in the thread. Email address of an individual is shown on the left with darker colour if he/she has sent at least one email. Rows can be ordered by time (individuals sent earlier messages are at top) and by level of engagement (individuals sent more messages are at top). To have a higher scalability, individuals with identical send/receive patterns are merged together. At the top of Figure. 5.23, a group of 23 individuals is received one (the first message in the thread) and only one message. When a mouse is hovered the visualisation, the corresponding message and individual are highlighted with light yellow background.

Multiple Threads View

Similar to visualising a single thread, we also focus on the three thread characteristics: time, individual and engagement. However, we need to sacrifice the detailed information to accommodate for the number of threads. A grid-based layout is also used, similar to a single thread, with rows representing individuals and columns representing threads (Figure. 5.24). Each thread connects to the timeline with the sent times of the first and the last message in the thread. Each grid cell contains two bars: the darker/lighter bar represents the number of messages that the corresponding individual sent/received in the corresponding thread. Individuals with identical sent/received patterns are merged.

A few interesting observations can be made from Figure. 5.24. Individual "m..tholt" (highlighted with light yellow background by mouse hovering) is involved and is an active sender in the first two threads. The third thread is quite short in terms of both duration and number of messages. The first thread is more like a one-way conversation; whereas, both individuals in the second thread make similar contribution.

Thread Features View

In this view, we considered visualisation-assisted feature engineering, as the email data we wanted to analyse from threads perspective had limited features. It is a numeric way to

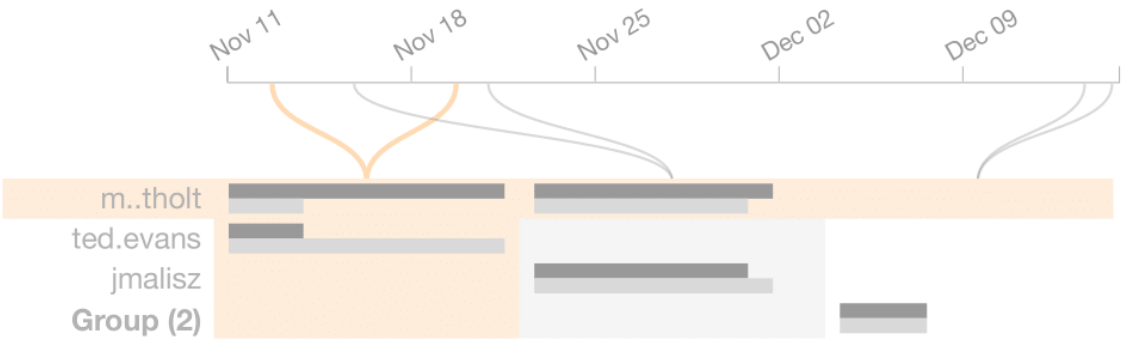


Figure 5.24: A visualisation of multiple threads. Individual and thread are highlighted according to mouse position.

guide the search of interesting threads when we have thousands of them. In this process, we use domain specific knowledge and human insight to derive relevant features from the raw data to support analytical tasks (as discussed in Chapter 3). In the high fidelity prototype, to reveal the characteristics of threads and collections of threads, we considered two categories of thread features (individuals-based and interaction dynamic based), based on the analytical tasks, which are further classified into the following below. The thread features view is shown in Figure. 5.25.

1. Thread-based Individual Related Features: These features investigate the changes in the set of individuals involved across a thread.

Sender Diversity - Proportion of unique senders to the total number of messages sent in a thread

Sender Diversity Entropy - The variation in the number of messages sent by different senders

Participant Growth - Proportion of number of individuals at the end of a message and the start of a message to indicate whether the discussion grow or shrink over time.

Participant Volume Variation - Variation in the size of the set of number of individuals involved in messages along a thread.

2. Thread-based Interaction Dynamics Features: These features try to capture the temporal and behavioural dynamics of the discussions across a thread.

Level of engagement - The number of active individuals in proportion to the number of all individuals involved in a thread.
Pace of interaction - The median value of all the temporal gaps within messages in seconds

Why scatter plot diagrams?
Based on the interviews with the experts (Appendix A.7), requirements captured (R7-R10), tasks abstracted (T7-T10) and literature review (Chapter 3, Section 3.3.2), we learnt scatter plot diagrams will help in representing relationship between two variables in the data. Each position in the scatter plot can be defined according to two dimensions produced in the form of either points, dots or symbols. Since we identified six categories of thread features (mentioned above), based on the analytical tasks, we considered scatter plot diagrams with small multiples approach (discussed in Chapter 3) that can help in comparison analysis, reveal a range of potential patterns in the visualisations and identify a particular thread of interest. For example, in the investigation domain [75], small multiples allow analysts to easily compare the differences in charts that are placed in a matrix form (horizontal and/or vertical, rows/columns) and find points/areas of interest. The visualisation technique [75] is well suited to represent trends and correlations in the communication data. Discussing with the experts throughout the paper sketches phase (Appendix A.7) and design observations from the ScatterPlot approach [125], we proceeded with the scatter plot diagrams to represent multiple thread features in a small multiples form based on the requirements captured and the characteristics (specific tasks) of interest.

We observed in our interviews with the experts (Appendix A.7) that the features mentioned play an important role in examining threads individually and collectively. A series of scatter plots using the small multiples concept, each for a feature is stacked on top of each other as shown in Figure. 5.22 (left). These scatter plots share the same horizontal time axis, allowing to observe the temporal trend within and between features more effectively. The vertical axis is mapped to the feature value. On the left of each scatter plot, a 1D rug plot is shown as an extra encoding of the feature values.

To be able to see the interaction between all features, we construct a feature vector

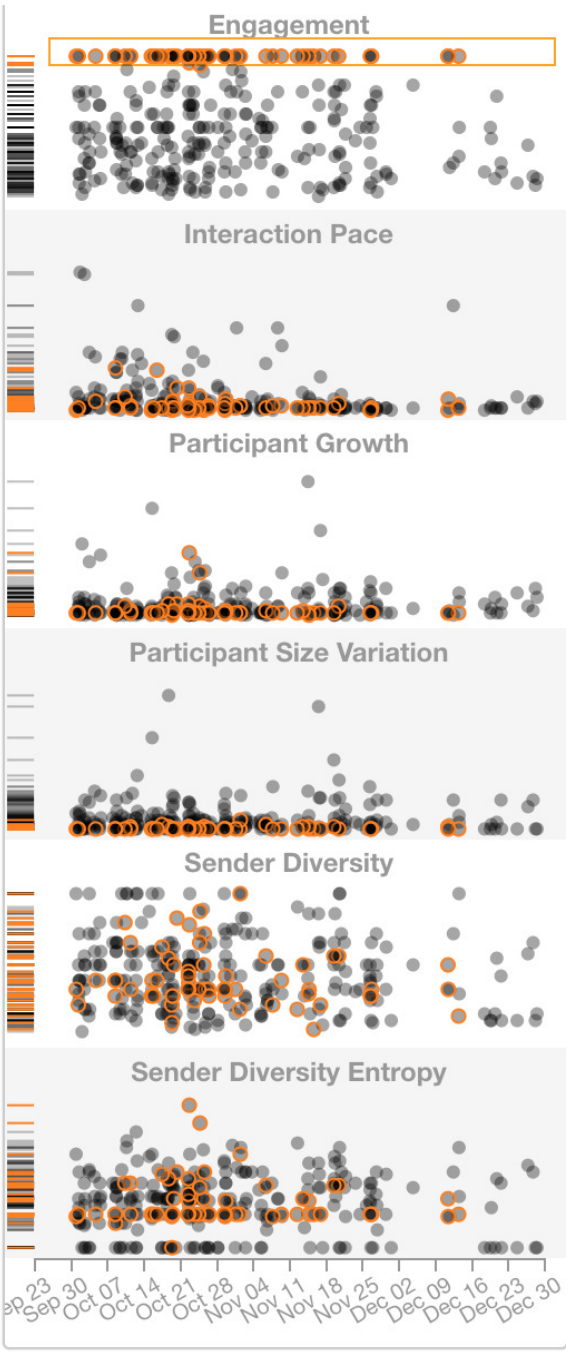


Figure 5.25: A visualisation of thread features. When high-level engagement is selected across time, the other thread features are highlighted indicating the nature of thread.

for each thread and apply a t-SNE projection [125] to 2D space. The projected threads are shown as a scatter plot in 5.22 (top-right). When threads are assigned classes, either manually or automatically through modelling as described later in Appendix A.3, the classes are encoded by filling thread circles with different colour hues. Threads that are suggested for labelling from the active learning model are highlighted with grey crosses.

These two feature views work on numerical feature values, which can be used to explore the features and check if they make sense and be useful in understanding the threads. They can also be used as starting points to identify interesting patterns.

Content View

From a single thread view, a particular node can be selected (where each node represents a message) to read the content exchanged which will be displayed in the content view. This gives an additional support to find if a particular thread selected is interesting or not.

In addition to these, we included "Data Labeling" as well. Analyst can label interesting threads based on their exploration and understanding. The marked/selected threads are highlighted and the threads that do not belong to that category are delighted once labeled.

5.3.2 Validation & Findings

This section demonstrates how the aforementioned views can be used together to discover interesting patterns in email communication. We demonstrate a scenario as shown in Figure. 5.26. A cluster of threads in Feature Projection view (top-right) is brushed, shown as yellow borders, the same threads are highlighted in the same way in Thread Features view and are shown in the Thread Overview as well. It turns out that these threads all have low engagement as evidenced in both views. Each column has one long dark bar and many short light bars, indicating that there is one single individual sends emails to all recipients and no one replies. Some of the individual names can help explain this such as

Figure 5.26: An example of combining visualisations to discover interesting threads. These threads are low-engaged and single-sender.

"enron.payroll" and "noreply". Selecting one thread in the Thread Overview (highlighted with dark orange) for further examination in the Thread Messages view. It is confirmed that there is a single individual sending email to the rest. Interestingly, it sends messages to each person separately at different time through out a day. Hovering each message to quickly check its content shown in the Message view: this is an advertising thread. Clicking on other brushed threads, we can confirm they share same send/receive pattern. Figure. 5.27 5.28 shows two other examples of interesting threads that can be discovered using our visualisations.

5.3.3 Empirical Evaluation

To evaluate our visualisation and the user experience of our system, we conducted a user evaluation to understand how the experts use the system in exploring, discovering and

Figure 5.27: An example of combining visualisations to discover interesting threads. These threads contain a small number of individuals and they all actively discuss, in almost a "ping-pong" style (one-to-one communication)

Figure 5.28: An example of combining visualisations to discover interesting threads. These threads contain a higher number of individuals and there are many individuals who are actively discussing.

interpreting the features and patterns that can be interesting in a given E-mail dataset (with threads). We report the process, how they approached giving examples of the kinds of patterns identified and categories generated. We had two experts from the company who had knowledge of both the domains (E-discovery) and technicality. Both were involved in the initial requirements analysis study and were regularly consulted during the development of the framework thus having a better understanding about the system. In this evaluation, we aimed to:

1. Understand what strategies experts employ when they are exploring the data set for discovering. Also, to understand to role of the specialised interactive visualisation views and features are perceived to be useful for experts in their investigation (R7-R9).

2. Understand how experts specify the characteristics of threads and name the cluster of threads (R8-R9).

3. Understand to what extent experts make use of active learning and how they respond to the labels returned (R10).

The experts were all introduced to the tool's features, using demo dataset (100 threads), in about 30 minutes before being given the tasks. The task was to use system to explore, identify threads of interest, and then categorise/label threads. We used two different datasets (with 300 threads in each dataset) for two experts to avoid any learning effect. Instead of asking abstract questions, we provided the tool to reflect the exploratory nature of the analysis. We asked the experts to explore the threads according to their own interests, discuss their thoughts aloud on the discovery and interpretation. The experts spent about 45 minutes to complete the task.

During the study, we primarily focused on gathering qualitative data such as experts' views and observations. The recorded observations were transcribed and coded to facilitate analysis. In addition, we captured interface actions to better understand the usage of the system, exploration and their interpretation.

Results and Analysis

We first explained to the experts our conceptual overview, interface, and analytical process of our system, and then, presented the Enron case study. The experts were expected to think-aloud when using the tool and we made observations. The experts spent a good amount of time in exploring and investigating various types of threads that could be of interest to them. As per our explanation to the experts and their investigation strategies, we express our analysis into three stages:

Pattern discovery and investigative process. For interactive visualisation exploration, the two experts employed two different strategies for discovering, finding threads of interest and communication types and made good use of the visualisations.

The expert (E1) started by skimming and scanning all the threads (T7) in the "Feature Projection" view to find different communication thread types (Fig. 5.29(a)). He said "I am interested to see all the different types of thread types using Feature Projection as a starting point". He observed the selection in the "Thread Features" view (multi-faceted thread features) to understand the nature of the threads selected (for example, high engagement, fast interaction in threads). The expert (E1) identified a portion of threads as felt those are interesting from a subset of threads (T8). Further, he viewed the threads in the "Thread overview" (group of threads) to compare the individuals' engagement, people-overlaps and time-overlaps in the threads (T10); and to understand the activities of a particular thread in the "Thread Message" view and then reading the messages linked to that thread in the "Message" view. From the exploration, he feels there are many types of thread communications and he was able to understand the changes in the conversation/thread characteristics there is something unusual (T9). The expert (E1) commented that "the interactive interface assists in understanding, exploring and comparing various threads and patterns of interest. It is helps in seeing different thread communication types". Furthermore, the expert (E1) appreciated the idea of having multiple visualisation views such as thread features, overview of threads selected and a detailed single thread view that helps in understanding the types of threads. The expert said the tool's ability to show a visual overview of

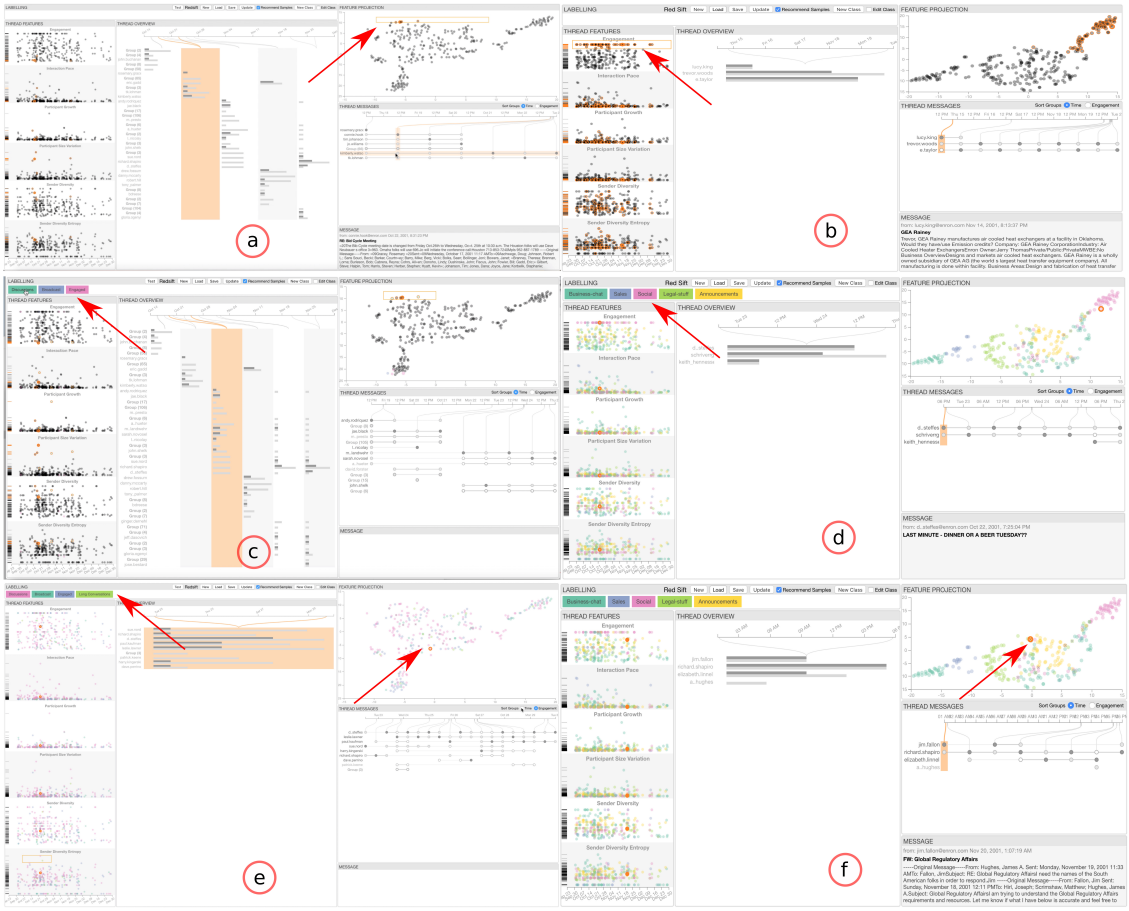


Figure 5.29: Screenshots of our system taken during the study with the experts (E1) & (E2) for investigating multi-faceted E-mail data. (a) E1 started with a selection of random cluster of threads in the Feature Projection view. (b) E2 started with a selection of random cluster of threads in the Feature Threads view (high engagement). (c) After quick exploration, E1 created three classes immediately namely "Discussions", "Broadcast" and "Engaged". (d) After investigation a group of threads in multiple visualisation views, E2 created three classes immediately namely "Business-chat, "Social, "Sales, "Legal-stuff and "Announcement. (e) With further investigation, based on the Active Learning recommendation, E1 identified some of the threads identified were long and he created a new class called "Long Conversations". (f) Based on the Active Learning recommendation, E2 worked on the samples very closely, and even continued all were exhausted.

all the thread features (multi-stacked scatter plots), thread comparison (bars in the thread overview), individual thread view (dots representing senders, recipients and the types with sorting feature), it's conversation and allowing the user to explore and navigate through facets. The expert said, "the tool is interesting, very efficient and comprehensive". This helps establish the email visualisation helps in exploring and comparing various threads in an effective and comprehensive way. Moreover, the visualisation tool was found to be easy to learn by the expert and execute modelling by exploring, owes to the fact of the 18-month long study where the expert was constantly informed about our approaches and visualisation views. He agreed that selecting a set of threads and then visually linking them to all the views helped him to quickly understand what a conversation is about and to focus on its most interesting parts, which is quite useful and helpful in the investigative analysis. This helps establish our visualisation helps in discovering interesting parts that can be helpful in investigation. The expert also acknowledged that in the conversation, he is interested to know who is replying to the emails initiated and who were bcc'd. However, the expert mentioned, when more data points are selected/considered, it becomes difficult to comprehend, explore smoothly and label the right threads.

The expert (E2) started exploring (T7) by selecting the high range of the "Participant Growth" in the "Thread Features" view and then checked the other visualisation views to find different communication thread types (Fig. 5.29(b)). The expert said "I select participant growth because more and more people are being involved and topic of discussion might be of interest". He checked the "Thread Overview", "Message" view and found Sheila.Nancey as a person of interest and identified the whole thread communication as business chat (T8). The expert found changes in the conversation which are unusual (T9). He further continued to compare multiple threads to understand Sheila's behaviour (T10). The expert worked mainly on the features rather than the clusterings on the "Feature Projection" space. He said, "I understand what the features are, and I can explain & relate to the features when constructing". He made good use of all the interactive visualisations except feature projection but he used feature ranges to look for interesting cases. This helps establish the email visualisation supports in exploring and comparing various thread

features to find some interestingness. Though the expert was comfortable with selecting "sparse data points" rather a big cluster of data in some of the thread features but he felt it was difficult to understand & interpret, selecting a particular thread in a cluster was a challenge.

Pattern specification and investigative process. For interactive visualisation specification, the two experts employed different approaches for specifying threads.

The expert (E1) selected different groups (clusters) of threads in the "Feature Projection", especially different clusters in different positions. After sifting through several threads in a selected cluster and investigating other supporting views, he often relied on the thread features (meta-data of the threads), thread comparisons displayed in the thread overview, and a detailed view of a particular thread and the contents exchanged. The expert labelled the thread types as "Discussions", "Broadcast" and "Engaged" based on the communication types in the single thread messages and the contents exchanged (Fig. 5.29(c)). That is, if the messages were being exchanged continuously for a long period of time, the expert labelled it as "Discussion". If the message was sent by one person to many individuals in the company without much of replies, the expert labelled it as "Broadcast". If all the recipients in the email were actively involved in the email thread, the expert labelled it as "Engaged". For many other clusters in the feature projection, based on his understanding about each thread or a group of threads, he assigned them to one of the classes created and updated the system to see how all the threads are assigned to the three classes created. He continued to investigate the threads by going back and forth to all the views and hovered over different features to find if the threads belong to the same class or to create a new class. Doing so, he created a new class "Long Conversation" and re-marked "Engaged" threads to this new class (Fig. 5.29(e)). The expert commented that "it is an innovative idea to label threads based on the exploration, navigation and analysis. Given the fact, it is iterative interactive modelling using other supporting views such as features view, thread comparison view and single-thread view, the process and accuracy of classifying the threads can be improved immensely". This navigational behaviour can be

observed from the sequence of actions made by the expert. However, at the same time he acknowledged that he tended to coordinate with the threads selected in the Thread Features, Thread Overview and Feature Projection where the related items were highlighted, so that he could get a sense of the type of communication. As a limitation, the expert mentioned, "it will be good to have multi-labelling for each thread". So, we understand thread need multiple labels (this was not an original requirement - we had to build the software to know this).

The expert (E2) was always focusing on the outliers in the "Thread Features" view. Making good use of the multi-faceted features, the expert immediately checked the thread overview, individual threads and the contents exchanged, the expert labelled the thread types (Fig. 5.29(d)) as "Business-Chat" (high-growth could be an indicator of business discussions which are of importance), "Social" (if the messages were related to farewell and parties, that is a single sender with people replying back only without much further engagement), "Sales" (if the messages were related to some sales & marketing), "Legal-trouble" (if the messages had legal information about the Enron case) and "Announcements" (if the messages were sent to a large group of people without any replies to it). The expert focuses on individual cases and tries to find representative threads and looks for expanding from them. He also said "I don't want an incorrect labelling to spread across the classification, so I mark only individual cases first and then expand checking carefully. If I marked groups, I would be making mistakes easily". He made good use of all the interactive visualisations except feature projection. As a limitation, the expert mentioned, "it will be good to have all the threads related to a particular label being highlighted when selected".

The experts were able to identify meaningful patterns in different ways which gives us some specific detailed design requirements that we can feed back into iterative system redesign. The experts made good use of the labelling. This helps establish exploratory visual labelling of threads is useful and important.

Pattern modelling and investigative process. For interactive visualisation specification, the two experts again employed different approaches for modelling threads.

The expert (E1) did not make much use of the active learning and the recommended samples. The response to the labels returned from the model was not deeply investigated. There was only one instance where relabelling was performed when an issue was spotted. The expert mentioned the active learning helped him re-label the thread.

The expert (E2) made good use of the active learning and the recommended samples. The expert worked on the samples very closely, and even continued all were exhausted (Fig. 5.29(f)). The expert mentioned the active learning helped him to a greater extent to mark many of the threads.

Other observations. The expert (E1) considered that the interface of our system can be a useful tool for the analysts in the organisations, in the field of investigation and can be used as a personal analytic tool as well. He specifically pointed out that the integration of task relevance analysis and data modelling enhances the applicability of the system for transferring to real-world tasks, which can be used for labelling emails in the cyber security domain (to prevent cyber-attacks), for analysing Slack communication and any other communication/social media threads. He mentioned "we would like to use this tool on our platform to conduct cyber security analysis".

The expert (E2) said "the interface is good and helps in navigation & exploration. Easy to identify and classify the threads based on the conversations or emails exchanged. The complete pipeline of the workflow is good and it is definitely useful for investigation and we will use it as a solution for our products planned".

Preference. When the expert was asked to express his experience using the visual thread interface, the answers were generally in favour of our tool, due to its ability to show a visual overview of all the thread features (multi-stacked scatter plots), thread comparison (bars in the thread overview), individual thread view (dots representing senders, recipients and the types with sorting feature), it's conversation and allowing the user to explore and navigate through facets. The expert said, "the tool is interesting, very efficient and comprehensive". Moreover, the visualisation tool was found to be easy to learn by the expert and execute modelling by exploring. He agreed that selecting a set of threads and then visually linking

them to all the views helped him to quickly understand what a conversation is about and to focus on its most interesting parts, which is quite useful and helpful in the investigative analysis. He also acknowledged that in the conversation, he is interested to know who is replying to the emails initiated and who were bcc'd. The expert considered that the interface of our system can be a useful tool for the analysts in the organisations, in the field of investigation and can be used as a personal analytic tool as well. He specifically pointed out that the integration of task relevance analysis and data modelling enhances the applicability of the system for real-world transfer learning tasks, which can be used for labelling emails in the cyber security domain (to prevent cyber-attacks), for analysing Slack communication and any other communication/social media threads. He mentioned "we would like to use this tool on our platform to conduct cyber security analysis".

Overall. The Red Sift experts were satisfied with our visualisation to analyse thread patters and discover interesting information in the threads. The solution is currently deployed in the GMAIL to analyse organisation emails related to cyber attacks to discover interesting thread patterns and individuals (mentioned in the previous paragraph). Again, due to commercial sensitivity and confidentiality clauses, we are not able to include the organisation's email visualisation. The suggestions from the experts are discussed in the Chapter 6.

Summary of Evaluation

To evaluate our visualisation, visual exploration of thread information, and the user experience of our system, we conducted a user evaluation to understand how the experts use the system in exploring, discovering and interpreting the features and patterns in a given E-mail dataset. We reported the process, how they approached giving examples of the kinds of patterns identified and categories generated. We had two experts from the company who had knowledge of both the domains (E-discovery) and technicality. Both were involved in the initial requirements analysis study and were regularly consulted during the development of the framework thus having a better understanding about the system. Dur-

ing the study, we primarily focused on gathering qualitative data such as experts' views and observations. The observations recorded were transcribed and coded to facilitate analysis. In addition, we captured interface actions to better understand the usage of the system, exploration and their interpretation. Some of the positive findings are:

1. We were able to understand what strategies experts employ when they are exploring the data set for discovering. Also, to understand to role of the specialised interactive visualisation views and features are perceived to be useful for experts in their investigation.

2. We were able to understand how experts specify the characteristics of threads and name the cluster of threads.

3. We were able to understand to what extent experts make use of active learning and how they respond to the labels returned.

5.3.4 Learnings

L_{dv3}**: Understanding the dynamics of topics in threads must be considered.** In the current visualisation, analysts have a good understanding of how a thread structure of interest looks like and which individuals are active/passive and included/excluded in the conversation. For example, while expert (E2) was working out a task (T7), he explored and "Participant Growth" in the "Thread Features" view as more number of people were involved and topic of discussion might be of interest. The expert found Sheila.Nancey as a person of interest and identified the whole thread communication as business chat (T8). During the design study process, we learnt introducing visualisation-assisted feature engineering will help in deriving features such as participant growth, response time in replying and so on and support in finding/discovering interesting information. Also, we learnt introducing visualisation-assisted active learning will help in classifying emails based on exploration. During the design process, the three key challenges (C1-C3) considering the design requirements (R7-R10), analysis goal (AG3) and tasks (T7-T10) were addressed and the visualisation was able to reveal things that were interesting. This confirmed our

solutions were working specifically addressed the question *"To what extent visualisation can support analysts in discovering interesting threads information in the E-mail communication data?"*. When we demonstrated the functionality of the tool that supports pattern discovery (for discovering interestingness in the patterns), the expert (E1) liked the thread features which helps in drilling down the data of interest and observe the individuals in the single thread novel visualisation. The expert also liked the idea of labelling the communication type based on one's interest which helps in classifying the complete data. All the points are mentioned in the Appendix A.7. Also, the expert (E2) said "the interface is good and helps in navigation & exploration. Easy to identify and classify the threads based on the conversations or emails exchanged. The complete pipeline of the workflow is good and it is definitely useful for investigation and we will use it as a solution for our products planned". Both the experts, expressed their liking towards the multi-faceted exploration and multi-granular approach, where the flow from the high-level view of all the threads to the low-level view of each thread and the content exchanged. For analysts to identify interesting points and seamlessly switch between the different levels of overview and the actual thread view, integrating the high level view of the thread features, the visual overview of the thread (for comparing all the threads selected), and the actual thread view (detailed view) helps in the investigation process. The prototype addressed all the challenges (C1-C3) and the solution was deployed on their platform. Though the Enron use cases and tasks fit in with what Red Sift wanted to do, the expert (E1) came up with questions such as "Can you find what kind of topics the senders and receivers discussed in a thread?", "Can you find which individuals in a thread discussed a topic on California Energy Deal?". The main lesson we learnt is to focus on topics to understand why two individuals contact each other, what are their common interests/topics and context. The topics can be extracted from both subjects and body of the email which will analysts a way to delve interesting stories between two individuals of interest. The current visualisation helps in exploring threads based on various features, investigate a multiple sets of threads and deeply investigate a single thread of interest. In that, individuals communication pattern can be observed, that is how they have sent/received, whether the individuals

were passive/active in the complete conversation and which are the individuals who were included/excluded in a thread. Interestingly, the analysts can specify the communication types (such as announcements, meetings, ping-pong discussion, etc.). Adding a topic modelling approach would add more value to the communication types, which can support investigation cases. Based on the personal validation (also with the experts) and empirical evaluation, we understood visualising thread patterns can be extended to topic analysis that can be effective for investigations. This has given us a base to move into the next phase of building topics-based Email Visualisation by again extending the set of requirements, analysis goals and tasks (which will be our future work). The further details are discussed in the next chapter (Chapter 6).

5.4 Conclusion

In this chapter, we discussed the design process which helps in facilitating data abstraction, visual encoding and interaction mechanisms [156] and addressed our second objective in the study (**O2**). The design process had developing designs (visualisations) and validation which we report on an 36-month long design study conducted as a multidisciplinary team of visualisation researchers and E-discovery experts. The complete design process was broken down into three phases (temporal, individuals and thread analysis). In each phase, we took a user-centric approach in designing visualisations (low-fidelity to high-fidelity) and in each stage, we validated the solution with the experts from the organisation, which addresses our third objective in the study (**O3**). The complete stage was iterative, as the analysts tried to understand the research question that is part of the three phases throughout the three years of the design study in a regular interview session.

The design & validation phases led into "Triangulation" [84] which facilitates validation of each separate design and reflect upon this based on iterative user-centered design approach, interviews, feedbacks and features that are adopted and deployed by organisation collaborators (Red Sift). The solutions from each of the phases are deployed as standalone solutions without integrating each of them to address the 3 different development phases.

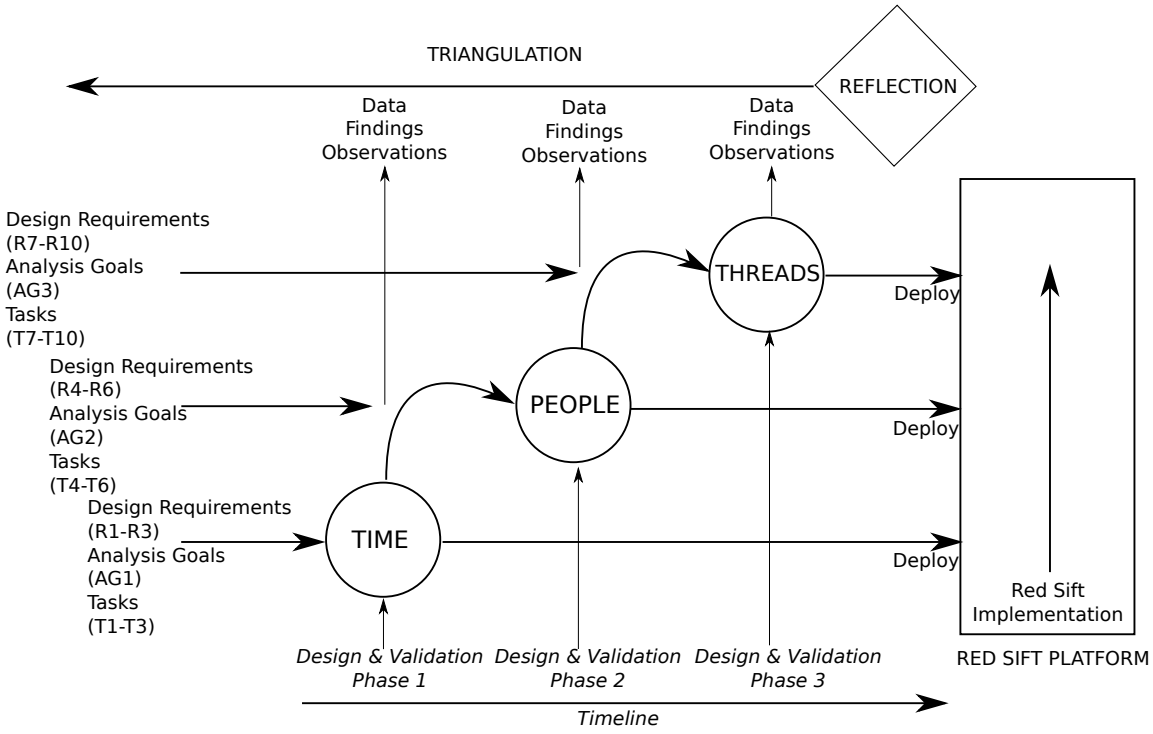


Figure 5.30: Our design & development stage has three main phases (time, people and threads) for visualising patterns in the E-mail data. Each of the phase is mapped to a particular set of design requirements, analysis goals, and tasks based on the interviews with the experts. We iterate through each individual design that addresses a different sub problem, to observe, generate qualitative data, suggest findings and learnings. We validate each separate design and reflect upon this based on feedback and features that are adopted and deployed by Red Sift. And this learning feeds into the next design study.

Triangulation not only supports validation but helps in multi-perspective analysis to reflect on the work conducted [68]. We use Triangulation analysis to increase confidence in the findings through iterative interviews conducted, confirmation of validation and through the confirmation of results achieved [84], as shown in Figure 5.30. In our research, we have both validation and confirmation of interactive visual solutions being adopted and deployed in the collaborators' platform [84]. The combination of findings from two or more phases/methods/approaches will provide a more comprehensive picture of the results than either approach could do alone [84]. Our conclusion from each phase (L_{dv1}, L_{dv2} and L_{dv3}) are as follows:

- Each phase was built on the other. Thread-based visualisation L_{dv3} was built on Individuals-based L_{dv2} and this was built on temporal-based visualisation L_{dv1}. There is a level of complexity relationship between the three of these (increasing) and we learned from one and moved to the other. In the final phase, we delivered a final solution that address all of these in a much more integrated manner. We don't think of temporal information in isolation or threads in isolation. We presented a narrative that with the analysis of threads (in phase 3) we looked at tasks that require analysts to understand, time, people, content and conversations/communications concurrently.

- We came up with the three phases based on the data availability, design requirements and task complexity. In the Phase 1, the data identified did not have organisation roles and it is was in a non-threaded form. In the Phase 2, we identified a table of organisation roles for each of the employee in the Enron E-mail data. We manually parsed the data, merged two E-mail datasets into one to understand how designations / organisation roles can help in our analysis (based on the requirements captured). In the Phase 3, with the help of our collaborators, the data was engineered to group all the messages forming a thread based on the subject header in the email. This formed a thread-based email data for further analysing time, individuals, threads and content. In this process we learnt, as there was a level of complexity increase from

Phase 1 to 3, new set of design requirements and tasks were considered and the data attributes were added as well.

- We also came up with the three phases based on the validation (which is nothing but the current problems). In the Phase 1, temporal-based email visualisation, the three key challenges (C1-C3) considering the design requirements (R1-R3), analysis goal (AG1) and tasks (T1-T3) were addressed and the visualisation was able to reveal things that were interesting (that is, discovering and characterising time period(s) of interest). This confirmed our solutions were working and specifically addressed the question *"To what extent visualisation can support analysts in discovering interesting temporal information in the E-mail communication data?"*. However, the questions raised by the experts helped us move into the next phase of building Individuals-based Email Visualisation by extending the set of requirements, analysis goals and tasks. Similarly, in the Phase 2, the three key challenges (C1-C3) considering the design requirements (R4-R6), analysis goal (AG2) and tasks (T4-T6) were addressed and the visualisation was able to reveal things that were interesting (that is, discovering and characterising individual(s) of interest). This also confirmed our solutions were working for the question *"To what extent visualisation can support analysts in discovering interesting individuals information in the E-mail communication data?"*. Again, the questions raised by the experts helped us move into the next phase of building Threads-based Email Visualisation by extending the set of requirements, analysis goals and tasks. During the third phase of the design process, the three key challenges (C1-C3) considering the design requirements (R7-R10), analysis goal (AG3) and tasks (T7-T10) were also addressed and the visualisation was able to reveal things that were interesting (that is, discovering and characterising threads(s) of interest). This also confirmed our solutions were working specifically addressed the question *"To what extent visualisation can support analysts in discovering interesting threads information in the E-mail communication data?"*. Another set of questions raised by the experts made us think from future direction.

- After each phase of the design study, the Red Sift experts explored the tool by themselves and were satisfied with our visualisation to discover and find interesting patterns based on the analysis goals and tasks extracted. All the three solutions are deployed in the Google Suite to analyse their organisation emails to discover interesting patterns related to their business collaborations. Due to commercial sensitivity and confidentiality clauses, we are not able to include the organisation's email visualisation. In this process, we learnt if the solutions are usable, then it can be deployed.

- In all the three phases, the design process was same. We started designing with low-fidelity, moved to medium and later high-fidelity, which led to develop an interactive visualisation that supports in multi-faceted exploration and multi-granular analysis. We also addressed all the three challenges - finding interesting subsets within the large volume of data (C1); detecting changes in the E-mail communication due to its complex and dynamic nature (C2); and open-ended data exploration to find interesting communication patterns (C3). Iterative interviews conducted and the solutions validated constantly are the strengths in our research which addressed our research question *"To what extent visualisations can support analysts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation?"*. We also learnt, the systematic approach to address our research question helped us deploy the solutions in the organisation's platform.

We reflect on our research by discussing our contributions, findings and limitations in Chapter 6.

Chapter 6

Post-condition Phase: Reflection & Conclusion Stage

In the Design Study Methodology (DSM) [156], the final stage is the reflection and conclusion. In the preceding chapter, design study has been described to explore different aspects of patter-oriented interactive visualisation, covering three main features (time, individuals and threads), through three phases (DV Phase 1 - DV Phase 3), considering three challenges (**C1-C3**), ten design requirements (**R1-R10**), three main analysis goals (**AG1-AG3**) and ten tasks (**T1-T10**). In this chapter, we take a step back to think in-depth about the design study, reflect on the three objectives (**O1-O3**), and further generalise those practical experiences, by providing theories, learnings, findings, implications and principles to the future design and study of email communication analysis.

6.1 Reflection

In this section, we reflect on observations made during the complete three years of study and discuss a number of suggestions to guide the utilisation of the visualisations. We discuss the reflections based on the aim of the study, research question and three objectives (O1-O3) we considered in the thesis.

6.1.1 Revisiting Research Question and Objectives

In this thesis we tackled the problem of the analysis of email communication data and investigated the use of interactive visualisation to address the challenges faced by the domain experts. The overall goal of the thesis was to design and develop interactive visual solutions to explore and find/discover relevant/interesting information from an investigation perspective to support in an organisation specialising in E-discovery. As discussed in Chapter 1, despite the use of visualisations in various domains, there were no optimal solutions to support organisations in E-discovery compliance [63]. Visually identifying/finding/discovering various information or groups of data objects from multiple perspectives in E-mail communications were under-explored and under-investigated [63]. There was a need to analyse the data in efficient ways and visualisation was found to be a good solution [63] to support the organisation compliance and analysts such that the whole process could be made proactive, preventive and support legal evidence. So, this research was interested in answering the question *"To what extent visualisation can support analysts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation?"*. The research question helped us generate three objectives (O1-O3) to answer through understanding knowledge gap (learning), domain characterisation, design & validation, deployment and reflection. We are re-introducing the three objectives again for the benefit of readers.

Objectives of the research

O1: *Develop design requirements:* understand the E-discovery domain, identify the knowledge gap and develop a rich understanding of challenges, tasks and requirements (specific to E-mail communication data).

O2: *Design and develop visual methods:* design and develop interactive visualisations based on the domain requirements to effectively navigate and explore within data to uncover relevant/interesting information and relationships within the multi-facets such that solutions can be used as an evidence in investigation.

O3: *Validation of visual methods:* validate and re-access the developed interactive visualisation prototypes by conducting validation and empirical studies. Express findings based on the user-centric approach & evaluation that can help analysts to investigate and navigate E-mail data productively and identify various interesting information relevant to the investigation.

Objective 1 (O1):

To answer the research question and the objective 1, we first examined various visualisation and interaction techniques related to digital communication data analysis, the problems connected with their use and a number of methods/approaches for addressing the E-discovery problems, which is reported in Chapter 3 (Section 3.3.2). We also conducted unstructured interviews with the experts to understand the E-discovery domain, identified the challenges (C1-C3), captured the design requirements (R1-R10) and extracted the analysis goals (AG1-AG3) and tasks (T1-T10) to design interactive visualisation solutions, which is reported in Chapter 4 (Section 4.2.3 to 4.2.5) and the notes captured are in the Appendix A.7. In this way, we accomplished our first objective (O1). The finding (F1) is discussed in Section 6.1.3.

Based on Sedlmair et al. [156], we understood if there is a specific real-world problem faced by the domain experts using real data, we need to first capture user requirements and identify knowledge gap which can be iterative. BubbleNet [126], Concept Explorer [97] and

Polimaps [166] are visualisations built for investigation considered iterative user-entered design process to record all the user requirements before designing and implementing visual solutions.

Our first objective (O1) gave us a base to employ user-centered design (UCD) which involved iterative design process for three years and we built several interactive visual solutions based on the requirements and tasks (this formed second objective - O2). The iterative interviews helped us to capture three questions which eventually formed three phases of our design study based on the data availability, design requirements and task complexity:

DV Phase 1: To what extent visualisation can support analysts in discovering interesting temporal information in the E-mail communication data?

DV Phase 2: To what extent visualisation can support analysts in discovering interesting individuals with their designations (organisation roles) in the E-mail communication data?

DV Phase 3: To what extent visualisation can support analysts in discovering interesting individual behaviour (conversations) in the E-mail communication data?

The three phases led into "Triangulation" [84] which facilitates validation of each separate design and reflect upon this based on iterative interviews, feedbacks and features that are adopted and deployed by organisation collaborators (Red Sift). As discussed in Chapter 5, the triangulation analysis helped us to increase confidence in the findings through iterative interviews conducted, confirmation of validation and through the confirmation of results achieved [84], as shown in Figure 5.30.

So, we claim *"Iterative user-centered design approach supported in understanding user requirements (from the experts) which helped us achieve Objective 1 (O1)"*.

Objective 2 (O2):

Based on Sedlmair et al. [156], we understood for real-world problem faced by the domain experts using real data, we need to design and develop solutions that will be

iterative. BubbleNet [126], Concept Explorer [97] and Polimaps [166] are visualisations built for investigation. All the three papers considered iterative user-entered design process to design and implement visual solutions.

In our work, the objective 2 was addressed based on the objective 1. The analysts moved between the three questions (DV Phase 1-3) throughout the three years of the design study in a regular interview session, as we followed iterative user-centered design approach. For example, some of the requirements had a mix of time, individuals, threads and contents during the three years. In the first design phase (building Temporal-based Email Visualisation), as discussed in L_{dv1} of Chapter 5, the expert (E1) was able to explore the temporal features, discover patterns and find some interestingness. However, searching and selecting a particular individual of interest and finding/investigating their connections was not possible. This made us focus more on finding various individuals and understanding their connections in terms of emails being sent or received over a period of time, understanding individuals and their connections with/without their designations (organisation roles), comparing and finding who sent/received most/least (discussed in the Appendix A.7).

In the second design phase, as discussed in L_{dv2} of Chapter 5, the expert (E1) was able to explore the individuals, discover patterns and find some interestingness considering both with/without organisation roles. However, understanding a thread structure of how individuals behave/treated was not possible, that is in terms of individuals being active/passive or included/excluded in a conversation. This made us focus more on visualising thread patterns to understand the communication structure, such that we will be able to infer commonalities and differences within sets of threads for the purposes of specifying the communication types (such as announcements, one-way communication, two-way communication, etc.). The notes are captured in the Appendix A.7.

In the third design phase, as discussed in L_{dv3} of Chapter 5, the expert (E1) was able to explore the thread features which helped them in drilling down the data of interest and observe the individuals in the single thread novel visualisation. The expert also liked the idea of labelling the communication type based on one's interest which helps in classifying

the complete data. The experts were able to investigate and navigate within communication data, identify/find/discover various patterns, trends, anomalies and information relevant to investigation (based on their interest). The expert (E2) said "the interface is good and helps in navigation & exploration. Easy to identify and classify the threads based on the conversations or emails exchanged. The complete pipeline of the workflow is good and it is definitely useful for investigation and we will use it as a solution for our products planned". The prototype addressed all the challenges (C1-C3) and the solution was deployed on their platform. The notes are captured in the Appendix A.7.

Though the Enron use cases and tasks fit in with what Red Sift wanted to do, the expert (E1) came up with various questions while validating the solutions in all the phases. This shifted our focus from temporal analysis to individuals and then to threads with individuals. The objective 2 was addressed based on the objective 1 (by capturing domain requirements). In all the three phases, considering user-centered design approach, we addressed all the challenges (C1-C3) identified, design requirements (R1-R10), analysis goals (AG1-AG3) and tasks (T1-T10) captured/extracted by designing an interaction visualisation solution that can support analysts to navigate and explore in finding/discovering interesting information relevant to an investigation. The solutions from each of the three phases are deployed as standalone solutions by organisation collaborators (Red Sift) without integrating each of them to address the 3 different development phases.

So, we claim, *"Iterative user-centered design approach supported in designing interactive visualisation solutions in an applied context with the experts which helped us achieve Objective 2 (O2)"*.

Objective 3 (O3):

In each of the three phases, the solutions were validated personally (personal validation) and discussed with the experts, as we followed iterative user-centered design approach, discussed in the Appendix A.7). Based on the new set of design requirements and tasks,

we moved from one phase to the other. As discussed in Chapter 5 (conclusion section), the design & validation phases led to "Triangulation" [84] which facilitates validation of each separate design and reflect upon this based on feedback and features that are adopted and deployed by organisation collaborators (Red Sift). The combination of findings from the three phases provides a more comprehensive picture of the results. Our learnings from each phase (L_{dv1}, L_{dv2} and L_{dv3}) helped us understand the level of complexity involved - in terms of email data, problems, design requirements, analysis goals and tasks.

By validating the solutions with the experts iteratively (discussed in Chapter 5 and interview notes captured in the Appendix A.7)), we addressed all the three challenges - finding interesting subsets within the large volume of data (C1); detecting changes in the E-mail communication due to its complex and dynamic nature (C2); and open-ended data exploration to find interesting communication patterns (C3). After each phase of the design study, the Red Sift experts explored the tool by themselves and were satisfied with our visualisation to discover and find interesting patterns that are relevant to the investigation. All the three solutions are deployed in the Google Suite to analyse their organisation emails to discover interesting patterns related to their business collaborations.

Based on Sedlmair et al. [156], we again understood for real-world problem faced by the domain experts, we need to iteratively validate the designs and solutions. BubbleNet [126], Concept Explorer [97] and Polimaps [166] work revolved around iterative user-entered design process to validate the designs and solutions developed.

So, we claim *"Iterative user-centered design approach supported in validating our solutions with the experts which helped us achieve Objective 3 (O3)"*.

Through evidence in achieving our objectives (O1-O3), we demonstrate our interactive visualisations, applied through design study [156] supports analysts in making sense of a corpus of E-mail, facilitate discovery and insight in the exploration of email communication for E-discovery to the extent that our partners are now implementing the techniques that have been established to do just this in a commercial context. Through this, we demon-

strate our success in addressing our research question though we have several limitations that can be addressed in the future (discussed in the next section).

Contribution to Knowledge

In this section, we summarise the contributions based on the three years of design study we conducted. We discuss the knowledge we have contributed based on the previous available knowledge.

- **Characterised the domain and tasks for E-discovery.** Based on our exploration, we understood E-discovery Compliance in an organisation is one of the chief drivers for organisations to take on any legal actions against individuals working in their organisation or legal actions against other organisations [67]. Based on the interviews, we also understood the current model of organisation compliance with E-discovery for investigating email communication is manual. The complete process (data gathering to legal actions, including finding interestingness or evidence in the data) is tedious and time consuming. Based on the interviews with the experts, we identified three key challenges (C1-C3), captured ten design requirements (R1-R10) and extracted ten tasks (T1-T10) related to investigation that needed to be addressed. Iterative user-centered design approach helped in understanding user requirements (from the experts). Chapter 4 contributed to answering the questions such as "What are the current challenges?", "What questions can be answered?" and "What can be visualised?". These kind of analysis helped in conducting interviews with the experts and examine various techniques to support analysts in investigation.

- **Identified knowledge gap and provided overview of the existing techniques.** We conducted a literature review that helped us in gaining a good knowledge and understanding of the domain, state-of-the-art, visual design principles, interaction techniques used, visualisation tasks, methods, techniques etc. and assisted in comparing findings that helped us in designing our visualisations for email communication. The research work also helped us to understand the concepts, ideas and reasons

for using different approaches to address investigation challenges and identify the research/knowledge gap. Based on the visualisation features and taxonomic harmonisation, we identified four main entities/features in visualising email communication, that is temporal, individuals and contents, including thread features. All the papers considered four main features in various combination for analysing email communication but there is no smooth interaction between all the four. However, harmonising the taxonomy of entities based on the visualisations related to email communication data aided in understanding the different entities, the association between them and their limitations. As a positive, the work also helped us to understand the visualisation approaches that can be used for discovering interesting information that can be helpful in an investigation. We examined various visualisation techniques related to digital communication data analysis, the problems connected with their use and a number of methods/approaches for addressing these problems. Chapter 3 contributed to answering the questions such as "How the data can be visualised?", "Which representations can be effective?" and "What interactions can be used?". These kind of analysis helped in understanding how to build an interactive visualisation that can help in discovering interesting information related to the investigation.

- **Designed and Developed interactive visualisation solutions.** Iterative collaborative design approach (user-centric design approach) supported in designing and developing interactive visualisation solutions to support E-discovery analysts. In the three years of collaboration with the experts, we had three phases (temporal, individuals and thread analysis). In each phase, we took a user-centric approach in designing visualisations (low-fidelity to high-fidelity) and in each stage, we validated the solution with the experts from the organisation. Each phase was built on the other, where we delivered an integrated solution that supports analysts to understand time, people, content and conversations/communications at the same time. We observed that interactive visualisation solutions (multi-faceted exploration and multi-granular analysis) can support analysts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery

Investigation. Specifically, interactive visualisation assisted active learning helped in classifying communications relevant to investigation. The novel visualisation for a single thread analysis helped in revealing hidden patterns of several individuals who were secretly copied (bcc) when sensitive information was broadcasted. This thesis contributes with insights and reflections on the effectiveness of particular design choices where we learnt conventional visualisations and novel visualisations must be considered based on the design requirements. We argue that using conventional visualisations considering all the features will help in exploring and discovering interesting information that can be relevant to investigations. To find nuances in a communication, novel visualisations can be considered which can help in discovering hidden information relevant to an investigation case. Chapter 5 contributed to answering the three questions (which formed three phases) such as "To what extent visualisation can support analysts in discovering interesting temporal information in the E-mail communication data??", "To what extent visualisation can support analysts in discovering interesting individuals with their designations (organisation roles) in the E-mail communication data?" and "To what extent visualisations can support in discovering interesting individual behaviour (conversations) in the E-mail communication data?". We answered the three questions by demonstrating our interactive visualisations, applied through design study can support analysts in making sense of a corpus of E-mail that can facilitate discovery and insight by exploring email communication for investigations.

- **Validated the solutions** - Iterative user-centered design approach supported in validating our solutions with the experts throughout the study. To evaluate our visualisation and the user experience of our system, we conducted an empirical evaluation to understand how the experts use the system in exploring, discovering and interpreting the features and patterns that can be interesting in a given E-mail dataset. We also made observations on how experts interact with information, how they use interactive classification methods and discover various information of interest. We draw upon and demonstrate to what extent visualisation can support analysts and

how the solutions are effective in an applied context (discussed in Chapter 5).

- **Deployed Solutions in the Collaborator's Platform.** After each phase of the design study, the Red Sift experts explored the tool by themselves and were satisfied with our visualisation to discover and find interesting patterns. All the three solutions were deployed in Red Sift's platform connecting to Google Suite to analyse their organisation emails to discover interesting patterns related to their business collaborations. The design study helped in critically understanding various design, developing them and understanding its effects through analytical and practical way of inquiring to support email investigation, through the iterative process of deploying solutions in the collaborators platform (discussed in Chapter 5 & 6). Initially, we questioned ourself "Can we provide a deployable solution that can support real live data?". Our research question, objectives and three design phases helped us achieve it.

- **Lesson Learnings & Principles** - Based on each phase of the study, we learnt many lessons and proposed principles to inform future use of visualisation in E-discovery and Digital Forensics (discussed in the next section).

6.1.2 Limitations

Visual scalability: Though the organisation experts liked our solutions, the main limitation of our work is the scale of the data. The raw Enron corpus contained 619,446 messages with 158 users (sender & receivers) over a 4 year period. In the cleaned Enron corpus, there are 200,399 messages with 30,091 threads with the same number of users in the corpus with timestamp and the message type (to, cc, bcc) in which the messages were sent to recipients. Specifically, we worked with a subset of a hundred users for a period of 2 years (2000 and 2001), and these hundred users were chosen from a subset of 300 threads, which had close to 10,000 emails. In our current tool, we have provided a drill-down approach to filter down the information of interest and label them for classification using active learning approach. The tool can lead users from a large data chunk to a portion where there

is interesting information. Since we aim to support organisations of any size to address investigations concerning email communications, we can expect datasets to be much larger than the current one. So, we can consider the classic visual information-seeking mantra proposed by Shneiderman [158] that summarises many guidelines for information design and interaction techniques for effective visualisation of information: overview first, zoom and filter (& sampling), then details-on-demand. Creating an overview and providing cues for further exploration is challenging with large datasets. Search, show context, expand on demand [174] is a more appropriate approach that can support in visual E-discovery and engage in scalable pattern discovery on large communication data. We can also introduce data aggregation techniques [66] that can represent higher number of threads and use a number of heuristics to automatically filter the large volumes of information that is unlikely to be useful. Also, develop different strategies to create models and classes effectively that can support investigations. By experimenting with small data, we were able to explore and discover various patterns, classified information, found interesting communication types such as long discussions, announcements, ping-pong discussions and identified several individuals who were secretly copied in a particular thread. Small data methods cannot be directly applied to the big data, we need better and effective techniques. As a future work, merging both visualisation and data aggregation techniques might aid in producing effective results while exploring the data and extracting deeper insights [66]. So, visual scalability techniques needs to be implemented to improve exploration and classification. We argue that this is a positive start towards a great journey in deploying to big data.

Visual clutter: The experts felt the chronology/time-ordering of the email sequence for each of the thread (single thread analysis) helped in understanding the communication type. The arrival sequence of messages which create a thread helps to illustrate the evolution of a thread, including which messages came first, and which is the most recent message. For example, we were able to examine a particular message, and we were able to trace back through the chain of earlier messages which led back to the root of the thread. In addition, we were able to see which messages subsequently respond to a particular message clearly.

These relationships gave important contexts for each message in a thread, their evolution, time-periods and the individuals involved that are relevant for an investigation. However, there were clutter in the thread features and multi-thread analysis. Though we developed a visual recommendation system using an active learning algorithm to find various categories for communication in the complete dataset, as there were no known groupings and visual clutter doesn't give any detailed indications. Structure of data and effectiveness of visual representation are not unrelated. If we have a dataset or a subset of a data with a high number of individuals, their connections over a large period of time, the axes (x and y axis) might become overcrowded and unreadable, causing it to get cluttered. We might need to have more iterative interviews/workshops with the organisation experts to address this problem. From our previous interviews, we could either split the data into smaller parts (based on time periods of interest) or introduce scrolling and/or panning. In the thread visualisation, when more data points are selected/considered, the system gets complicated. It becomes difficult for an analyst to comprehend, explore smoothly and classify. So, basically large sets would be an issue, so sampling would be a good idea. We are planning to work on the data engineering process and introduce additional clutter reduction techniques [66] so the system works for higher number of threads as well as larger ones found in the communication and support evolution of threads. We argue introducing visual clutter reduction techniques will help in improving readability.

Visual comparison: From the iterative interviews (in the design & validation phase 2), we understand the E-discovery analysts have difficulty in defining anomalies/abnormalities. In fact, "anomalous behaviour" is hard to define and we need a robust model of normality to be define what is "normal" and be able to effectively detect anomalies. However, in the case of multi-faceted data, there can be many ways to model normality and different data objects can be marked as anomalous from different perspectives, hence the need for flexible ways of defining normals is of utmost important. Some of the current techniques/approaches on anomaly detections are not easy to fit into real-world application due to their cumbersome approach, especially when considered multi-faceted communi-

cation data over time. So, representations must be simple and efficient to identify and detect anomalous behaviours in data over time. One of the example questions is "How to find out whether an individual behaved "differently" when compared with another individual or a group of individuals at a given/selected time-frame?". To address this kind of question, we can think of introducing comparison strategies [104][105] to understand the changes between the two selected subsets of data to find more meaningful interestingness. To the best of our knowledge, there are no comparative techniques/approaches for multi-faceted email analysis that can aid in supporting comparison of subsets of data. So, we can design solutions that are effective for displaying multiple relationships and also help in comparing information when placed close together or side by side to find similarities and differences. We argue introducing comparison techniques and strategies will aid in finding and discovering anomalous behaviour.

6.1.3 Findings

The findings of our research is based upon the Design Study Methodology (DSM) [156] we applied to gather information which forms a knowledge contribution are discussed below.

F1: An iterative user-centered design approach helped in understanding E-discovery domain and the investigation needs: The question such as "How to collaborate with the experts and provide a deployable solution that can support real live data?" made us consider an iterative user-centered approach. Regular meetings and interviews with the domain experts helped us understand the E-discovery domain and what is real-world investigation all about. Following Sedlmair et al. [156], the iterative iterviews helped us understand the investigator characteristics (needs, wants and limitations) throughout the whole design process (discussed in Chapter 4 & 5). We also learnt about expert's background, expertise, behaviour, goals, as well as their work environment and familiarity with technologies to support organisation compliance (notes captured in the Appendix A.7). For example, BubbleNet [126] is a visualisation built for investigation, adopted DSM [156], that considered iterative user-entered design process to record all the user requirements before designing and implementing visual solutions. We followed the

same approach in understanding E-discovery domain and the investigation needs. As discussed in Chapter 4 (Section 4.2.1), in the Figure 4.1, the complete process (data gathering to legal actions) of the current E-discovery model for organisation investigation is cumbersome (also discussed in Chapter 3). The real-world Enron case helped us in coming up with real use cases and tasks that are used in investigating a case from Digital Forensics and E-discovery point of view. To best of our knowledge, we fully implemented a UCD approach through an iterative collaborative design process involving prototype validation and an evaluation of tool efficacy from the expert's perspective. The helped us to deploy E-discovery solutions in the organisation's platform, which aids in finding interesting patterns relevant to the investigation (discussed in Chapter 5 and in the Section 6.1.1). We learnt just talking with the analysts/experts is not enough, involving them actively within a participatory design (also called "co-design") process [146] throughout the design-cycle from capturing design requirements to building solutions and deploying them. However, we could have provided an effective and efficient solution by providing various strategies and guidelines for carrying out an in-depth investigation based on a structured co-design process.

F2. Multi-faceted exploration and multi-granular analysis helped in discovering interestingness: From the initial user requirements, based on the research question, to identify interesting samples where the notion of interesting was vague and not defined was a challenge. The information search itself was challenging due to lack of established pointers for starting an investigation, as the need to engage with information comes in various channels (time, communication, text). The question such as "How to drill-down and establish a relationship with the E-mail data to support E-discovery analysts?" made us introduce a multi-faceted and multi-granular approach that can support in exploration and finding interesting connections within the data (discussed in Chapter 5). While designing and developing a multi-faceted and multi-granular interactive visualisation, during the iterative user-centered design process, we observed how the multiple facets such as time, people, threads and content can be inter-linked to support browsing and perceiving

information which is needed for any investigation. The latest email investigative solution, Beagle, developed by Koven et al. [111], is a multi-faceted and multi-granular linked visualisations to find information relevant to the investigation. We also understood granularities must be inter-linked, so there is a smooth high-level to low-level flow (for example, time to be visualised in a granular form - years, months, weeks, days and hours). By capturing the design requirements from the interviews with the experts, we were able to build conventional visulisations (considering multi-facetedness and multi-granular) that can support analysts to explore and analyse temporal patterns and analyse individuals' communication patterns. Specific to thread analysis, representing the evolution of communication was a challenge, so we had to develop novel visualisations (considering multi-facetedness and multi-granular) that can help in representing individuals who are included/excluded constantly, passive/active during the continuous conversation and who are secretly copied (bcc). In the user study with the experts, the expert (E1) feels there are many types of features and granularities and he was able to understand unusual changes or patterns. The expert (E1) commented that "the interactive interface assists in understanding, exploring and comparing various threads and patterns of interest. It is helps in seeing different thread communication types". Furthermore, the expert (E1) appreciated the idea of having multiple visualisation views such as thread features, overview of threads selected and a detailed single thread view that helps in understanding the types of threads. The expert said the tool's ability to show a visual overview of all the thread features (multi-stacked scatter plots), thread comparison (bars in the thread overview), individual thread view (dots representing senders, recipients and the types with sorting feature), it's conversation and allowing the user to explore and navigate through facets. The expert said, "the tool is interesting, very efficient and comprehensive". Also, while further exploring, the expert (E2)said "I select participant growth because more and more people are being involved and topic of discussion might be of interest". The expert checked the "Thread Overview", "Message" view and found Sheila.Nancey as a person of interest and identified the whole thread communication as business chat. The expert found changes in the conversation which are unusual. He further continued to compare multiple threads to understand

Sheila's behaviour. He made good use of all the interactive visualisations except feature projection but he used feature ranges to look for interesting cases. This helps establish the email visualisation using multi-faceted exploration and multi-granular analysis helped in in exploring, comparing and discovering various patterns (interestingness) in an effective and comprehensive way. Based on the validation with the experts (Appendix A.7) and the evidence of success (solutions deployed), we understand the importance of multi-faceted exploration with multi-granular analysis that helps in drilling-down to an information of interest and helps in finding the connections within the data. However, we can further improvise visualisation by introducing comparison techniques [104][105] that can explicitly depict the relations between multiple granulars and multiple facets of the threads with the related message/topic, individuals and their connection/network.

F3: Interactive visualisation assisted active learning helped in classifying communications: The most interesting idea we encountered in this research is the use of inter-active visualisation assisted active learning where E-discovery experts can categorise/classify complex communication patterns in the E-mail data (with threads) without any prior knowledge of the underlying categories. We followed Hoferlin's [93] work, an inter-active learning approach, which extends active learning by integrating users' expertise for posing queries of data instances for labelling, annotating manually/automatically and adjusting complex classifier models. This helps in the detection and correction of inconsistencies between the classifier model trained by examples and the user's mental model of the class definition. We followed the same approach as discussed in Appendix A.3. The question such as "How emails can be classified as short discussions, information leaks or announcements to support E-discovery analysts?" made us introduce a number of novel visual representations of email threads along with data features that were engineered through a human-centric approach through an iterative user-centered design process (discussed in Chapter 5). We demonstrated the validity of our approach through a case study conducted together with experts on email thread classification (notes recorded in the Appendix A.7). In the user study, we observed that experts were able to build a clustering model with various labels

such as "Business-Chat", "Legal-trouble" and so on in a short session of less than 5 minutes (Appendix A.7). The expert (E2) made good use of the active learning and the recommended samples. The expert worked on the samples very closely, and even continued all were exhausted (Fig. 5.29(f)). The expert (E2) commented that "This was one of the first 'visual expert generated' classification model they were able to build" (Appendix A.7). The expert mentioned the active learning helped him to a greater extent to mark many of the threads. He mentioned "the visual-assisted active learning can help in finding interesting behaviours quickly which will help in taking proactive decision" (Appendix A.7). Based on the case study conducted with the experts, we observed the act of annotating interesting threads enabled the analysts for further deep investigation on particular individuals, their connects and the content exchanged. We observed and found that our framework help experts in inferring relevant categories of communication, and in turning their observations into effective models that can then be used to classify threads at scale. Based on the validation with the experts (Appendix A.7) and the evidence of success (solutions deployed), we understand the importance of interactive visualisation assisted active learning that helps in filtering non-interesting/uninteresting subsets of data and helps in focusing on areas that are likely to be more interesting. To the best of our knowledge, we know this is an under-explored direction for using interactive visualisation assisted active learning for classifying email threads. Further studies can be conducted to improve accuracy and efficiency [98] of the classification and also its capabilities.

6.1.4 Learnings

During the three years of the design study, we reflect on the entire research process, from requirement analysis, design process to evaluation. We learned many lessons that can be potentially useful to the visual analysis / visualisation community. These lessons come from both struggles and success we faced while developing the visualisations for the E-discovery and Digital Forensics as well as comments received from our collaborators. In addition to the three learnings from the design and validation phases (L_{dv1}, L_{dv2} and L_{dv3}), we have three more lessons learnt (**L1, L2** and **L3**). Each of the lessons learnt are mapped to the

objectives and they are as follows.

L1. Domain-specific requirements in depth should be focussed. The iterative interviews conducted with the experts helped us understand the domain problem and its challenges, which helped us come up with design requirements, analysis goals and tasks to find interestingness in the data. The domain experts contributed actively in discussions. The iterative interviews helped the domain experts understand how visualisation solutions can add value to their existing workflow. For example, one of the design requirements (R9) question "Who are the senders and receivers in a thread who are secretly added as BCC? when are they excluded and included back?". The iterative interviews while designing prototypes using three stages, that is low-medium-high fidelity stages (paper designs, tableau designs and web-based designs), helped us understand the E-discovery domain, also get more valuable suggestions and requirements (O1) to develop an industry deployable interactive visualisation solution which can help in finding or discovering interesting individuals who are secretly copied (bcc), included and excluded in a conversation within a thread. We observed conducting regular interviews with the experts was effective for the purpose of understanding the high-level domain problems, their current challenges, design issues, which helped us address our research question. However, one of the first lessons we learnt during end of the design study was to focus deeply on the domain-specific problems and challenges, understand both high-level and low-level issues and not restrict only to problems/challenges where visualisation/visual analysis could be used. A recent paper has conducted a systematic literature review on quality criteria for conducting user requirements [85], which will help in requirements collected to be complete, unambiguous, specific, time-bounded, consistent, etc. In our work, we filtered several challenges based on relevancy where visualisation techniques could be implemented immediately (discussed in the Appendix A.7). For example, challenges such as "effective search techniques to find an important information that can improve accuracy", "effective strategies to find pertinence that can minimise cost and time" should have been focussed in conjecture with the main challenges identified in this thesis (**C1-C3**). Jiang et al. [98] studied a special

paradigm of active learning, called cost effective active search, which helps in minimising time and labeling cost by using Bayesian approaches. We should have investigated further how our designs/solutions could have supported "*cost effectiveness*" to current email cases using latest techniques. In order to improve focusing on domain-specific requirements, we can also use a qualitative data analysis (QDA) computer software package, NVivo [1] for analysing the data captured and support in iterative interviews which can be effective. The QDA software [181] can help in searching for an accurate and transparent picture of the data which will help in focussing on all the relevant information. In this way, we reflect on our first objective (O1) and conclude saying, we should have focussed more on the domain-specific problems and requirements to address effectiveness.

L2. Conventional visualisations and novel visualisations should be carefully analysed. One of the major mistakes we made during the early part of our project was to insist on building novel visualisation(s) to address the initial requirements captured in the interviews and contribute novelty to visualisation domain, which consumed lot of our time. Also, we lost time in thinking how novel visualisations can contribute to the E-discovery domain. Later, we realised focusing on conventional visualisations that solves the problems rather novel visualisations is the best way to provide a good usable solution. And novelty has many facets - novelty in visualisation is not only merely as visual encoding of information but more holistically as how various parts of visual communication of information complement one another, different interaction techniques are used, how the data is transformed, derived and at times re-purposed where also analytical/statistical operations are considered to support the data to be visually displayed. To a greater extent, for the design and implementation, following low-fidelity to high-fidelity prototyping helped in building solutions quickly based on the requirements and we were able to deploy the solutions on their platform (mentioned in the Appendix A.7). For example, to address the research question, in the design & validation phase 1 (temporal analysis), based on the design requirements (R1-R3) and tasks (T1-T3), we considered conventional heat matrix

[1] https://www.qsrinternational.com/nvivo/home

charts and bar graphs to support in exploration and discovering interesting patterns. In this phase, we did not consider any novel visualisations, interaction techniques or data operations. We designed a conventional pattern-oriented interactive visualisation that can provide various multi-facets such as temporal information, individuals and the messages exchanged in easy navigation and browsing. We used conventional filtering techniques to drill-down and roll-up to find more subsets of data that might help in discovery and finding patterns of interest (O2). The expert (E1) liked the idea of using small multiples to visualise relationship between multiple granularities (years, months, days, days of the week) that helps in finding interestingness in the data (mentioned in the Appendix A.7). Similarly, in the design & validation phase 2 (individuals analysis), based on the design requirements (R4-R6) and tasks (T4-T6), we developed a conventional pattern-oriented interactive visualisation that can support in exploration and aid in discovering interesting patterns (O2). The expert (E1) liked the idea of using a bubble matrix to visualise relationship between the selected individual and their connections, their number of occurrences that helps in finding interestingness in the data (mentioned in the Appendix A.7). Interestingly, in the design & validation phase 3 (thread analysis), we included thread ids to the existing email data (based on reverse engineering). Though we considered conventional visualisations (scatterplots, line-dot graphs), based on the design requirements (R7-R10) and tasks (T7-T10) we had to consider feature engineering (to represent thread features), active learning (to support in labeling and classifying communication types in threads) and design a novel visualisation to represent a single thread which can characterise active/passive individuals, constantly included/excluded individuals during a conversation and secretly copied (bcc) individuals that can help in discovery and finding patterns of interest (O2). The expert (E1) liked the thread features which helps in drilling down the data of interest and observe the individuals in the single thread novel visualisation. The expert also liked the idea of labelling the communication type based on one's interest which helps in classifying the complete data (mentioned in the Appendix A.7). All the interactive visualisation solutions in all the three phases were validated, adopted and deployed by our collaborative partners. So, we argue conventional visualisations could be

used where necessary to address a particular problem (requirements) based on the data type, problems/tasks. In our view, many of the challenges or design requirements can be accomplished using conventional visualisation techniques with conventional interaction facilities. In many other challenges, data has to be engineered and visually represented in a novel way to address a particular problem/requirement. Irrespective of considering conventional or novel visualisation, we must be using a specific methodology to sketch paper designs during the initial design phase will lead to better direction in building an effective visualisation tool. Though we considered design sketching as part of the design process, we did not follow a specific methodology. Through our learning, we understood considering Five Design Sheet (FdS) methodology [145] will enable us to decide whether to select conventional or novel visualisation. In this way, we reflect on our second objective (O2) and conclude saying, the designers must carefully contemplate whether to consider conventional or novel visualisation for their research.

L3. Evaluation with non-experts should be considered: The personal validation helped in understanding the use of our visualisation to compare multiple time periods / temporal dimensions, individual information, thread information to summarise the communication patterns across all the facets to infer patterns and trends. As discussed in Chapter 2, one of the advantages of first carrying out personal validation followed with a quick use case walkthrough with the experts saves time in conducting empirical/user studies. So, personally validating a tool, reporting on findings, then quickly walking through a use case with experts and then deploying a solution immediately to check for engineering issues saves a lot of time and effort. This approach allowed us to continue with the subsequent phases of the design process. The tasks helped to determine the potential effectiveness of our techniques in email investigations. Our validations were discussed with the experts and they explored the tool to understand the same. Though personal validation was successful, it is tough as making sense of your own sense-making is fraught with difficulty. However, we conducted an empirical study with the experts. This helped us understand how our solution can help in their E-discovery requirements to find interestingness in the data. We also

observed, the experts did not feel pressurised, they were confident in using the tool, they were satisfied with its features and they started planning to deploy the solution in their organisation platform. The experts contributed actively in discussions and also in suggesting various improvements in the tool (mentioned in the limitations). We observed conducting the study with the experts was effective for the purpose of using the solution/interface in their organisation. Through this evaluation, we have reflected on our third objective (O3), we have also addressed our research question and helps us establish that interactive visualisation has helped experts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation to support in the E-discovery Investigation. However, we also learnt evaluating with non-experts would add more value if the solutions are planned to be deployed in various other organisations, especially individuals from compliance and auditing teams have limited technical/visualisation knowledge. Through the process, we learnt conducting empirical studies with the non-experts will help in validating interactive visualisation-assited active learning solutions more effectively [185].

6.1.5 Principles

Informed and motivated by the lessons learned and findings, we propose the following general principles as the basis of our approach that can support other researchers who are interested to design and develop interactive visualisations to explore and find/discover relevant/interesting information in the digital communication data from an investigation perspective.

P1. Focus on domain-specific requirements. The iterative interviews and collaborative design phase helped us understand design requirements and we deployed solutions in the organisation's platform. We should focus deeply on the domain-specific problems and challenges, understand both high-level and low-level issues and not restrict only to problems/challenges where visualisation/visual analysis could be used. In this way, we can deploy an effective and efficient solution by providing various strategies and guidelines for carrying out an in-depth investigation (**based on L1 and F1**).

P2. Consider iterative user-centric design throughout design-cycle. Even though we introduce a generalisable, domain-agnostic analysis framework, our approach also builds significantly on an iterative user-centric design phase that recognises the specifics of the data and the problem domain throughout the design-cycle. Using real users, case studies, design requirements and tasks helped us in understanding how visualisation can help in identifying interesting time points, individuals, threads and their activities which involves messages sent/received, especially investigating communication data to support investigation cases (**based on L1 and F1**).

P3. Generate system-based features for pattern characterisation. Due to the non-numeric nature of the E-mail data, one needs to make use of system generated measures that part-explain the characteristics of the patterns. One central mechanism we leverage in our framework is to support analysts in their investigation and specification of underlying patterns through custom-build visual representations and semantically relevant heuristic features. Based on the design requirements, analysis goals, tasks and iterative design interviews, we decided to build a novel visualisation that can represent individuals who are active/passive and included/excluded in the threads (conversations) which can be facilitated through a selection of system generated features. So, generating system-based features for pattern characterisation aids in investigations (**based on L2 and F2**).

P4. Build pattern-oriented interactive visualisations for discovering interestingness. For discovering and finding interesting patterns in the data, based on the interviews and analysis, we can build conventional or novel visualisation. For example, to help support the specification of the thread patterns, novel visual representations helps to reveal the structure and characteristics of a single thread. The interactive visual interface must display a comprehensive set of user/system generated metadata (e.g., thread length, sender diversity etc.). The analysts must be able to understand different characteristics of the patterns from various perspectives that can then be translated into a formal description for investigations (**based on L2 and F2**).

P5. Leverage multi-facetedness and multi-granularity for exploration and discovery. Considering the exploratory nature of email/complex pattern analysis, the inter-

face must provide various facets (e.g., temporal, individuals, thread information etc.) as a means for navigation and browsing. If the multiple facets are inter-linked, browsing and perceiving information becomes easy. If the facets are effectively presented, analysts will be able to explore complex patterns, by quickly navigating through a particular facet. For analysts to identify interesting points and seamlessly switch between the different levels of overview and the detailed view (multi-granularity), integrating the high level view of the email features, the visual overview of the email communication (for comparing all the emails selected), and the detailed view is the key thing in investigations (**based on L2 and F2**).

P6. Important to represent evolution of communication. The chronology/time-ordering of the communication sequence for each of the thread helps in understanding the communication type. So, representing the arrival sequence of messages which create a thread helps in understanding the evolution of a thread and the relationships of the individuals. These relationships give important contexts for each message in a thread, their evolution, time-periods and the individuals involved that are relevant for an investigation. So, it is important to represent evolution of communication for investigation cases (**based on the scalability limitation**).

P7. Evaluate the system with both experts and non-experts. conducting evaluation studies with both experts and non-experts (have limited technical/visualisation knowledge) will help in understanding how visualisation can help in solving E-discovery problems by finding interesting information and/or relevant information in the data for any domain (**based on L3**).

6.2 Conclusion

The main aim of this research was to investigate critical aspects of E-mail communication within an organisation compliance by designing and developing interactive visual solutions to support E-discovery analysts. More specifically, this research aimed to answer the question *"To what extent visualisation can support analysts in finding relevant and/or*

discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation?".

To address the above research question, we adopted the Design Study Methodology (DSM) [156, 61] which provides a methodological framework and practical guidance for conducting a design study and research. We explained all the three phases (nine stages) of the DSM used in our work. Each stage in this nested form helps in analysing the problem and validate the solution independently. Figure. 6.1 represents how we made use of the three phases in this study, addressed our questions and contributed.

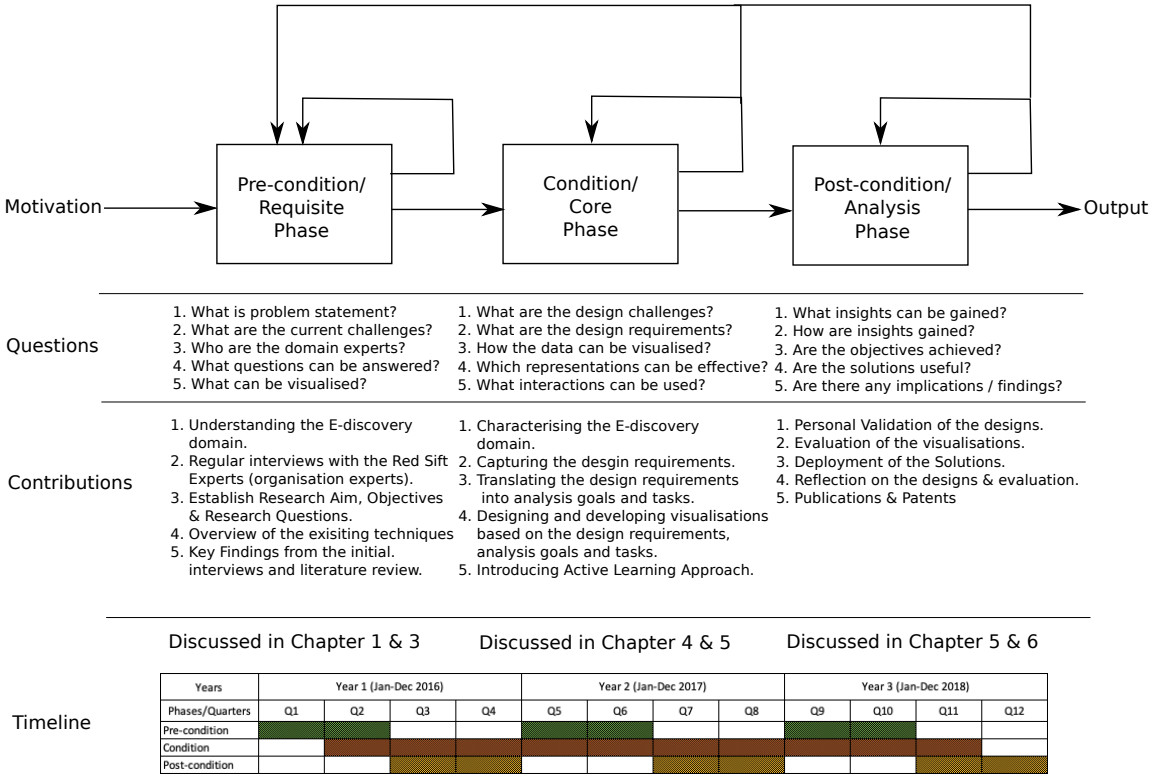


Figure 6.1: The three main phases of our design study in this project (adapted from Design Study Methodology (DSM) [156] are Pre-condition, Condition and Post-condition phase which had various questions to be addressed. All the contributions are mentioned in each of the phases with a timeline.

In the Pre-condition phase, we identified the domain as Digital Forensics & E-discovery

and the users as organisational experts from Red Sift (cybersecurity company). We report on an 36-month long design study conducted as a multidisciplinary team of visualisation researchers and E-discovery experts. We conducted a series of bi-weekly sessions with two experts/analysts from a cybersecurity company who are familiar with the type of investigation carried out in E-discovery, where we take a user-centric approach in designing visualisations. In the field of Digital Forensics & E-discovery, organisations play an important role to take on legal actions against individuals working in their organisation or legal actions against other organisations based on the communicated E-mails [143]. In E-discovery litigations or investigations, practically every analyst/investigator finds a vast and semantically meaningful collection of data, that is uncategorised and unlabelled, in E-mail inboxes to investigate which makes it tedious and time-consuming to classify, identify and/or discover various information [30], eventually makes it expensive for an organisation. So, this phase helped us to understand the domain, high-level problems and establish research aim and objectives.

In the Condition phase, we characterised the domain, captured the design challenges, requirements and tasks (**O1**). To support organisations in the E-discovery process, we designed interactive visualisation prototypes to explore and find/identify/discover interesting patterns/information (which could be temporal, individuals, threads). We considered three levels of fidelity for designing prototypes: low, medium and high. This allowed us to quickly throw away designs/codes and spend more time on analysing the design challenges and requirements, as we moved from low to high fidelity. We presented our designs (**O2**) through an applied case of discovering interesting E-mail communication patterns. We demonstrated the validity of our approach through a real case study on email communication that is conducted together with experts. Each of the interview session lasted around one to two hours where in the earlier meetings the focus was on the requirements which later shifted towards discussions on the different design alternatives. In total, we had three design phases (DV Phase 1 - DV Phase 3), addressing all the ten requirements (**R1-R10**). Each phase was built on the other. There was a level of complexity relationship between the three of these (increasing) and we learnt from one and move to the other. The

first two design phases were related to temporal and individual analysis which helped in understanding various patterns of interest and led us to the third design phase considering complete threads which involved time and individuals to find interestingness. In this third design phase, we presented an integrated approach starting from interactive visual exploration of threads to understand threads and creating various classes efficiently for investigation purposes. During the exploratory phase analysts were supported by highly interactive visual means to explore various thread features in order to design classifiers based on the domain and task specific needs of the analysts. Our approach incorporated novel methods that take advantage of thread features, with interactive visualisation that supports multifaceted exploration and classifying communication types. The experts' feedback from our evaluation (**O3**) suggested that our tool can help the analysts to identify threads of interest, find individuals of interest, discover their activities and classify them accordingly. So, based on the design phases, validation, empirical evaluation and experts' feedback, we argue that our visualisation can support analysts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation within well-understood problems. However, there are several other problems/challenges (discussed in the reflection section, 6.1), such as visually analysing topics within threads to quickly understand the co-relation between the threads, individuals and the topics exchanged within a particular time which could add value to the investigation. We also identified several limitations (discussed in the reflection section, 6.1), such as scalability, improvised visual interactive active learning approaches etc to quickly discover anomalies or interesting information relevant to an investigation. So, this phase helped us learn how to successfully iterate over designs and build a solution that can support analysts in investigation.

In the Post-condition phase, we reflected on the three objectives (**O1-O3**) mentioned in the Chapter 1, to reflect on our work, gain insights, identify our lessons learnt and also identify positives, limitations, and direction for future work. We report a number of lessons learnt - as mentioned in **L1**, domain-specific requirements needs to be focussed in depth. In this way, we could have identified more complicated challenges that might require

innovative visualisation techniques. Also in **L2**, we mentioned to focus on conventional visualisations over novelty. Such that, we respond with effective designs based on the nature of the problem we are addressing. For example, we considered conventional pattern-oriented visualisation (matrices, node-links, bar charts etc.) for finding interestingness in the data. Similarly, we identified the limitations (scalability, clutter, interpretability, comparability, etc) which can be addressed using novel visual encoding of information or analytical/statistical operations. The third lesson (**L3**) was about considering non-experts for evaluating solutions, which will add value for deploying solutions in various domains. So, this phase helped us learn how to reflect on our work, find positives, identify limitations and how we can improvise our work to support analysts in investigation.

As discussed in Chapter 1, we call our solutions as "data visualisation as evidence" which are interactive visualisations built using data. In an E-discovery case, an organisation lawyer who wants to win a legal case against the opposition organisation lawyer, data visualisation as evidence can be demonstrated to defend the case, as visuals have the potential to convey more complex meanings and often represent concepts that are challenging to express. A jury that can visualise the case while the expert is testifying will be much more likely to comprehend the occurrence and give a fair judgement. In this way, the whole process will be made proactive, preventive and/or support legal evidences [63], [120]. Through our design process and deployment, we have answered our research question - *"The interactive visualisation solutions can support analysts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation".*

6.2.1 Future Work

We envision extending our thread-based email visualisation in several ways that can support analysts in developing strategies to create models and classes effectively for investigation purposes. To improvise on the visualisation that can explicitly depict the relations between multiple facets of the threads with the related message/topic, individuals and their connection/network. We plan to devise an interactive model-driven visualisation technique [82]

that allows analysts to dynamically change the facets of interest and reveal relations between them. On visual scalability, we are planning to work on the data engineering process and introduce additional clutter reduction techniques [66] that are needed to represent higher number of threads. Merging both visualisation and data aggregation techniques might aid in producing effective results while exploring the data and extracting deeper insights [66].

In the desing & validation phase 3 (thread-based analysis), we have demonstrated visual interactive active learning in the context of email communication pattern analysis; it remains an open question how successfully it can be applied to other domains. For example, we can consider using it in other communication analysis such as social network analysis (Twitter tweet-threads, Facebook post-threads, Slack message analysis, etc.) that contains similar multi-faceted features such as time, individuals, threads and content. Most of the tasks (T1-T10) can be tested if our solutions can help in finding interestingness in the data. It will be intriguing to see how our solutions can be applied for social network analysis. For applying it in other domains successfuly, we would also require a more comprehensive evaluation of active learning [36] considering scalability [158][66] (as discussed in the limitation). So, we would also like to design and integrate more advanced methods on active learning techniques [37] and conduct a formal user study to further evaluate the usability of our system using different data types and tasks [36],[163]. In addition to these, we will extend our multi-class classification to support comparison of data instances from multiple classes using smart labeling approach [74]. We are also planning to introduce effective comparative techniques [105] for investigators to compare subsets of data to find more interesting changes/behaviours.

Though we have message box integrated with the other facets (time, individuals, threads), we did not consider topic analysis as part of the design studies due to the initial challenges we faced in designing and building the tool. The text in the email data is rich and likely to contain a lot of information that can help on investigate. As a side track, for our investigation purpose, we worked on topic analysis and modelling (screenshots attached in the Appendix A.4) but it is currently not deployed in the organisation's platform.

Introducing an effective and effecient topic modelling & visualisation would add a lot of value to the investigation - we can extract topics/keywords/sentiments and understand the co-relation between the threads, individuals and the topics exchanged. A change in these features may imply an anomalous behaviour or something is interesting. The latest Beacon system [142] uses text mining methods with social network analysis in the emails communicated. It extracts information about messages sent, subjects, topics and phrases. We can improvise this technique to understand sudden changes in the topics. For example, talking about an organisation's confidential policy (sudden change of topics) to an external party. Also, to find more interesting cases related to the emails exchanged and individuals involved in the emails, it might be worth adding a text filter to find specific words of interest. For example, if we search the word "finance", the tool must visualise only emails which contain it. In this way, we can find and related emails with similar topics. This will link all the threads related to the same topic/subject. We also aim to re-access the following in the future:

1. The domain challenges, tasks and requirements (interview insights) will be re-accessed to improve the prototype.

2. The design challenges, strategies, framework and visualisations will be re-accessed to improve the prototype that can support scalability, clutter, interpretability and comparability.

3. The results of the empirical studies will be re-accessed to improve the prototype and provide a concrete guidelines/recommendations for analysts to carry our investigations effectively and efficiently.

Along with our company partners Red Sift, we intend to open source parts of the output of this project and by providing the parts as an extension for GMAIL and for email services at the City, University of London (serve as an academic research tool). This will result in deploying smart apps (Sifts) to work on one's data to automate workflows and surface actionable insights. This will also result in new products originating from researchers and that it will significantly shorten the current gap between academic research

and commercial deployment within the domain of email data analysis. The academic validation and publicity of the platform will help in further development of the tool.

6.2.2 Impact

Contribution to the Industrial Partner: The tools currently available in the market are based on simple keyword search and legal firms charge companies based on the volume of information produced by the search, which is then manually reviewed [63]. This results in significant costs for the company or in a number of cases settlement because they can't afford the costs of E-discovery in digital communication [63]. We collaborated with Red Sift, a cyber security company in London that has a unique vision for the future of E-discovery. Our novel visual design in the Thread-based visualisation (single thread view) to visualise individuals who are active/passive or individuals included/excluded during a conversation will support in E-discovery investigations. The expert (E1) from Red Sift said, "the tool is interesting, very efficient and comprehensive". This helps establish our email visualisation helps in exploring and comparing various threads in an effective and comprehensive way to discover interesting information. The experts are keen to combine their platform (streaming processing) with static and dynamic human data interactive visualisations, which aims to revolutionise the industry - allowing companies to adopt proactive, preventive, more accurate and ultimately more cost-effective E-discovery procedures. The company also intends to open source parts of the output of this project and by providing the parts as an extension for GMAIL and for email services at various academic institutions. The work can be further extended by integrating latest innovations that can improve investigation. For example, Gupta et al. [83] has patented a work on "Visual representation of an email chain" in 2018 and Bai et al. [31] has patented a work on "Analysing email threads" in 2019. We will be filing a patent on our work "Email Thread Analysis using Interactive Visualisation Assisted Active Learning" in 2020, which will be another contribution to our industrial partner.

Contribution to Digital Forensics & E-discovery: Using visual analysis in Digital Forensics and E-discovery compliance for E-mail communication within an organisation

is still an under-researched topic, this research aims to develop visual solutions that can help in the investigation/analysis for analysts to come to a conclusion and build a legal case, in a way making the whole process proactive and preventive [63, 120, 3]. We closely worked with a team of E-discovery experts in the UK (**O1**). Our solution that helps in classifying emails based on the communication (using active learning approach) is a contribution that can support in E-discovery to discover interesting information. Based on the experts' opinion, our solutions will help in the investigation/analysis by facilitating in the generation of visual evidence to come to a conclusion and build a legal case, in a way making the whole process proactive and preventive. The expert (E2) said "*the interface is good and helps in navigation & exploration. Easy to identify and classify the threads based on the conversations or emails exchanged. The complete pipeline of the workflow is good and it is definitely useful for investigation and we will use it as a solution for our products planned*". Our solutions in information discovery and decision making within an investigation domain was not fully investigated in the past thus an innovation that will be an useful contribution.

Contribution to HCI and Visualisation Domain: As discussed earlier, HCI clearly overlaps with Visualisation, particularly in topics relating to data analysis. Many of the concepts and datasets, especially in the area of investigation, motivates the need for HCI for interaction with data and the algorithms processing them. The work identified the knowledge gap, challenges, requirements and tasks in Digital Forensics and E-discovery involving the analysis of E-mail communication data from the unstructured interviews with the organisation domain experts and from the literature. We considered an iterative user-centered design (UCD) approach which involved experts in the complete design-cycle for 3 years and we built several visual solutions based on the requirements and tasks (from low-fidelity to high-fidelity designs). One of our main contribution is the interactive visualisation solution in combination with active learning which can support investigators. We call it as "Interactive Visualisation-assisted Active Learning", which can be a new domain and problem in the field of Visual Analytics and HCI [118]. We evaluated the interactive visualisation-assisted active learning solution by conducting an empirical study with

experts to understand E-discovery tasks, visual solutions and the interface that can help analyst, to investigate and navigate within communication data, to identify/find/discover various patterns, trends, anomalies and information relevant to investigation. A recent PhD thesis by Linder [123] has proposed a new approach "Grounded Visual Analytics" that integrates Information Visualisation and interactive machine learning, and Grounded Theory that can support investigators to discover patterns in the data. We know these kind of new approaches will arise in the area of E-mail analysis, with resultant effects for how analysts/users should interact and interpret these data. Recent interest in the "Quantified Self" also relies on data collation and analytics about email activities of an individual. Through our methods and solutions, researchers will be able to gather insight on individuals' interactions with their E-mail. Such valuable information could help researchers build more effective personalised inboxes and could inform the design of email management software, may lead to more effective solutions to could help people with this time-consuming activity. Since our research has considered the intersection of two broad areas - Visualisation and Human-Computer Interaction, we know the results will yield more flexible and scalable solutions in the future.

6.3 Closing Remarks

As the quantity and complexity of email information continues to grow, new challenges and demands for developing effective interactive visualisation techniques to explore and discover interesting information in the data grows. As discussed earlier, worldwide E-mail use continues to grow at a healthy pace, at an average annual rate of 4% over the next four years, reaching over 347 billion by the end of 2023 [1]. So, more the emails more are the complexity to identify/discover various information and relationships within the data. Hence there is a need of simplifying the investigation process by providing visual solutions [63], which is undoubtedly critical. This thesis investigated to what extent visualisation can support analysts in finding relevant and/or discovering interesting information in a corpus of E-mail within an organisation supporting in the E-discovery Investigation. We characterised the

E-discovery domain, understood the visualisation principles and techniques, design and developed interactive visualisation solutions, validated them constantly and deployed the solutions in the organisation's platform. Through the evidence of success in this research, we argue visualisations (with active learning) combined with Human-Computer Interaction can be a vehicle for E-discovery. This intersection of two broad areas - Visualisation and HCI, indicates that the bond of the two fields will be even closer in the future to yield more flexible and scalable solutions in an interdisciplinary approach. Our investigations have shown that visualisation, applied through design study can significantly facilitate discovery and insight in the exploration of email communication for E-discovery to the extent that our partners are now implementing the techniques that have been established to do just this in a commercial context. However, our work is only a start of the emerging research of developing interactive visualisations to assist investigators with making sense of the email data. When we consider streaming big data, there is a huge potential and big opportunities in this research space.

Bibliography

[1] https://www.radicati.com/wp/wp-content/uploads/2018/12/Email-Statistics-Report-2019-2023-Executive-Summary.pdf.

[2] https://gdpr.eu/email-encryption/.

[3] http://www.legaltechnology.com/latest-news/guest-comment-visualisation-technology-ligh

[4] http://enterprise.brainspace.com/discovery.

[5] http://www.lexisnexis.com/litigation/products/ediscovery/concordance.

[6] http://in-spire.pnnl.gov/.

[7] http://www.ftitechnology.com/radiance-visual-analytics-software.

[8] http://zovy.com/solutions/ediscovery/.

[9] https://www.nngroup.com/articles/goal-composition/.

[10] http://EnronData.org.

[11] http://cs.ubc.ca/labs/lci/bc3.html.

[12] http://w3c.org.

[13] http://www.ins.cwi.nl/projects/trec-ent/wiki/index.php/Main_Page.

[14] http://verbs.colorado.edu/enronsent/.

[15] https://www.kaggle.com/kaggle/hillary-clinton-emails.

[16] https://snap.stanford.edu/data/email-EuAll.html.

[17] http://www.cs.jhu.edu/\~mdredze/code.php.

[18] http://www.cs.cmu.edu/\~einat/datasets.html.

[19] http://khorshid.ece.ut.ac.ir/\~m.dehghani/emaildataset.html.

[20] https://catalog.ldc.upenn.edu/LDC2015T03.

[21] http://fcir.org/2014/12/29/search-jeb-bush-email/.

[22] http://www.politifact.com/florida/statements/2015/aug/31/jeb-bush/jeb-bush-says-he-has-released-all-his-emails/.

[23] http://www.dfki.de/~neumann/resources/omqdata.html.

[24] http://csmining.org/index.php/spam-email-datasets-.html.

[25] http://www.dit.ie/computing/staff/sjdelany/datasets/.

[26] http://bailando.sims.berkeley.edu/enron_email.html.

[27] Tableau software, inc. https://www.tableausoftware.com/, 2011.

[28] Daniel Archambault, Derek Greene, Pádraig Cunningham, and Neil Hurley. Themecrowds: Multiresolution summaries of twitter usage. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, pages 77–84. ACM, 2011.

[29] Daniel Archambault, Helen Purchase, and Bruno Pinaud. Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):539–552, 2011.

[30] Simon Attfield and Ann Blandford. Discovery-led refinement in e-discovery investigations: sensemaking, cognitive ergonomics and system design. *Artificial Intelligence and Law*, 18(4):387–412, 2010.

[31] Song Bai, Ming Qun Chi, Hui Huang, Hui Liu, Xiang Xing Shi, and Ang Yi. Analyzing email threads, May 21 2019. US Patent 10,298,531.

[32] Fabian Beck, Michael Burch, and Daniel Weiskopf. A matrix-based visual comparison of time series sports data. 2016.

[33] Richard A Becker, William S Cleveland, and Ming-Jen Shyu. The visual design and control of trellis display. *Journal of computational and Graphical Statistics*, 5(2):123–155, 1996.

[34] R Beecham, C Rooney, S Meier, J Dykes, A Slingsby, C Turkay, J Wood, and BLW Wong. Faceted views of varying emphasis (favves): a framework for visualising multi-perspective small multiples. In *Computer Graphics Forum*, volume 35, pages 241–249. Wiley Online Library, 2016.

[35] Melvin M Belli Sr. Demonstrative evidence. *Wyo. LJ*, 10:15, 1955.

[36] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics*, 24(1):298–308, 2018.

[37] Jürgen Bernard, Matthias Zeppelzauer, Markus Lehmann, Martin Müller, and Michael Sedlmair. Towards user-centered active learning algorithms. In *Computer Graphics Forum*, volume 37, pages 121–132. Wiley Online Library, 2018.

[38] Jürgen Bernard, Matthias Zeppelzauer, Michael Sedlmair, and Wolfgang Aigner. Vial: a unified process for visual interactive labeling. *The Visual Computer*, 34(9):1189–1207, 2018.

[39] Jacques Bertin. Semiology of graphics: diagrams, networks, maps. 1983.

[40] Hugh Beyer and Karen Holtzblatt. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[41] Mike Bostoc. Data-driven documents. `https://d3js.org/`, 2011.

[42] Michael Boyle and Saul Greenberg. Rapidly prototyping multimedia groupware. 2005.

[43] Ulrik Brandes and Bobo Nick. Asymmetric relations in longitudinal social networks. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2283–2290, 2011.

[44] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[45] Virginia Braun and Victoria Clarke. *Successful qualitative research: A practical guide for beginners*. sage, 2013.

[46] Matthew Brehmer, Stephen Ingram, Jonathan Stray, and Tamara Munzner. Overview: The design, adoption, and analysis of a visual document mining tool for investigative journalists. *IEEE transactions on visualization and computer graphics*, 20(12):2271–2280, 2014.

[47] Matthew Brehmer and Tamara Munzner. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19(12):2376–2385, 2013.

[48] Michael Brooks, Saleema Amershi, Bongshin Lee, Steven M Drucker, Ashish Kapoor, and Patrice Simard. Featureinsight: Visual support for error-driven feature ideation in text classification. In *Visual Analytics Science and Technology (VAST), 2015 IEEE Conference on*, pages 105–112. IEEE, 2015.

[49] Marion Buchenau and Jane Fulton Suri. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, pages 424–433. ACM, 2000.

[50] Nan Cao, Yu-Ru Lin, Xiaohua Sun, David Lazer, Shixia Liu, and Huamin Qu. Whisper: Tracing the spatiotemporal process of information diffusion in real time. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2649–2658, 2012.

[51] Nan Cao, Lu Lu, Yu-Ru Lin, Fei Wang, and Zhen Wen. Socialhelix: visual analysis of sentiment divergence in social media. *Journal of Visualization*, 18(2):221–235, 2015.

[52] Nan Cao, Conglei Shi, Sabrina Lin, Jie Lu, Yu-Ru Lin, and Ching-Yung Lin. Targetvue: Visual analysis of anomalous user behaviors in online communication systems. *IEEE transactions on visualization and computer graphics*, 22(1):280–289, 2016.

[53] Sheelagh Carpendale. Evaluating information visualizations. *Information visualization*, pages 19–45, 2008.

[54] Eoghan Casey. *Handbook of digital forensics and investigation*. Academic Press, 2009.

[55] Eoghan Casey. *Digital evidence and computer crime: Forensic science, computers, and the internet*. Academic press, 2011.

[56] Min Chen, Luciano Floridi, and Rita Borgo. What is visualization really for? In *The Philosophy of Information Quality*, pages 75–93. Springer, 2014.

[57] Alistair Cockburn. *Agile software development: the cooperative game*. Pearson Education, 2006.

[58] Christopher Collins, Sheelagh Carpendale, and Gerald Penn. Docuburst: Visualizing document content using language structure. In *Computer graphics forum*, volume 28, pages 1039–1046. Wiley Online Library, 2009.

[59] Jack G Conrad. E-discovery revisited: A broader perspective for ir researchers. In *Proceedings of the first international workshop on supporting search and sensemaking for electronically stored information in discovery proceedings at the 11th international*

conference on artificial intelligence and law (ICAIL07 DESI Workshop, Stanford University). DESI Press, CA, 2007.

[60] Kristin A Cook and James J Thomas. Illuminating the path: The research and development agenda for visual analytics. Technical report, Pacific Northwest National Laboratory (PNNL), Richland, WA (US), 2005.

[61] Anamaria Crisan, Jennifer L. Gardy, and Tamara Munzner. On regulatory and organizational constraints in visualization design and evaluation. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, pages 1–9. ACM, 2016.

[62] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE transactions on visualization and computer graphics*, 17(12):2412–2421, 2011.

[63] D. Lawton and R. Stacey and G. Dodd. Uk home office. `https://www.gov.uk/government/publications/ediscovery-in-digital-forensic-investigations`, 2014.

[64] Laura A Dabbish and Robert E Kraut. Email overload at work: an analysis of factors associated with email strain. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 431–440. ACM, 2006.

[65] Laura A. Dabbish and Zachary Gordon Lee Wise. Seemail: Visualizing email response. 2011.

[66] Rogério Abreu de Paula. *Visualization Techniques*, pages 1775–1786. Springer International Publishing, 2019.

[67] Chris Delgado. Facing e-discovery: Preparedness and compliance. 2011.

[68] Norman K Denzin. *The research act: A theoretical introduction to sociological methods*. Routledge, 2017.

[69] Pedro Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012.

[70] Steven Dow, Blair MacIntyre, Jaemin Lee, Christopher Oezbek, Jay David Bolter, and Maribeth Gandy. Wizard of oz support throughout an iterative design process. *IEEE Pervasive Computing*, 4(4):18–26, 2005.

[71] Niklas Elmqvist, Thanh-Nghi Do, Howard Goodell, Nathalie Henry, and Jean-Daniel Fekete. Zame: Interactive large-scale graph visualization. In *Visualization Symposium, 2008. PacificVIS'08. IEEE Pacific*, pages 215–222. IEEE, 2008.

[72] Niklas Elmqvist, Andrew Vande Moere, Hans-Christian Jetter, Daniel Cernea, Harald Reiterer, and TJ Jankun-Kelly. Fluid interaction for information visualization. *Information Visualization*, 10(4):327–340, 2011.

[73] Daniel Engelberg and Ahmed Seffah. A framework for rapid mid-fidelity prototyping of web sites. In *IFIP World Computer Congress, TC 13*, pages 203–215. Springer, 2002.

[74] Xin Fan, Chenlu Li, Xiaoru Yuan, Xiaoju Dong, and Jie Liang. An interactive visual analytics approach for network anomaly detection through smart labeling. *Journal of Visualization*, 22(5):955–971, 2019.

[75] Michael Farrugia, Neil Hurley, and Aaron Quigley. Exploring temporal ego networks using small multiples and tree-ring layouts. *Proc. ACHI*, 2011:23–28, 2011.

[76] Camilla Forsell and Jimmy Johansson. An heuristic set for evaluation in information visualization. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 199–206. ACM, 2010.

[77] Simone Frau, Jonathan C Roberts, and Nadia Boukhelifa. Dynamic coordinated email visualization. 2005.

[78] Yifan Fu, Xingquan Zhu, and Bin Li. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283, 2013.

[79] Tiago Garcia, João Aires, and Daniel Gonçalves. Who have i been talking to? In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 481–484. ACM, 2012.

[80] Sohaib Ghani, Bum Chul Kwon, Seungyoon Lee, Ji Soo Yi, and Niklas Elmqvist. Visual analytics for multimodal social network analysis: A design study with social scientists. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2032–2041, 2013.

[81] Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D Hansen, and Jonathan C Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[82] Matteo Golfarelli and Stefano Rizzi. A model-driven approach to automate data visualization in big data analytics. *Information Visualization*, page 1473871619858933, 2019.

[83] Saurabh Gupta, Sandeep Perumbuduri, Nancy A Schipon, and Jack P Yapi. Visual representation of an email chain, December 4 2018. US Patent 10,147,071.

[84] Roberta Heale and Dorothy Forbes. Understanding triangulation in research. *Evidence-Based Nursing*, 16(4):98–98, 2013.

[85] Petra Heck and Andy Zaidman. A systematic literature review on quality criteria for agile requirements specifications. *Software Quality Journal*, 26(1):127–160, 2018.

[86] Jeffrey Heer and Danah Boyd. Vizster: Visualizing online social networks. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 32–39. IEEE, 2005.

[87] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *Queue*, 10(2):30, 2012.

[88] Florian Heimerl, Charles Jochim, Steffen Koch, and Thomas Ertl. Featureforge: A novel tool for visually supported feature engineering and corpus revision. *Proceedings of COLING 2012: Posters*, pages 461–470, 2012.

[89] Florian Heimerl, Steffen Koch, Harald Bosch, and Thomas Ertl. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization & Computer Graphics*, (12):2839–2848, 2012.

[90] Nathalie Henry and Jean-Daniel Fekete. Matrixexplorer: a dual-representation system to explore social networks. *IEEE transactions on visualization and computer graphics*, 12(5), 2006.

[91] Nathalie Henry and Jean-Daniel Fekete. Matlink: Enhanced matrix visualization for analyzing social networks. *Human-Computer Interaction–INTERACT 2007*, pages 288–302, 2007.

[92] Nathalie Henry, Jean-Daniel Fekete, and Michael J McGuffin. Nodetrix: a hybrid visualization of social networks. *IEEE transactions on visualization and computer graphics*, 13(6):1302–1309, 2007.

[93] Benjamin Höferlin, Rudolf Netzel, Markus Höferlin, Daniel Weiskopf, and Gunther Heidemann. Inter-active learning of ad-hoc classifiers for video visual analytics. In *Visual Analytics Science and Technology (VAST), 2012 IEEE Conference on*, pages 23–32. IEEE, 2012.

[94] Mengdie Hu, Huahai Yang, Michelle X Zhou, Liang Gou, Yunyao Li, and Eben Haber. Opinionblocks: a crowd-powered, self-improving interactive visual analytic system for understanding opinion text. In *IFIP Conference on Human-Computer Interaction*, pages 116–134. Springer, 2013.

[95] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. Technology probes: inspiring design for and with families.

In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 17–24. ACM, 2003.

[96] Petra Isenberg, Torre Zuk, Christopher Collins, and Sheelagh Carpendale. Grounded evaluation of information visualizations. In *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, page 6. ACM, 2008.

[97] Wolfgang Jentner, Dominik Sacha, Florian Stoffel, Geoffrey Ellis, Leishi Zhang, and Daniel A Keim. Making machine intelligence less scary for criminal analysts: reflections on designing a visual comparative case analysis tool. *The Visual Computer*, 34(9):1225–1241, 2018.

[98] Shali Jiang, Roman Garnett, and Benjamin Moseley. Cost effective active search. In *Advances in Neural Information Processing Systems*, pages 4881–4890, 2019.

[99] Minoo Erfani Joorabchi, Ji-Dong Yim, and Christopher D Shaw. Emailtime: Visual analytics of emails. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 233–234. IEEE, 2010.

[100] Sachindra Joshi, Danish Contractor, Kenney Ng, Prasad M Deshpande, and Thomas Hampp. Auto-grouping emails for faster e-discovery. *Proceedings of the VLDB Endowment*, 4(12):1284–1294, 2011.

[101] Hyunmo Kang, Catherine Plaisant, Bongshin Lee, and Benjamin B Bederson. Netlens: iterative exploration of content-actor network data. *Information Visualization*, 6(1):18–31, 2007.

[102] James Max Kanter and Kalyan Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on*, pages 1–10. IEEE, 2015.

[103] Holtzblatt Karen and Jones Sandra. Contextual inquiry: A participatory technique for system design. In *Participatory design*, pages 177–210. CRC Press, 2017.

[104] Johannes Kehrer and Helwig Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE transactions on visualization and computer graphics*, 19(3):495–513, 2013.

[105] Johannes Kehrer, Harald Piringer, Wolfgang Berger, and M Eduard Gröller. A model for structure-based comparison of many categories in small-multiple displays. *IEEE transactions on visualization and computer graphics*, 19(12):2287–2296, 2013.

[106] Daniel Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, and Hartmut Ziegler. Visual analytics: Scope and challenges. *Visual data mining*, pages 76–90, 2008.

[107] Daniel A Keim, Florian Mansmann, Andreas Stoffel, and Hartmut Ziegler. *Visual analytics*. Springer, 2009.

[108] Bernard Kerr. Thread arcs: An email thread visualization. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 211–218. IEEE, 2003.

[109] J Kielman, J Thomas, and R May. The future of visual analytics. *Information Visulalization*, 2009.

[110] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In *Machine learning: ECML 2004*, pages 217–226. Springer, 2004.

[111] Jay Koven, Enrico Bertini, Luke Dubois, and Nasir Memon. Invest: Intelligent visual email search and triage. *Digital Investigation*, 18:S138–S148, 2016.

[112] Benjamin L Kovitz. *Practical software requirements: a manual of content and style*. Manning Publications Co., 1998.

[113] Josua Krause, Adam Perer, and Enrico Bertini. Infuse: interactive feature selection for predictive modeling of high dimensional data. *IEEE transactions on visualization and computer graphics*, 20(12):1614–1623, 2014.

[114] Kostiantyn Kucher, Teri Schamp-Bjerede, Andreas Kerren, Carita Paradis, and Magnus Sahlgren. Visual analysis of online social media to open up the investigation of stance phenomena. *Information Visualization*, 15(2):93–116, 2016.

[115] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[116] Heidi Lam, Tamara Munzner, and Robert Kincaid. Overview use in multiple visual information resolution interfaces. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1278–1285, 2007.

[117] Craig Larman and Victor R Basili. Iterative and incremental developments. a brief history. *Computer*, 36(6):47–56, 2003.

[118] Bongshin Lee, Kate Isaacs, Danielle Albers Szafir, GE Marai, Cagatay Turkay, Melanie Tory, Sheelagh Carpendale, and Alex Endert. Broadening intellectual diversity in visualization research papers. *IEEE computer graphics and applications*, 39(4):78–85, 2019.

[119] Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on Challenges in Representation Learning, ICML*, volume 3, page 2, 2013.

[120] Victoria L Lemieux and Jason R Baron. Overcoming the digital tsunami in e-discovery: is visual analysis the answer? *Canadian Journal of Law and Technology*, 9(1 & 2), 2011.

[121] Wei-Jen Li, Shlomo Hershkop, and Salvatore J Stolfo. Email archive analysis through graphical visualization. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pages 128–132. ACM, 2004.

[122] Ching-Yung Lin, Nan Cao, Shi Xia Liu, Spiros Papadimitriou, Jimeng Sun, and Xifeng Yan. Smallblue: Social network analysis for expertise search and collective

intelligence. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pages 1483–1486. IEEE, 2009.

[123] Rhema Promise Linder. *Grounded Visual Analytics: A New Approach to Discovering Phenomena in Data at Scale*. PhD thesis, 2019.

[124] Sheng Jie Luo, Li Ting Huang, Bing Yu Chen, and Han Wei Shen. Emailmap: Visualizing event evolution and contact interaction within email archives. In *Visualization Symposium (PacificVis), 2014 IEEE Pacific*, pages 320–324. IEEE, 2014.

[125] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[126] Sean McKenna, Diane Staheli, Cody Fulcher, and Miriah Meyer. Bubblenet: A cyber security dashboard for visualizing patterns. In *Computer Graphics Forum*, volume 35, pages 281–290. Wiley Online Library, 2016.

[127] Miriah Meyer, Michael Sedlmair, and Tamara Munzner. The four-level nested model revisited: blocks and guidelines. In *Proceedings of the 2012 BELIV Workshop: Beyond Time and Errors-Novel Evaluation Methods for Visualization*, page 11. ACM, 2012.

[128] Miriah Meyer, Michael Sedlmair, P Samuel Quinan, and Tamara Munzner. The nested blocks and guidelines model. *Information Visualization*, 14(3):234–249, 2015.

[129] Julia Moehrmann and Gunther Heidemann. Efficient annotation of image data sets for computer vision applications. In *Proceedings of the 1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*, page 2. ACM, 2012.

[130] Bill Moggridge, JF Suri, and D Bray. People and prototypes. *Designing interactions*, pages 641–735, 2007.

[131] Tamara Munzner. A nested model for visualization design and validation. *IEEE transactions on visualization and computer graphics*, 15(6), 2009.

[132] Tamara Munzner. *Visualization analysis and design*. CRC Press, 2014.

[133] Galileo Mark Namata, Brian Staats, Lise Getoor, and Ben Shneiderman. A dual-view approach to interactive network visualization. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 939–942. ACM, 2007.

[134] Jakob Nielsen. The usability engineering life cycle. *Computer*, 25(3):12–22, 1992.

[135] Adam Perer, Ido Guy, Erel Uziel, Inbal Ronen, and Michal Jacovi. Visual social network analytics for relationship discovery in the enterprise. In *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pages 71–79. IEEE, 2011.

[136] Adam Perer and Ben Shneiderman. Beyond threads: Identifying discussions in email archives. Technical report, MARYLAND UNIV COLLEGE PARK HUMAN COMPUTER INTERACTION LAB, 2005.

[137] Adam Perer and Ben Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12(5), 2006.

[138] Adam Perer and Marc A Smith. Contrasting portraits of email practices: visual approaches to reflection and analysis. In *Proceedings of the working conference on Advanced visual interfaces*, pages 389–395. ACM, 2006.

[139] Adam Perer and Jimeng Sun. Matrixflow: temporal network visual analytics to track symptom evolution during disease progression. In *AMIA annual symposium proceedings*, volume 2012, page 716. American Medical Informatics Association, 2012.

[140] Jorge Poco, Aritra Dasgupta, Yaxing Wei, W Hargrove, C Schwalm, R Cook, Enrico Bertini, and C Silva. Similarityexplorer: A visual inter-comparison tool for multi-faceted climate data. In *Computer Graphics Forum*, volume 33, pages 341–350. Wiley Online Library, 2014.

[141] H. Purchase. *Experimental Human Computer Interaction: A Practical Guide with Visual Examples.* Cambridge University Press, 2012.

[142] Tim Repke, Ralf Krestel, Jakob Edding, Moritz Hartmann, Jonas Hering, Dennis Kipping, Hendrik Schmidt, Nico Scordialo, and Alexander Zenner. Beacon in the dark: A system for interactive exploration of large email corpora. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 1871–1874. ACM, 2018.

[143] Osterman Research. Key issues for e-discovery and legal compliance. An Osterman Research White Paper, 2017.

[144] Marc Rettig. Prototyping for tiny fingers. *Communications of the ACM*, 37(4):21–27, 1994.

[145] Jonathan C Roberts, Chris Headleand, and Panagiotis D Ritsos. Sketching designs using the five design-sheet methodology. *IEEE transactions on visualization and computer graphics*, 22(1):419–428, 2015.

[146] Toni Robertson and Jesper Simonsen. Challenges and opportunities in contemporary participatory design. *Design Issues*, 28(3):3–9, 2012.

[147] Steven L Rohall, Daniel Gruen, Paul Moody, and Seymour Kellerman. Email visualizations to aid communications. In *Proceedings of InfoVis 2001 The IEEE Symposium on Information Visualization, IEEE*, volume 12, page 15, 2001.

[148] Jim Rudd, Ken Stern, and Scott Isensee. Low vs. high-fidelity prototyping debate. *interactions*, 3(1):76–85, 1996.

[149] Sébastien Rufiange and Guy Melançon. Animatrix: A matrix-based visualization of software evolution. In *Software Visualization (VISSOFT), 2014 Second IEEE Working Conference on*, pages 137–146. IEEE, 2014.

[150] Per Runeson and Martin Höst. Guidelines for conducting and reporting case study research in software engineering. *Empirical software engineering*, 14(2):131, 2009.

[151] Dominik Sacha, Matthias Kraus, Daniel A Keim, and Min Chen. Vis4ml: An ontology for visual analytics assisted machine learning. *IEEE transactions on visualization and computer graphics*, 25(1):385–395, 2019.

[152] Dominik Sacha, Manuel Stein, Tobias Schreck, Daniel A Keim, Oliver Deussen, et al. Feature-driven visual analytics of soccer data. In *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*, pages 13–22. IEEE, 2014.

[153] Michael Sedlmair, Annika Frank, Tamara Munzner, and Andreas Butz. Relex: Visualization for actively changing overlay network specifications. *IEEE transactions on visualization and computer graphics*, 18(12):2729–2738, 2012.

[154] Michael Sedlmair, Petra Isenberg, Dominikus Baur, and Andreas Butz. Information visualization evaluation in large companies: Challenges, experiences and recommendations. *Information Visualization*, 10(3):248–266, 2011.

[155] Michael Sedlmair, Petra Isenberg, Dominikus Baur, Michael Mauerer, Christian Pigorsch, and Andreas Butz. Cardiogram: visual analytics for automotive engineers. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1727–1736. ACM, 2011.

[156] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics*, 18(12):2431–2440, 2012.

[157] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012.

[158] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*, pages 336–343. IEEE, 1996.

[159] Ben Shneiderman and Catherine Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings*

of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization, pages 1–7. ACM, 2006.

[160] Barry Smith. Foundations of gestalt theory. 1988.

[161] Ian Soboroff. A comparison of pooled and sampled relevance judgments in the trec 2006 terabyte track. In *EVIA@ NTCIR*, 2007.

[162] GJ Socha and T Gelbmann. The electronic discovery reference model (edrm). `http://edrm.net/`, 2009.

[163] Fabian Sperrle, Jürgen Bernard, Michael Sedlmair, Daniel Keim, and Mennatallah El-Assady. Speculative execution for guided visual analytics. *arXiv preprint arXiv:1908.02627*, 2019.

[164] John Stasko, Carsten Görg, and Zhicheng Liu. Jigsaw: supporting investigative analysis through interactive visualization. *Information visualization*, 7(2):118–132, 2008.

[165] Chad A Steed, Margaret Drouhard, Justin Beaver, Joshua Pyle, and Paul L Bogen. Matisse: A visual analytics system for exploring emotion trends in social media text streams. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 807–814. IEEE, 2015.

[166] Florian Stoffel, Hanna Post, Marcus Stewen, and Daniel A Keim. polimaps: supporting predictive policing with visual analytics. In *Proceedings of the EuroVis Workshop on Visual Analytics*, pages 43–47. Eurographics Association, 2018.

[167] Chris Stolte, Diane Tang, and Pat Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, 2002.

[168] Guodao Sun, Yingcai Wu, Shixia Liu, Tai-Quan Peng, Jonathan JH Zhu, and Ronghua Liang. Evoriver: Visual analysis of topic coopetition on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1753–1762, 2014.

[169] Min Tang, Xiaoqiang Luo, and Salim Roukos. Active learning for statistical natural language parsing. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 120–127. Association for Computational Linguistics, 2002.

[170] Edward R Tufte. Envisioning information. *Optometry & Vision Science*, 68(4):322–324, 1991.

[171] Edward R Tufte and PR Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

[172] VALCRI. Visual analytics for sense-making and criminal intelligence analysis, 2017.

[173] Stef van den Elzen and Jarke J van Wijk. Small multiples, large singles: A new approach for visual data exploration. In *Computer Graphics Forum*, volume 32, pages 191–200. Wiley Online Library, 2013.

[174] Frank Van Ham and Adam Perer. search, show context, expand on demand: supporting large graph exploration with degree-of-interest. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):953–960, 2009.

[175] Frank Van Ham, Hans-Jörg Schulz, and Joan M Dimicco. Honeycomb: Visual analysis of large scale social networks. In *IFIP Conference on Human-Computer Interaction*, pages 429–442. Springer, 2009.

[176] Fernanda B Viégas, Scott Golder, and Judith Donath. Visualizing email content: portraying relationships from conversational histories. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988. ACM, 2006.

[177] Fernanda B Viégas, Scott Golder, and Judith Donath. Visualizing email content: portraying relationships from conversational histories. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 979–988. ACM, 2006.

[178] Fernanda B Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the*

SIGCHI conference on Human factors in computing systems, pages 575–582. ACM, 2004.

[179] Fernanda B Viegas, Martin Wattenberg, and Jonathan Feinberg. Participatory visualization with wordle. *IEEE transactions on visualization and computer graphics*, 15(6):1137–1144, 2009.

[180] Harold Weiss and James B McGrath. *Technically speaking: Oral communication for engineers, scientists, and technical personnel*. McGraw-Hill, 1963.

[181] Elaine Welsh. Dealing with data: Using nvivo in the qualitative data analysis process. In *Forum qualitative sozialforschung/Forum: qualitative social research*, volume 3, 2002.

[182] Steve Whittaker and Candace Sidner. Email overload: exploring personal information management of email. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 276–283. ACM, 1996.

[183] Yingcai Wu, Shixia Liu, Kai Yan, Mengchen Liu, and Fangzhao Wu. Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1763–1772, 2014.

[184] Panpan Xu, Yingcai Wu, Enxun Wei, Tai-Quan Peng, Shixia Liu, Jonathan JH Zhu, and Huamin Qu. Visual analysis of topic competition on social media. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2012–2021, 2013.

[185] Qian Yang, Jina Suh, Nan-Chen Chen, and Gonzalo Ramos. Grounding interactive machine learning tool design in how non-experts actually build models. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 573–584. ACM, 2018.

[186] Ji Soo Yi, Youn ah Kang, and John Stasko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6):1224–1231, 2007.

[187] Ji Soo Yi, Niklas Elmqvist, and Seungyoon Lee. Timematrix: Analyzing temporal social networks using interactive matrix-based visualizations. *Intl. Journal of Human–Computer Interaction*, 26(11-12):1031–1051, 2010.

[188] Kelvin S Yiu, Ronald Baecker, Nancy Silver, and Byron Long. A time-based interface for electronic mail and task management. *ADVANCES IN HUMAN FACTORS ERGONOMICS*, 21:19–22, 1997.

[189] Jian Zhao, Nan Cao, Zhen Wen, Yale Song, Yu-Ru Lin, and Christopher Collins. # fluxflow: Visual analysis of anomalous information spreading on social media. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1773–1782, 2014.

[190] Jian Zhao, Liang Gou, Fei Wang, and Michelle Zhou. Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 203–212. IEEE, 2014.

[191] Jian Zhao, Zhicheng Liu, Mira Dontcheva, Aaron Hertzmann, and Alan Wilson. Matrixwave: Visual comparison of event sequence data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 259–268. ACM, 2015.

[192] Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 1137–1144. Association for Computational Linguistics, 2008.

[193] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, Carnegie Mellon University, PA, US, 2002.

[194] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

[195] Hartmut Ziegler, Tilo Nietzschmann, and Daniel A Keim. Visual analytics on the financial market: Pixel-based analysis and comparison of long-term investments. In *2008 12th International Conference Information Visualisation*, pages 287–295. IEEE, 2008.

[196] Torre Zuk, Lothar Schlesier, Petra Neumann, Mark S Hancock, and Sheelagh Carpendale. Heuristics for information visualization evaluation. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–6. ACM, 2006.

Appendix A

Appendix A

A.1 Publications

Toward the overall goal of supporting E-discovery analysts through interactive visualisations, we presented papers at various International forums. They are as follows:

[1] Sathiyanarayanan, M., Turkay, C. and Dykes, J. Visualising E-mail Communication to Improve E-discovery. IEEE VIS, InfoVis Conference, 2018.

[2] Sathiyanarayanan, M., Turkay, C. and Fadahunsi, O. Design and implementation of small multiples matrix-based visualisation to monitor and compare email socio-organisational relationships, IEEE, COMSNETS, 2018.

[3] Sathiyanarayanan, M. Improving Visual Investigation Analysis of Digital Communication Data within E-discovery. IEEE Conference on Information Visualisation (InfoVis) 1-6 October 2017, Phoenix, Arizona, USA.

[4] Sathiyanarayanan, M. and Turkay, C. Challenges and Opportunities in using Analytics Combined with Visualisation Techniques for Finding Anomalies in Digital Communications. DESI VII Workshop on Using Advanced Data Analysis in eDiscovery & Related Disciplines to Identify and Protect Sensitive Information in Large Collections 12 June 2017, London, UK.

[5] Sathiyanarayanan, M. and Turkay, C. Determining and Visualising E-mail Subsets to Support E-discovery. IEEE VIS, InfoVis Conference, 2016.

[6] Sathiyanarayanan, M. and Turkay, C. Is Multi-perspective Visualisation recommended for E-discovery Email Investigations? IEEE VIS 2016 Workshop - Creation, Curation, Critique and Conditioning of Principles and Guidelines in Visualization 23 October, Baltimore, Maryland, US.

To be Submitted

[1] Sathiyanarayanan, M., Nguyen, P., Turkay, C. and Dykes, J. Discovery, Specification and Modelling of Communication Patterns using Interactive Visualisation Assisted Active Learning. IEEE VIS, 2020 (March 30).

[2] Sathiyanarayanan, M., Turkay, C. and Dykes, J. Design and Implementation of Multi-faceted Interactive Visualisation to Discover Interesting Information in E-mail Communication Data, IEEE, TVCG, 2020 (May 30).

A.2 Technologies Used

Since the project is in collaboration with the Red Sift company in London, they have provided a core infrastructure for email management that is required for the project (IMAP connectors, stream processing engine, application framework, persistence, etc.) which is mostly in place within the Red Sift platform. This significantly de-risks the project and allows to focus on the actual visualisation and interaction models which are the core subjects of this project. We will adopt an incremental technical complexity approach by starting with an individual inbox analysis and then extending it to multiple inboxes.

Due to the web-based nature of the organisation platform (Red Sift's requirement), we employed web-enabled interaction and visualisation capabilities and we will use Data-driven Documents JavaScript library (D3.js) by Mike Bostock [41] as the framework for data visualisation. D3 is recognised as the current best in class platform for building interactive data visualisation which uses JavaScript. In order to ensure a modular structure that facilitates the extendibility our solutions, we also incorporated a high level visualisation

grammar module called Vega[1] and/or Vega Lite[2]. The solution utilised a number of open source solutions for data presentation. To test initial features, interactions, and other analysis we used Tableau[3] and R[4]. However, for designing specialised views, we considered D3.

Importance of using Tableau: Tableau is an interactive explorative visualisation tool, business intelligence system, mainly used in a business setting for making decisions through an innovative means of visualising their data. Tableau provides immediate insight by transforming data into visually engaging, interactive views in dashboards. The tool helps in saving development time and one can quickly understand the importance and power of visualisation for quicker decision making and for finding the real value in the data. Tableau can be integrated with R for more statistical analysis.

Importance of using R: R is the most comprehensive statistical analysis package available that provides a comprehensive language for managing and manipulating data. R is extensible and offers a rich functionality, graphics and charting capabilities built for developers to build their own tools and methods for analysing data. The dplyr and ggplot2 packages for data manipulation and plotting, respectively, will add more value for analysis and prototype development. The underlying modular architecture will also enable information and skill exchange between researchers. Similar modular (or package based) systems where researchers can exchange their "modules" have shown tremendous impact within tools such as R which has now become the main analysis environment for a wide spectrum of researchers from computer science, biology, finance, and even social science. As such, the modular architecture of the project has the same potential to grow quickly and become a platform where many share methods and insight. We were able to integrate R and Tableau to find interesting communication patterns.

Importance of using Python: Python is a developer-friendly language which has numerous libraries/packages for machine learning and other computations such as numpy,

[1] http://vega.github.io/vega/
[2] https://vega.github.io/vega-lite/
[3] https://www.tableau.com/
[4] https://www.rproject.org/

pandas, keras, tensorflow etc. These packages are well-documented, which is helpful in starting with new projects and solutions - data analysis, pattern recognition, etc. The documents and online support helps in speeding up the process of fixing bugs.

Importance of using D3: D3.js is a JavaScript library used for displaying dynamic, interactive visualisation in a web browser (front-end) by taking data in txt, csv or json file in the back-end (the server). It is also known as Data-Driven Documents. It uses HyperText Markup Language (HTML), Cascading Style Sheets (CSS), and Scalable Vector Graphics (SVG), which focuses on binding data to DOM elements, to create visual representations of data by providing interaction and animation features. D3.js is open-source can be used with other JS framework such as Angular.js, React.js, Ember.js etc. D3 is lightweight, and works directly with web standards, it is extremely fast and works well with large datasets. Our collaborator's Red Sift platform (front-end) takes in JavaScript, so the transferability will be quick.

A.3 Interactive Visual Active Learning

The complete feature engineering and active learning was developed by Dr Phong Nguyen. After an exploratory analysis phase with the features and the views discussed in Chapter 5, analysts moved to the modelling phase which comprises of a number of stages as also illustrated in Figure. A.1:

Stage-1: Class Generation – After consolidating the understanding from the interactive analysis, analysts declare the names of the classes that they decide to have in the classification model. This is supported by a simple menu (top left in the interface as seen in Figure. 5.22) where they name a new model and declare the class names.

Stage-2: Initial labelling and model creation – Once the set of classes are declared, the analyst selects the most representative threads for each class using either the thread projection or the proxy feature views, and labels them through the interface. Once the initial labelling is performed, a model is created and trained on the labelled threads. Important to note here that the feature space F that is generated for all the threads (Figure. A.1:)

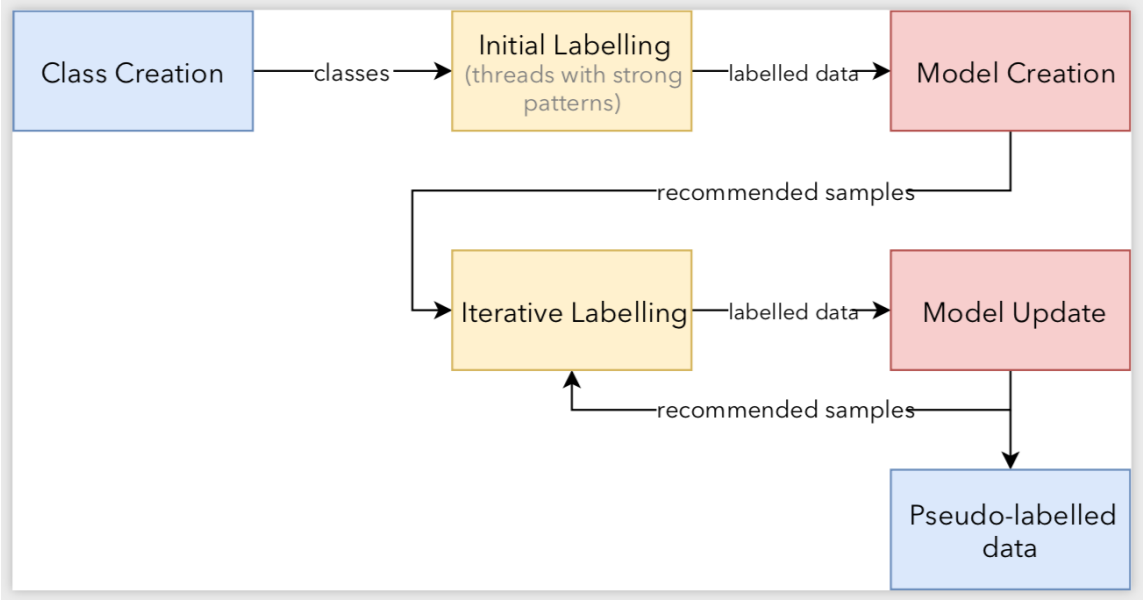


Figure A.1: The active learning and pseudo-labelling process

is fed into the model for the computations.

In this work, we utilise a semi-supervised algorithm called *Label Propagation* [194, 193]. This algorithm is preferred for its capability to work with small numbers of labelled data instances which are then "*propagated*" to unlabelled instances depending on their similarity. In order to achieve this, all the data instances are represented as vertices in a weighted graph and the similarity between the instances are treated as edge weights.

Stage-3: Iterative model building through active learning – Once the initial model is built, this model is used to predict the classes of all the threads. These *predicted* labels are then transferred to the visualisation side for further investigation. Such an approach where approximately labelled instances are considered as potential true labels (mostly for training purposes) is often referred to as pseudo-labelling [119] and our approach here makes use of a similar attitude and returns a full labelling of the data. These labels are then overlaid on all the thread feature-based views as colours of the classes they belong to. The results of such a full labelled set could be seen in Figure. 5.22 in the thread features and projection views.

Uncertainty sampling: In addition to the pseudo-labels, we also support analysts in choosing suitable threads to label in each iteration. Selecting the most informative instances to label within each iteration of the learning cycle is a key aspect of active learning methods and several strategies exist [78]. In our approach, we make use of the uncertainty sampling method and identify the instances for which the model is most uncertain for as recommended samples for labelling [192]. In order to identify the most uncertain samples, we make use of an entropy-based uncertainty proxy introduced by Tang et al. [169]. Since label propagation is a probability based classification model, we can estimate to what extent each instance belongs to the different classes, and the instances that exhibit equal levels of belonging to each of the classes, i.e., high entropy in the distribution of class membership, are selected for recommendations for labelling in the next iteration of the active learning phase. We indicate the recommended instances, i.e., threads in this case, with crosses in all the thread projection plot as can be seen in Figure. 5.22.

Once these recommendations are made, the analysts have the freedom to either to label them or label any instance that they think are mis-labelled or are important to be labelled. Through the visualisation of the pseudo-labels and the recommended samples, we aim to support analysts in converging towards satisfactory models that they can understand and trust in a short training phase as possible. In Chapter 5, Empirical Evaluation section, we have demonstrated how this iterative process works and how it supports analysts in building classification models.

A.4 Topic Analysis Visualisation

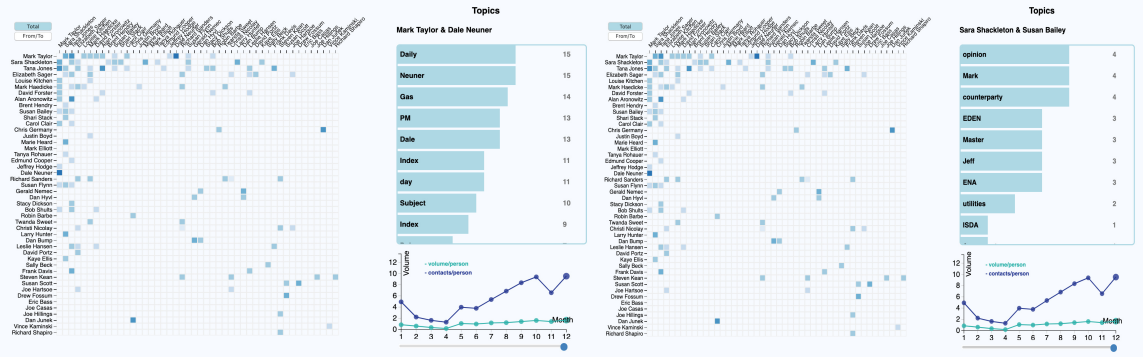


Figure A.2: Random screenshots of the topics discussed between two individuals

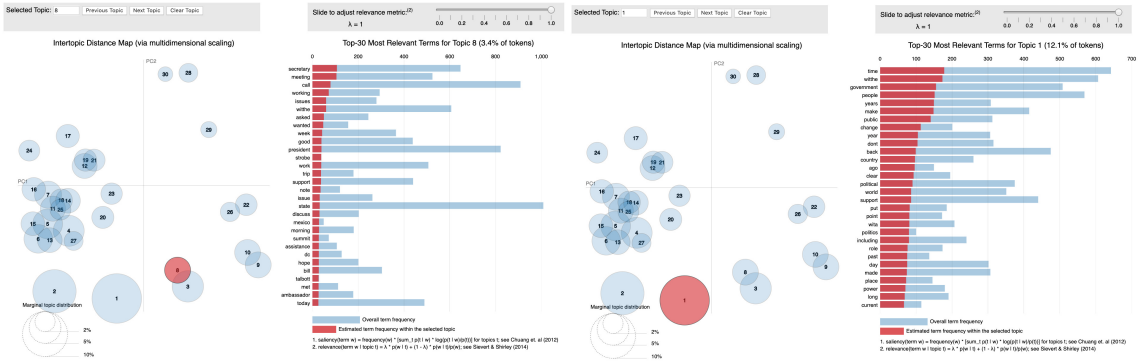


Figure A.3: Random screenshots of the topics discussed between individuals based on LDA

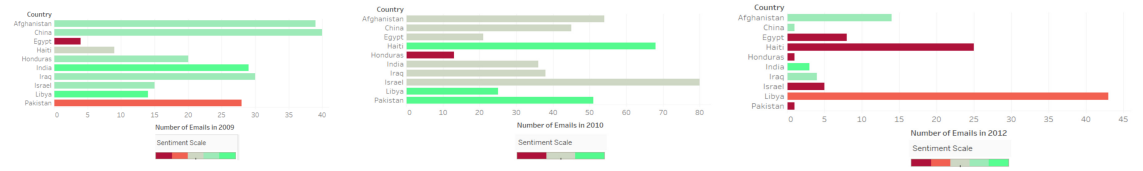


Figure A.4: The screenshots represent sentiment for 10 countries discussed by two individuals for 3 years

A.5 Ethical Approval

City, University of London
Low Risk Application for Approval of Research Involving Human Participants

PLEASE NOTE THE FOLLOWING:

- Ethical approval **MUST** be obtained before any research involving human participants is undertaken. Failure to do so may result in disciplinary procedures being instigated, and you will not be covered by City's indemnity if you do not have approval in place. It may also result in the degree not being awarded or the data not being published in a peer review journal.
- The Signature Sections **MUST** be completed by the Principal Investigator (the supervisor and the student if it is a student project).

I confirm that I have reviewed the relevant checklist(s) and that my research project is suitable for low risk review. **YES** ☒ **NO** ☐

Tick this box if you do not grant City permission to use your application form for training purposes. ☐

1. Applicant Details	
1.1 Name of Principal Investigator(s) (if this is a student project, please note that the Principal Investigator is the supervisor and all correspondence will be with the supervisor):	Mithileysh Sathiyanarayanan
1.2 Name of student (if student research)	
1.3 Degree programme (if student research)	
1.4 Department/School	Computer Science
1.5 City email address (**not** private email)	mithileysh.sathiyanarayanan@city.ac.uk
1.6 Name and status of others taking part in the project, e.g. students, research assistants, external collaborators	
1.7 Do any of the investigators listed have any direct personal involvement (e.g. financial, shareholding, personal relationship etc.) **in the organisations sponsoring or funding the research that may give rise to a possible conflict of interest?**	YES ☒ NO ☐ If yes, please provide details and attach the supporting correspondence. Redsift Inc. is sponsoring this research (letter attached).
1.8 Will any of the investigators receive any personal benefit or incentives, including payment above normal salary, from undertaking the research or the results of the research above those normally associate with scholarly activity?	YES ☐ NO ☒ If yes, please provide details and attach the supporting correspondence.
1.9 Will the research data collected be shared with the outside	YES ☐ NO ☒

organisation? If so, will the outside organisation be a joint data controller with City or a data processor on behalf of City?	If yes, please provide details and attach the supporting correspondence.
1.10 Please detail any data sharing arrangements (joint data controller) or data processing contract (data processor on behalf of City) with the outside organisation. Please attach a copy of the data sharing agreement (joint data controller) or contract (data processor) with your application. The Information Compliance Team at dataprotection@city.ac.uk can provide further advice.	YES ☐ NO ☐ If yes, please provide details and attach the supporting correspondence.

*In order to apply for ethical approval for a research project you must be a current member of City, University of London; this means that you must be a currently registered student (not suspended or having completed your studies) or a current member of staff (not on a leave of absence or retired). In special circumstances those holding honorary contracts with City, University of London may apply for ethical approval for research, but this must be agreed with the Chair of senate REC in advance.

2. Project Overview	
2.1 Project title	Visual Analysis of E-Mail Communication Data to Support E-Discovery
2.2 Duration of project Please note that no data collection can take place until the study has been approved.	Start date: **15/09/2018** Estimated end date: **30/05/2019**
2.3 Non-technical summary Please provide a brief outline of the background, aims, key questions and significance of the project (maximum 500 words).	
The main aim of the research is to design and develop visual solutions to unravel the information in E-mail communication data to support E-discovery in an organisation by facilitating the generation of visual evidence for users/analysts. This will enable analysts to compare various communication features from multiple perspectives, identify relevant subsets of data and find anomalous communication behaviour. As visualisation tools and solutions continue to be improved and optimised, analysts are increasingly calling for novel techniques that can improve in identifying and understanding of various communication features to understand unusual behaviour in E-mail communication data. In this research, we will be developing techniques and implementing investigative strategies in software prototypes through a structured process of abstraction, design and testing, by using a well-known methodology called Design Study Methodology. Doing so is intended to explore and answer a series of research questions in ways that will improve the role of visualisation and visual evidence in E-discovery.	
2.4 Research Methodology Please provide a summary and brief explanation of the research design, including overall aims, methods data collection and data analysis (maximum 500 words).	
The goal of the unstructured interviews is to characterise the problems faced by analysts in E-discovery and characterise the tasks that they perform. In these types of series of interviews, just talking and contextual inquiries along with deep literature study will provide interesting and relevant information where the researcher observes users working in their real-world context and interrupts to ask questions when clarification is needed, also clarifies many points by referring/conducting literature review.	

Method (procedure, including data gathering & analysis):
We aim to conduct in-person biweekly meetings (once in two weeks) with the Red Sift experts starting from 15th September 2018 to 30th May 2019, to understand the potential use-cases and requirements on how visualisation can be used in the context of the analysis of email data, in particular for the purposes of E-discovery. Moreover, we expect to get their expert feedback on the prototypes that are iteratively built and improved based on their feedback. Each meeting would last up to one hour (maximum) at their location (Red Sift, London).

The experts will be provided with a detailed study participant information sheet, consent form and all potential participants will be screened for eligibility, including their age and computing experience required for this project. These will be signed by the experts only once and not every time. That is, if these forms have been signed once by the interviewed expert, it will cover all the future meetings I have with them. I will not be recruiting anyone under the age of 18, and vulnerable participants will be screened out.

The discussion would be informal without any structured approach. The researcher may or may not have a clear approach/plan in mind regarding the focus and goal of the discussion. The discussions tend to be open-ended and express little control over experts' responses. The researcher jots down the notes/points while the discussion is taking (hand-written). Since informal discussions occur 'on the fly', it is difficult to tape-record this type of interview. The researcher engages in the discussion to develop an understanding of the tasks and requirements. This type of discussions need to be included immediately in the researcher's field notes. The discussions will also help to uncover new areas or topics of interest that may have been overlooked by previous research. The discussions are highly informal because the researcher's understanding is still evolving, it is helpful to anticipate the need to speak with experts on multiple occasions.

In an effort to better understand, we aim to collect suggestions, comments and feedbacks from the Red Sift experts, by making notes in the diary book. Then, the researcher can elaborate on the points based on the conversation and observation with the experts (based on the hand-written notes). Finally, reflecting on the discussions, researchers can decide to have another discussion session or not. All the recorded notes will be transcribed, encoded and elaborated in detail in the report. Any additional notes or diagrams made/showed by participants will be digitised following the session. A thematic analysis of the notes will be completed after all sessions have taken place. The findings will be used to refine the design of the support materials.

In the interviews, we aim at focusing "How do E-discovery analysts/investigators gain insights from large and complex email data?. For task and data abstraction, we use the Why?, What?, How? framework to abstract the tasks, explore visualisations and develop interaction paradigms that would satisfy these tasks. We will make it flexible based on the nature of the project.

As a starting point to discuss about the email communication analysis, the Enron real-case will be unveiled and some of the following open-ended questions will be asked:
Q1: What are the challenges of E-discovery (E-mail communication investigations) with respect to visualisation?
Q2: How will you investigate on the key time-frame, key words and key individuals/players involved?
Q3: How will you categorize normal and abnormal E-mail communication data? How will you characterize suspicious behaviours?
Q4: How can visualisation inform unexpected behaviours? Do you think there are useful tools for investigations?

The above questions along with the "topic guide" (attached) questions/task can co-occur. The discussions in the series of interviews will help us understand the tasks and requirements in E-discovery that the analysts are interested in the email communication data investigations. The research questions along with the E-discovery requirements will help us abstract out some generalisable tasks and build visual solutions.

2.5 Ethical Issues
What do you consider are the ethical issues associated with conducting this research and how do you propose to address them?

The ethical issues considered for conducting this research are:

- **Anonymity:** no personal data will be collected and all other data captured will be reported in a completely anonymised format, with researchers having access only to their interview transcripts. Participant names will not be associated with the notes or any other data, and will not appear in any reports or presentations.

- **Consent:** we will be taking the consent from the experts before we begin the study, consisting of a series of interviews.
- **Confidentiality**: all the data will be kept private, both from other participants and when reporting findings. We will not be revealing the experts/analysts' names. Data will be held securely and kept confidential.
- **Disclosure:** the names of the participants and other personal information will not be recorded in the notebooks or the transcriptions. It will be completely anonymised. The company name is not anonymized and we have received the permission from the company to publish the same.
- **Permission:** all the recorded contribution, in written form, taken from the interview by the researcher, will be used in in line with the participant's preferences as recorded in our consent form.
- **Data Protection:** all the data will be password protected, stored securely, and backed up. Only myself, my supervisory team (Dr Cagatay Turkay and Prof Jason Dykes), and my examiners, will have access to the data. If a participant decides to withdraw from the study at any point, I will destroy any data already gathered from them.
- **Others:**

a) participants will be given a participant information sheet and be asked to consider and sign a consent form.
b) The names will not be recorded, we will use pseudonyms to retain anonymity in our notes.
c) no audio, video or screen capture will be recorded.
d) No personal information will be collected, identities will not be recorded or revealed.
e) The project will adhere to the requirements of data protection rules in terms of data labelling, storage and security relevant based on the current General Data Protection Regulation (GDPR).
f) Data will be stored for 10 years, and then destroyed (which is a standard), in locked filing cabinets and on password protected computers.

Dissemination

The project is designed to inform multiple stakeholders, including the academic community, governments and industries. In addition to the usual academic publications, all participants in the study will be offered access to the findings and resulting recommendations, via the school, in a form appropriate to their interests. Most likely this will be in the form of a short and accessible report providing an overview of findings and noting conclusions and any recommendations.

No discomfort or misrepresentation is anticipated as a consequence of the dissemination process. However, the researchers will take care to present the material in a manner that treats the reported experiences with respect. The outcome of the interviews along with our results will be published in journals, conferences papers as well as the researchers' thesis/report. The thesis will be made available on the City Research Online.

2.6 Where will the research take place (e.g. at City, in a library, café, not in person etc.)?
If the research is taking place in participant's homes, please describe the policy for lone working that you will be following. (Please contact the Health & Safety Office (safetyoffice@city.ac.uk) for information about City's lone working policy.)

2.7 Are there any health or safety issues? If yes, please provide details and information about how these will be mitigated.	YES ☐ NO ☒
2.8 It is a requirement that at least an initial assessment of risk be undertaken for all research and if necessary a more detailed risk assessment be carried out. Are there hazards associated with undertaking this activity where a risk assessment would be required? If no, please provide a brief explanation why. Safe office environment, no greater risk than everyday work.	YES ☐ NO ☒

If yes, it is mandatory to complete a risk assessment. Please contact the Health & Safety Office (safetyoffice@city.ac.uk) for advice on risk assessments and/or how to complete it.	
2.9 Is the research funded? If yes, please provide details.	YES ☒ NO ☐
If the project is funded, does the funding body (e.g. ESRC) require that the data be stored and made available for reuse/sharing? *This is an industrially funded project with no requirements on the reuse/sharing of the data.*	
2b If you have responded yes to any of the questions above, explain how you are intending to obtain explicit consent for the reuse and/or sharing of the data and how consent can be revisited if it becomes clear later that the original consent does not cover future reuse? *The details on how the dissemination of the results from the project could be made are provided in the funding agreement attached in accordance with the clauses 6.2 and 6.3.* *The funders have already provided their consent to support the study (see the confirmation letter attached) and unlikely to object to the sharing of the results when Clause 6.3 is followed during the dissemination of the results. If any further issues comes up, we will get advice from the Research and Enterprise office at City and find a solution that works best both for our dissemination purposes and the potential conflicting interests of the company.*	

3. External approvals/international research Please note that if you are travelling outside the UK to conduct your research in a country with a green travel advisory from the Foreign & Commonwealth Office (FCO) you are not eligible for low risk and should fill in the full application form. Research taking place in a country where the the FCO has issues a red travel advisory, you need to apply to Senate Research Ethics Committee using a full application form.	
3.1 If any part of the investigation is being carried out under the auspices of an outside organisation, involves collaboration between institutions or individual external researchers, or institutions/organisations where interviews/fieldwork will take place, please give details and address of organisation(s). Red Sift is a Cyber Security & Email Analytics Company, the interviews will be conducted at Wayra, Red Sift, 20 Air St, Soho, London, W1B 5AN, UK	YES ☒ NO ☐ N/A ☐
3.2 If applicable, has permission to conduct research in, at or through another institution or organisation been obtained? If yes, please provide details and attach the supporting correspondence. **Attached.**	YES ☒ NO ☐ N/A ☐
3.3 Is any part of the research being conducted with participants outside the UK, e.g. Skype interviews? If yes, please provide details of where. If no, please go to section 4.	YES ☐ NO ☒
3.4 Have you identified and complied with all local requirements concerning ethical approval & research governance and data protection*?	YES ☒ NO ☐

*Please note that many countries require local ethical approval or registration of research projects, further some require specific research visas. You must also ensure you are aware of and abide by the national data protection legislation including legal requirements around research using patient data and transfer of data to and from the UK. If you do not abide by the local rules of the host country, you will invalidate your ethical approval from City, and may run the risk of legal action within the host country.

4. Participants & recruitment	
4.1 Please provide information about the participants, including age, gender, any inclusion/exclusion criteria and how many participants will be recruited.	
Within Red Sift, there are three experts with experience in the area of email communication (inclusion criteria) Expert 1, Male, 43 years old. Expert 2, Male, 41 years old. Expert 3, Male, 40 years old.	
4.2 How are the participants to be identified and approached, and by whom? Please note that you must consider whether potential participants will feel under any undue pressure to participate.	
In the Red Sift company, they have three experts who have worked in the area of email communication. Since, the company is funding the research, they helped me identify the experts who will not feel under any undue pressure to participate.	
4.3 Describe the procedure that will be used when seeking and obtaining consent, including when consent will be obtained. Include details of (a) who will obtain the consent, (b) how you are intending to arrange for a copy of the signed consent form to be given to the participants, (c) when they will receive the participant information sheet, and (d) how long the participants have between receiving information about the study and giving consent. ? Please note that if you are relying on consent as the lawful basis for processing special category (sensitive) personal data, consent has to be freely given, specific, informed and unambiguous indication of the individual's wishes.	
(a) The participants will sign the consent form at the beginning of the study. Since the study has iterative interviews, they will sign only once when they first take part in the study. (b) The participants will sign two copies of the consent form. One copy of signed consent form for the participants and one for our reference. (c) The participant information sheet will also be given before the start a session that is part of the study. (d) The participants have 15-20 minutes between receiving information about the study and giving consent. However, if they wish to use more time to read through the information, they will be offered the option use a longer period and attend one of the next iteration of the meetings	
4.4 Please how you will comply with the data protection legislation for processing information provided by children?	**YES ☐ NO ☒**
4.5 Are you offering any incentives, rewards or payment for participating? If yes, please provide details	**YES ☐ NO ☒**
4.6 Will the results be made available to the participants? YES ☒ NO ☐ **If yes, please give details of in what form.**	
The outcome of the interviews along with our results will be published in journals, conferences papers as well as the researchers' thesis/report. The thesis will be made available on the City Research Online.	

5. Data Collection, Confidentiality and Data Handling	
5.1 Please indicate which of the following you will be using to collect your data ***Please tick all that apply***	
Questionnaires (paper based)	☐

Questionnaires (computer based)	☐
Interviews	☒
Participant observation	☒
Covert observation	☐
Observation of specific organisational practices	☐
Focus groups	☐
Audio/digital-recording interviewees or events	☐
Video recording	☐
Physiological measurements	☐
Digital/computer data	☐
Other	☐
Please give details if you have ticked other	

5.2 Will the research involve:	
• **complete anonymity of participants** (i.e. researchers will not meet, or know the identity of participants, as participants are a part of a random sample and are required to return responses with no form of personal identification)**?**	☐
• **anonymised sample or data** (i.e. an *irreversible* process whereby identifiers are removed from data and replaced by a code, with no record retained of how the code relates to the identifiers. It is then impossible to identify the individual to whom the sample of information relates)**?**	☒
de-identified samples or data (i.e. a *reversible* process whereby identifiers are replaced by a code, to which the researcher retains the key, in a secure location)? Please note that de-identified data may be treated as personal data under GDPR depending on how difficult it is to attribute a pseudonym to a particular individual.	☐
• **subjects being referred to by pseudonym in any publication arising from the research?**	☐
• **participants will be named**	
• **any other method of protecting the privacy of participants?** (e.g. use of direct quotes with specific permission only; use of real name with specific, written permission only)	☐
Please give details if 'any other method of protecting the privacy of participants' is used.	

5.3 Which of the following methods of assuring confidentiality of data will be implemented? *Please tick all that apply.*	
• data to be kept in a locked filing cabinet	☒
• data and identifiers to be kept in separate, locked filing cabinets	☐
• access to computer files to be available by password only	☒
• storage on an encrypted device (e.g. laptop, hard drive, USB)	☒
storage at City	☒
• storage at other site	☐
If stored at another site, please give details.	
5.3a Will the data be accessed by people other than the named researcher? (E.g. supervisor, translator, transcription service, colleagues, reviewers, etc.) If yes, please explain by whom and for what purpose and ensure you have consent for this. Dr Cagatay Turkay, supervisor of this project, will review the interview notes to improve the designs, tasks and the solutions. Also, to plan the iterative interviews effectively.	YES ☒ NO ☐
5.3b For how long will the data be kept? *Note that the institutional guidelines on retention state a minimum of 10 years,* ***but some funding bodies require a longer retention period.***	10 years

6 Insurance	
Does the research involve any of the following:	
Children under the age of 5 years	**YES** ☐ **NO** ☒

Over 500 participants?	**YES** ☐ **NO** ☒
Specifically recruiting pregnant women	**YES** ☐ **NO** ☒
Excluding information collected via questionnaires (either paper based or online), is any part of the research taking place outside of the UK? (E.g. Skype interviews.) Note that this does not include the researcher's location being outside the UK.	**YES** ☐ **NO** ☒

If you have answered 'yes' to any of the above questions you will need to check that the City's insurance will cover your research. You should do this by submitting this application and any accompanying documentation to insurance@city.ac.uk before applying for ethics approval.

*Note that it is the Committee's prerogative to ask to view risk assessments.

7. Final Checks	
Before submitting your application, please confirm the following, noting that your application may be returned to you without review if the reviewer/committee feels these requirements have not been met.	
There are no discrepancies in the information contained in the sections of the application form and the materials for participants.	☒
There is sufficient information regarding the study and materials to enable proper ethical review.	☒
The application form and materials for participants have been checked for grammatical errors, typos and clarity of expression.	☒
For students, the application form has been signed off by your supervisor.	☒

8. Documents

You are expected to provide copies of relevant documents including all letters to be sent to participants and other individuals (such as GPs) and organisations involved in the research. Please follow the guidelines and templates.

8.1 Document Checklist		
Please place an 'X' in all appropriate spaces for all documents you are submitting		
	Attached	**Not applicable**
Copy of study advertisement (including recruitment emails/letters)		☒
Participant information sheet	☒	
Participant consent form	☒	
Questionnaire(s)		
Topic guide(s) for interviews and/or focus groups	☒	
Confirmation letter(s) from / correspondence with external organisations	☒	
Confirmation that insurance is in place		☒
Product information		☒
GP Letter		☒
Data sharing agreement (with partner organisations)		☒
Contract with data processor (e.g. transcribing service)		☒
Other (please provide details)		
Permission from collaborating organization "Interviews Confirmation - RedSift.pdf"	☒	
Funding agreement	☒	

9. Declarations by Investigator(s)	
• I certify that to the best of my knowledge the information given above, together with any accompanying information, is complete and correct.	☒
• I have read City's guidelines on human research ethics, and accept the responsibility for the conduct of the procedures set out in the attached application.	☒
• I have attempted to identify all risks related to the research that may arise in conducting the project.	☒
• I have read and will comply with City's Data Protection and Information Security policies	☒

• I understand that **no** research work involving human participants or data can commence until **full** ethical approval has been given	☒

	Print Name	Signature
Principal Investigator(s) (student and supervisor if student project)	Mithileysh Sathiyanarayanan	
Date	21/08/2018	

Ethics Proportionate Review Application: Staff and Research Students

Computer Science Research Ethics Committee (CSREC)

Staff and research students in the Department of Computer Science undertaking research that involves human participation must apply for ethical review and approval before the research can commence. If the research is low-risk, an application can be submitted to CSREC for proportionate review through the University's process (and form) for *Low Risk Application for Approval of Research Involving Human Participants*. Applicants are advised to read the CSREC Policy and Procedure documentation in full prior to submitting an application. This is available online.

The process begins by completing this *Ethics Checklist,* to determine whether the research is low-risk.

If so, then City University *Low Risk Application for Approval of Research Involving Human Participants* process should be followed. If not, the checklist provides guidance as to where approval should be sought, but the checklist itself does not need to be submitted.

Completed forms should be returned to the Chair of CSREC by email (s.m.wilson@city.ac.uk).

Ethics Checklist

If your answer to any of the following questions (1 – 3) is YES, you must apply to an appropriate external ethics committee for approval:		*Delete as appropriate*
1.	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure, please check at http://www.hra.nhs.uk/research-community/before-you-apply/determine-which-review-body-approvals-are-required/*	~~YES~~ / NO
2.	Will you recruit any participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee http://www.scie.org.uk/research/ethics-committee/*	~~YES~~ / NO
3.	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	~~YES~~ / NO

If your answer to any of the following questions (4 – 11) is YES, you must apply to the Senate Research Ethics Committee for approval (unless you are applying to an external ethics committee):		*Delete as appropriate*
4.	Does your research involve participants who are unable to give informed consent. *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf?*	~~YES~~ / NO
5.	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	~~YES~~ / NO
6.	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	~~YES~~ / NO
7.	Does your research involve participants disclosing information about sensitive subjects?	~~YES~~ / NO

8.	Does your research involve the researcher travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning? (http://www.fco.gov.uk/en/)	~~YES~~ / NO
9.	Does your research involve invasive or intrusive procedures? *For example, these may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	~~YES~~ / NO
10.	Does your research involve animals?	~~YES~~ / NO
11.	Does your research involve the administration of drugs, placebos or other substances to study participants?	~~YES~~ / NO

If your answer to any of the following questions (12 – 18) is YES, you must submit a full application to the Computer Science Research Ethics Committee (CSREC) for approval (unless you are applying to an external ethics committee or the Senate Research Ethics Committee). Your application may be referred to the Senate Research Ethics Committee.		*Delete as appropriate*
12.	Does your research involve participants who are under the age of 18?	~~YES~~ / NO
13.	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	~~YES~~ / NO
14.	Does your research involve participants who are recruited because they are staff or students of City University London? *For example, students studying on a particular course or module. (If yes, approval is also required from the Head of Department or Programme Director.)*	~~YES~~ / NO
15.	Does your research involve intentional deception of participants?	~~YES~~ / NO
16.	Does your research involve participants taking part without their informed consent?	~~YES~~ / NO
17.	Does your research pose a risk to participants greater than that in normal working life?	~~YES~~ / NO
18.	Does your research pose a risk to you, the researcher(s), greater than that in normal working life?	~~YES~~ / NO

You must make a proportionate review application to the CSREC if your research involves human participation, is deemed to be low risk (*i.e. your answer to all questions 1 – 18 is "NO"*) and you are not submitting any other ethics application.

Please use the form titled *Low Risk Application for Approval of Research Involving Human Participants* – see the CSREC web pages.

Complete the form, remove all instructions and use the University templates for:

- **Participant Information Sheet**
- **Consent Form**

Check spelling, grammar and consistency - then send all documentation in a single file to the CSREC Chair - dlo_CSResearchEthics@cityuni.onmicrosoft.com

Jason Dykes, June 2018

Title: **Visual Analysis of Email Communication Data to Support E-discovery**

Overall Aim and Methods:

The main aim of the research is to design and develop visual solutions to unravel the information in E-mail communication data to support E-discovery in an organisation by facilitating the generation of visual evidence for users/analysts. This will enable analysts to compare various communication features from multiple perspectives, identify relevant subsets of data and find anomalous communication behaviour. As visualisation tools and solutions continue to be improved and optimised, analysts are increasingly calling for novel techniques that can improve in identifying and understanding of various communication features to understand unusual behaviour in E-mail communication data. In this research, we will be developing techniques and implementing investigative strategies in software prototypes through a structured iterative process of abstraction, design and testing, by using a well-known methodology called Design Study Methodology. Doing so is intended to explore and answer a series of research questions in ways that will improve the role of visualisation and visual evidence in E-discovery.

What do we seek approval on?

In this project, regular unstructured iterative interviews will be closely conducted with the domain experts/analysts of Red Sift London (funders of this project) to understand the challenges, needs and requirements for the investigation of E-mail communication data. The engagement would be a biweekly (once in two weeks) starting from 15th September 2018 to 30th May 2019. The plan with these iterative interviews is to both learn about the design requirements and to iteratively evaluate the solutions that are developed as part of the thesis. So, basically these iterative interviews can lead to new insights and help in developing designs and prototypes. With this application, we seek approval for conducting this iterative interview series with the experts from and associated with Red Sift.

Details of Unstructured Interviews:

The goal of the unstructured interviews is to characterise the problems faced by analysts in E-discovery and characterise the tasks that they perform. In these types of series of interviews, just talking and contextual inquiries [1] along with deep literature study will provide interesting and relevant information where the researcher observes users working in their real-world context and interrupts to ask questions when clarification is needed, also clarifies many points by referring/conducting literature review.

Method (procedure, including data gathering & analysis):
We aim to conduct in-person biweekly meetings (once in two weeks) with the Red Sift experts starting from 15th September 2018 to 30th May 2019, to understand the potential use-cases and requirements on how visualisation can be used in the context of the analysis of email data, in particular for the purposes of E-discovery. Moreover, we expect to get their expert feedback on the prototypes that are iteratively built and improved based on their feedback. Each meeting would last up to one hour (maximum) at their location (Red Sift, London).

The experts will be provided with a detailed study participant information sheet, consent form and all potential participants will be screened for eligibility, including their age and computing experience required for this project. These will be signed by the experts only once and not every time. That is, if these forms have been signed once by the interviewed expert, it will cover all the future meetings I have with them. I will not be recruiting anyone under the age of 18, and vulnerable participants will be screened out.

The discussion would be informal without any structured approach. The researcher may or may not have a clear approach/plan in mind regarding the focus and goal of the discussion. The discussions tend to be open-ended and express little control over experts' responses. The researcher jots down the notes/points while the discussion is taking (hand-written). Since informal discussions occur 'on the fly', it is difficult to tape-record this type of interview. The researcher engages in the discussion to develop an understanding of the tasks and requirements. This type of discussions need to be included immediately in the researcher's field notes. The discussions will also help to uncover new areas or topics of interest that may have been overlooked by previous research. The discussions are highly informal because the researcher's understanding is still evolving, it is helpful to anticipate the need to speak with experts on multiple occasions.

In an effort to better understand, we aim to collect suggestions, comments and feedbacks from the Red Sift experts, by making notes in the diary book. Then, the researcher can elaborate on the points based on the conversation and observation with the experts (based on the hand-written notes). Finally, reflecting on the discussions, researchers can decide to have another discussion session or not. All the recorded notes will be transcribed, encoded and elaborated in detail in the report. Any additional notes or diagrams made/showed by participants will be digitised following the session. A thematic analysis of the notes will be completed after all sessions have taken place. The findings will be used to refine the design of the support materials.

In the interviews, we aim at focusing "How do E-discovery analysts/investigators gain insights from large and complex email data?. For task and data abstraction, we use the Why?, What?, How? framework [1] to abstract the tasks, explore visualisations and develop interaction paradigms that would satisfy these tasks. We will make it flexible based on the nature of the project.

As a starting point to discuss about the email communication analysis, the Enron real-case [2] will be unveiled and some of the following open-ended questions will be asked:
Q1: What are the challenges of E-discovery (E-mail communication investigations) with respect to visualisation?
Q2: How will you investigate on the key time-frame, key words and key individuals/players involved?
Q3: How will you categorize normal and abnormal E-mail communication data? How will you characterize suspicious behaviours?
Q4: How can visualisation inform unexpected behaviours? Do you think there are useful tools for investigations?

The above questions along with the "topic guide" (attached) questions/task can co-occur. The discussions in the series of interviews will help us understand the tasks and requirements in E-discovery that the analysts are interested in the email communication data investigations. The research questions along with the E-discovery requirements will help us abstract out some generalisable tasks and build visual solutions.

Other Considerations:

- **Anonymity:** no personal data will be collected and all other data captured will be reported in a completely anonymised format, with researchers having access only to their interview transcripts. Participant names will not be associated with the notes or any other data, and will not appear in any reports or presentations.
- **Consent:** we will be taking the consent from the experts before we begin the study, consisting of a series of interviews.
- **Confidentiality**: all the data will be kept private, both from other participants and when reporting findings. We will not be revealing the experts/analysts' names. Data will be held securely and kept confidential.
- **Disclosure:** the names of the participants and other personal information will not be recorded in the notebooks or the transcriptions. It will be completely anonymised. The company name is not anonymized and we have received the permission from the company to publish the same.
- **Permission:** all the recorded contribution, in written form, taken from the interview by the researcher, will be used in in line with the participant's preferences as recorded in our consent form.

- **Data Protection:** all the data will be password protected, stored securely, and backed up. Only myself, my supervisory team (Dr Cagatay Turkay and Prof Jason Dykes), and my examiners, will have access to the data. If a participant decides to withdraw from the study at any point, I will destroy any data already gathered from them.
- **Others:**

a) participants will be given a participant information sheet and be asked to consider and sign a consent form.
b) The names will not be recorded, we will use pseudonyms to retain anonymity in our notes.
c) no audio, video or screen capture will be recorded.
d) No personal information will be collected, identities will not be recorded or revealed.
e) The project will adhere to the requirements of data protection rules in terms of data labelling, storage and security relevant based on the current General Data Protection Regulation (GDPR).
f) Data will be stored for 10 years, and then destroyed (which is a standard), in locked filing cabinets and on password protected computers.

Dissemination

The project is designed to inform multiple stakeholders, including the academic community, governments and industries. In addition to the usual academic publications, all participants in the study will be offered access to the findings and resulting recommendations, via the school, in a form appropriate to their interests. Most likely this will be in the form of a short and accessible report providing an overview of findings and noting conclusions and any recommendations.

No discomfort or misrepresentation is anticipated as a consequence of the dissemination process. However, the researchers will take care to present the material in a manner that treats the reported experiences with respect. The outcome of the interviews along with our results will be published in journals, conferences papers as well as the researchers' thesis/report. The thesis will be made available on the City Research Online.

References

[1] T.Munzner. Visualization analysis and design. CRCPress, 2014.
[2] 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.

PARTICIPANT INFORMATION SHEET

Title of study: Visual Analysis of Email Communication Data within E-discovery

Principal Investigator: Mithileysh Sathiyanarayanan

Supervisor: Dr Cagatay Turkay

We would like to invite you to take part in a research study. Before you decide whether you would like to take part it is important that you understand why the research is being done and what it would involve for you. Please take time to read the following information carefully and discuss it with others if you wish. If there is anything that is not clear or if you would like more information, please ask us.

What is the purpose of the study?

I am PhD research student in the Computer Science department at the City, University of London.

The main aim of the research is to design and develop visual solutions to unravel the information in E-mail communication data to support E-discovery in an organisation by facilitating the generation of visual evidence for users/analysts. This will enable analysts to compare various communication features from multiple perspectives, identify relevant subsets of data and find anomalous communication behaviour. As visualisation tools and solutions continue to be improved and optimised, analysts are increasingly calling for novel techniques that can improve in identifying and understanding of various communication features to understand unusual behaviour in E-mail communication data. In this research, we will be developing techniques and implementing investigative strategies in software prototypes through a structured process of abstraction, design and testing, by using a well-known methodology called Design Study Methodology. Doing so is intended to explore and answer a series of research questions in ways that will improve the role of visualisation and visual evidence in E-discovery.

Design Study Methodology requires inputs from you and so we will need to engage with you to find out what you are trying to do and what you think of our solutions. This will help us understand your needs and requirements and the challenges you face in investigating E-mail communication data.

The engagement would be a biweekly (once in two weeks starting from 15th September 2018 to 30th May 2019. These regular interviews can lead to new insights and help in developing designs and prototypes.

Why have I been invited?
I am looking for experts who have at least some – but not extensive – experience of using visual analytic tool for Email analysis. You must also be at least 18 years old and be able to undertake the activities described below.

Do I have to take part?
The participation is voluntary, you may withdraw at any stage, or avoid answering questions that you feel to be too personal or intrusive.

It is up to you to decide whether or not to take part. If you do decide to take part you will be asked to sign a consent form. If so, you are still free to withdraw at any time and without giving a reason - you will not be penalised or disadvantaged in any way and any data we have collected from you will be destroyed. Once the data has been published in reports/publication, participants will no longer be able to withdraw their data.

What will happen if I take part?

- There will be iterative unstructured interviews
- Each interview will last for maximum one hour (at a time convenient for you)
- The interviews will happen twice per month
- I will take notes to capture your reactions and responses
- The interviews will take place at your company premises and at City, University of London.
- You will be providing your opinion on particular aspects of a number of tasks, paper-based designs or prototypes we have created. Ranking the designs according to your preferences will also be considered.

What is the process?

You will informed in detail about the goal of the research. An unstructured interview will be conducted to gather information about your general and technical requirements for analysing Email communication data within E-discovery domain.

There will be regular interviews with you to understand the challenges in the E-discovery and to understand your tasks, requirements and how you use visualisation. In all the sessions, we will aim at focusing "How do E-discovery compliance team and/or analysts/investigators gain insights from large and complex E-mail data?". In an effort to better understand, we will collect suggestions, comments and feedback from you. The discussions will be open-ended and directed by your responses and reactions. We will jot down key notes and points while the discussion goes on. We hope that the discussions will help to uncover new areas or topics of interest that may have been overlooked by previous research.

We will discuss your challenges for one hour every two weeks during the study. Sometimes we will meet at City, and sometimes on your company premises, as required and agreed in advance.

During these sessions we will discuss your analytical needs. Sometimes we will structure discussions with specific questions. On other occasions we may let you lead the discussion to try to establish areas that we can explore with our solutions. As things progress we may give you some examples of visualisation techniques or software solutions and ask for your views on them as they may apply to your needs. We may ask you to use software and observe you doing so. This may include software that you currently use to support your work, or prototypes that we develop to understand your data and your needs. We may ask you to talk aloud during this activity and we may prompt you to do so or ask questions as you use software. At no time will we record you, your use of systems or any screen that you are using. We will not take photographs or collect any imagery, video or audio. We will record your responses by making notes in a notebook during, and subsequent to, the sessions. Your names will not be recorded, we will use pseudonyms to retain anonymity in our notes.

What are the possible benefits of taking part?
By taking part in this study you will contribute to our knowledge of how to design support for users of email analysis.

What will happen to the results of the research study?
We will use the information collected in the interviews and the reactions to the software to design, build and test systems that help support e-discovery tasks in order to answer our higher level questions about how visualization can support e-discovery.
This study will be written up to form part of my PhD thesis and the results will inform future studies. I also hope to submit my findings for academic publication in the journals and conferences that we use to inform others about the knowledge we have created and the techniques we have developed. The project is designed to inform multiple stakeholders, including the academic community, governments and industries. In addition to the usual academic publications, all participants in the study will be offered access to the findings and resulting recommendations, via the school, in a form appropriate to their interests. Most likely this will be in the form of a short and accessible report providing an overview of findings and noting conclusions and any recommendations.

What will happen when the research study stops?
Data will be stored for 10 years - the standard retention time for university studies - and then destroyed.

Who has reviewed the study?
This study has been approved by CSREC, the Computer Science Research Ethics Committee at City, University of London.

Further information and contact details

Mr Mithileysh Sathiyanarayanan
PhD Researcher
Address: A401, College Building
City, University of London
Northampton Square
London
EC1V 0HB
United Kingdom
Email:
mithileysh.sathiyanarayanan@city.ac.uk

Dr Cagatay Turkay
Project Supervisor
Address: A401b, College Building
City, University of London
Northampton Square
London
EC1V 0HB
United Kingdom
Email: cagatay.turkay@city.ac.uk
Phone: +44 (0)20 7040 84

Data Protection Privacy Notice: What are my rights under the data protection legislation?

City, University of London is the data controller for the personal data collected for this research project. Your personal data will be processed for the purposes outlined in this notice. The legal basis for processing your personal data will be that this research is a task in the public interest, that is City, University of London considers the lawful basis for processing personal data to fall under Article 6(1)(e) of GDPR (public task) as the processing of research participant data is necessary for learning and teaching purposes and all research with human participants by staff and students has to be scrutinised and approved by the City's Computer Science Research Ethics Committee (CSREC).

What if I have concerns about how my personal data will be used after I have participated in the research?

In the first instance you should raise any concerns with the research team, but if you are dissatisfied with the response, you may contact the Information Compliance Team at dataprotection@city.ac.uk or phone 0207 040 4000, who will liaise with City's Data Protection Officer Dr William Jordan to answer your query.

If you are dissatisfied with City's response you may also complain to the Information Commissioner's Office at www.ico.org.uk

What if there is a problem?

If the research is undertaken in the UK if you have any problems, concerns or questions about this study, you should ask to speak to a member of the research team. If you remain unhappy and wish to complain formally, you can do this through City's complaints procedure. To complain about the study, you need to phone 020 7040 3040. You can then ask to speak to the Secretary to Senate Research Ethics Committee and inform them that the name of the project is: **Visual Analysis of Email Communication Data within E-discovery**

You could also write to the Secretary at:
Anna Ramberg
Research Governance & Integrity Manager
Research & Enterprise
City, University of London
Northampton Square
London
EC1V 0HB
Email: Anna.Ramberg.1@city.ac.uk

City holds insurance policies which apply to this study. If you feel you have been harmed or injured by taking part in this study you may be eligible to claim compensation. This does not affect your legal rights to seek compensation. If you are harmed due to someone's negligence, then you may have grounds for legal action.

Thank you for taking the time to read this information sheet.

CONSENT FORM

Title of Study: *Visual Analysis of Email Communication Data within E-discovery*

Please initial box

1.	I confirm that I have had the project explained to me, and I have read the participant information sheet, which I may keep for my records. I have been given the opportunity to ask questions and have had them answered to my satisfaction. I understand this will involve the following: • being interviewed by the researcher • being observed by the researcher • allowing the interview to use a computer/book to make notes • completing questionnaires related to the project • evaluating the tasks, designs and the prototypes	
2.	This information will be held by City as data controller and processed for the following purpose(s): • PhD research • PhD assessment and examination • Research publications and presentations	
3.	I understand that any information I provide is confidential, and that no information that could lead to the identification of any individual will be recorded. As such no personal data will be disclosed in any reports on the project, or to any other party. No identities will be published. I understand that the thesis produced from this work will be made available in the City Research Online repository.	
4.	I understand that my participation is voluntary, that I can choose not to participate in part or all of the project, and that I can withdraw at any stage of the project without being penalized or disadvantaged in any way.	
5.	I agree to City, University of London recording and processing this information about me. I understand that this information will be used only for the purpose(s) set out in this statement and my consent is conditional on City complying with its duties and obligations under the General Data Protection Regulation (GDPR).	
6.	I agree to the arrangements for data labelling, data storage, archiving, sharing & security relevant based on the current GDPR rules and regulations. The details are provided in the Participant Information Sheet.	
7.	I agree to the use of anonymised quotes in publication.	
8.	I agree to take part in the above study.	

_____	_____	_____
Initials of Participant	Signature	Date
_____	_____	_____
Name of Researcher	Signature	Date

TOPIC GUIDE FOR UNSTRUCTURED ITERATIVE INTERVIEW

Main Questions	Additional Questions	Prompting for Details
• What are the challenges of E-discovery and investigations in general and with respect to visualisation? • What do you think about the role of visualisation in E-mail communication data within E-discovery? • What are the visualisation requirements analysts are calling for? • What type of visualisations to consider in the project? • What are the tasks carried out in E-discovery and investigations with respect to E- mail communication data? • Discuss different types of case studies for the project set-up. • What do you think about the initial designs/prototypes? • What do you think about the iterated designs/prototypes?	• Clarify what adaptations / changes were made e.g. changes to email analysis • Clarify what makes it difficult for the participant to use his / her current tools • Clarify what strategies the participant use to overcome those difficulties ... • Clarify what makes it easier for the participant to use current tools • Any suggestions/advice from the participant to help others not to use current tools • Clarify what reminders / cues / routines that the participant has tried ... • Clarify what reminders / cues / routines work best for the participant ... • Clarify what needs to be considered for the designs/prototypes? • Any perceived advantages of a particular reminders / cues / routines? • Any perceived disadvantages of a particular reminders / cues / routines?	• Can you please clarify what you meant by ...? • Can you please expand a little on ...? • Can you please give some examples of ...? • In particular, what do you think of ...?

A.6 Samples of the Note Taking

A.7 Thematic Analysis

Domain Characterisation

- We conducted several meetings/interviews (unstructured) with the Red Sift analysts to understand requirements and what they expect in visualisation
- We were making regular notes in our diary book. Later it was transcribed straight after the interview to consider any clarification. This process was carried out on Google Documents (by sharing it with my supervisors).
- We focussed on the coding and thematic analysis to identify themes.
- The methodology is described in Chapter 2.
- The Domain Characterisation is discussed in Chapter 4.

Encoding Meaning

ABCDEFGHIJK - Challenges

ABCDEFGHIJK – Design Expectations/Requirements

ABCDEFGHIJK – Guiding Observations

ABCDEFGHIJK – E-discovery Tasks

ABCDEFGHIJK – Techniques

Phase 1 (Jan 2016 - June 2016)

Tried to understand the basic idea of the Email Communication
Tried to understand their basic idea of the visualisations
Tried to understand the nature of the work carried out by Red Sift
Who are the users?

- Individuals?
- Analysts?

What are potential tasks for the users we identified (i.e., what do you want them to do)?
In what form will the results be best presented?
What sorts of data we might expect or would like to compute?

The E-discovery domain was new to us and we were keen to understand this a bit more. The Red Sift team explained the nature of their projects and the current visualisations they are working on.

One of their Email Communication Projects is “Personal Analytics” - to analyse one’s own inbox to identify their travel taxi receipts.

Some open-ended questions were presented

1. What are the challenges of E-discovery and investigations in general ?
2. What do you think about the role of visualisation in E-mail communication data within E-discovery?
3. What are the tasks carried out in E-discovery and investigations with respect to E-mail communication data? How will you investigate on the key time-frame, key words and key individuals/players involved?

Discussed E-discovery domain and the concept of E-discovery related to E-mail analysis was expressed.

The concept of email anomalies in the communication was discussed.

E-discovery plays an important role in investigating organisation’s email communication.

E-discovery requests are mostly conducted by Compliance Officer, Freedom of Information (FoI) Officer, Legal Counsel (E-discovery/legal officer), Human Resource officer, and/or IT Director/Manager.

The experts (E1) mentioned, ``some of the challenges of E-discovery and investigations are: the process is very expensive, time-consuming, complex and tedious, difficult to compare and identify/detect unusual communication behaviour, difficult to explore freely to identify changes and find some interesting behaviours related to a particular case".

Volume of the emails is the biggest challenge. Filtering and searching for connections and information to find interesting subsets of data needs to be addressed.

One of the experts (E2) mentioned ``Few legal experts and advanced users use E-discovery tools such as Jigsaw, Concordance by LexisNexis and/or IN-SPIRE to analyse electronic documents but for electronic mail data, we use only manual searching and excel for analysis. In many investigation cases, most of the E-discovery experts often do not know what they are looking for in their data which can be highly time-consuming to find relevance/interesting information and present it in legal proceedings”.

In organisation, E-discovery compliance is important to find various relationships in the data. In this project, we focus only on digital communication data, specifically E-mail communication data. Under this, we will focus on E-mail compliance. From investigation, we will work on Information discovery (email discovery).

Detecting changes to find interesting communication patterns in the complex and dynamic nature of emails can be a challenge.

Open-ended data exploration to find interesting information will be helpful and useful in navigating and finding relevant information for the cases.

There is a difference between ``finding'' and ``discovering'' information in the email communication data.

Exploring and understanding what makes a time period interesting from different dimensions. Also determining, identifying interestingness in a subset of time periods.

The potential case studies could be Enron Case and Hillary Clinton Case.We can use their data for the analysis

The visualisation requirements analysts are calling for:

1. simple, easy to use and quickly accessible. Basic charts like bars, line, pie, dot-line and matrix charts might help in understanding patterns quickly.
2. easy to understand, explore, search, detect and analyse
3. a means to discover and understand investigative stories, and then to present them to others (serve as visual evidence).
4. interactive and analyst-friendly
5. effective for displaying relationships and filtering subsets of data

Red Sift experts require Data-driven Documents (D3.js) for web-based data visualisations (basic charts such as bar graphs, line graphs, area graphs etc.). It is fine to start with paper sketches and quickly drag-and-drop visualisation tools for testing.

It might be good to have an overview to understand the semantics of time periods, individuals and find the structure of the emails exchanged over time. It must support in exploration, comparing and summarising email patterns across all temporal dimensions.

Explore and understand what makes an individual interesting. Then determine, visualise and identify interestingness in a subset/cohort of individuals.

The current E-discovery model is complicated and tedious to manually investigate and find some important or interesting points for the case. The organisations involved in the litigation need to hand-over the email raw data to the E-discovery experts to analyse. We must be able to provide a simpler model where experts can carry our their analysis within our platform and find the time, individuals and other key information. This becomes simple, time-saving, significantly cheaper and consumes less man power.

All the three experts mentioned the manual string search for E-mail investigation/analysis is strenuous, time-consuming and huge costs are involved.

The expert (E1), mentioned the other challenges could be``effective search techniques to find an important information that can improve accuracy",

The expert (E2), mentioned, there is also a challenges such as``effective strategies to find pertinence that can minimise cost and time"

Phase 2 (Jan 2017 - April 2017)

We need an overview to understand the individuals communication behaviour and find the structure of the emails exchanged over time. It must also support in exploration, comparing and summarising email patterns such as sending and receiving.

The tasks carried out in E-discovery and investigations with respect to E- mail communication data are:

- Is there any interesting trend/pattern before and after Oct 2001? how many messages were recorded and between whom during a particular time?
- Support a multi-faceted decision - eg., We need to consider Oct 2001 as that period had more interesting strong relationships.
- Are E-mails on track for this quarter?
- Explain and understand change in email communication behaviour
- What kind of topics employees/employers talk about with various email contacts?

Explore and understand threads and find what makes a conversation interesting. Classify the communication types based on the email patterns.

Phase 3 (Jan 2018 - April 2018)

Similar to the other prototypes, it will be nice to have an overview of threads which can help in getting a high-level view of the thread structures. Again, we need to see some comparison between the threads, understand the activities of individuals in a single thread. It might be interesting to find different email communication types within a thread and as a group of threads.

Some interesting tasks that will be useful in the investigation cases:

- Characterising clusters of communication and their changes over time and over multiple scales
  - Observing changes in communication groups
  - Groups might be different domains,
- Visual depiction of communication threads over time

- This relates to the depiction of individual email threads, e.g., who joins, who drops and when, etc.
- Will need to eventually relate to the above
- Validating/comparing the given structure of an organisation (schema)

Encodings	Coded Lines	Extracted Themes
Challenges	Some of the challenges of E-discovery and investigations are: the process is very expensive, time-consuming, complex and tedious, difficult to compare and identify/detect unusual communication behaviour, difficult to explore freely to identify changes and find some interesting behaviours related to a particular case Volume of the emails is the biggest challenge. Filtering and searching for connections and information to find interesting subsets of data needs to be addressed. Few legal experts and advanced users use E-discovery tools such as Jigsaw, Concordance by LexisNexis and/or IN-SPIRE to analyse electronic documents but for electronic mail data, we use only manual searching and excel for analysis. In many investigation cases, most of the E-discovery experts often do not know what they are looking for in their data which can be highly time-consuming to find relevance/interesting information and present it in legal proceedings. Detecting changes to find interesting communication patterns in the complex and dynamic nature of emails can be a challenge. Open-ended data exploration to find interesting information will be	1. Finding interesting subsets within the large volume of data 2. Complex and dynamic nature of communication patterns 3. Open-ended data exploration to find interesting communication patterns.

	helpful and useful in navigating and finding relevant information for the cases. effective search techniques to find an important information that can improve accuracy effective strategies to find pertinence that can minimise cost and time	
Design Expectations/Requirements	It might be good to have an overview to understand the semantics of time periods, individuals and find the structure of the emails exchanged over time. It must support in exploration, comparing and summarising email patterns across all temporal dimensions. We need an overview to understand the individuals communication behaviour and find the structure of the emails exchanged over time. It must also support in exploration, comparing and summarising email patterns such as sending and receiving. Similar to the other prototypes, it will be nice to have an overview of threads which can help in getting a high-level view of the thread structures. Again, we need to see some comparison between the threads, understand the activities of individuals in a single thread. It might be interesting to classify and find different email communication types within a thread and as a group of threads.	1. Explore temporal feature 2. Explore individuals feature 3. Compare temporal characteristics 4. Compare individual characteristics 5. Understand activities in a temporal selection 6. Understand activities in an individual selection 7. Explore thread feature 8. Compare thread characteristics 9. Understand activities in a thread selection 10. Classify thread communication types

Guiding Observations	The current E-discovery model is complicated and tedious to manually investigate and find some important or interesting points for the case. The organisations involved in the litigation need to hand-over the email raw data to the E-discovery experts to analyse. We must be able to provide a simpler model where experts can carry our their analysis within our platform and find the time, individuals and other key information. All the three experts mentioned the manual string search for E-mail investigation/analysis is strenuous, time-consuming and huge costs are involved. One of the experts mentioned “Few legal experts and advanced users use E-discovery tools such as Jigsaw, Concordance and/or Inspire to analyse electronic documents but for electronic mail data, we use only manual searching and excel for analysis. In many investigation cases, most of the E-discovery experts often do not know what they are looking for in their data which can be highly time-consuming to find relevance/interesting information and present it in legal proceedings”. In organisation, E-discovery compliance is important to find various relationships in the data. In this project, we focus only on digital communication data, specifically E-mail communication data. Under	1. Characterising the Domain 2. Area of focus 2. Visualisation Requirements

	this, we will focus on E-mail compliance. From investigation, we will work on Information discovery (email discovery). There is a difference between ``finding'' and ``discovering'' information in the email communication data. The visualisation requirements analysts are calling for: 1. simple, easy to use and quickly accessible. Basic charts like bars, line, pie, dot-line and matrix charts might help in understanding patterns quickly. 2. easy to understand, explore, search, detect and analyse 3. a means to discover and understand investigative stories, and then to present them to others (serve as visual evidence). 4. interactive and analyst-friendly 5. effective for displaying relationships and filtering subsets of data	
E-discovery Tasks	Exploring and understanding what makes a time period interesting from different dimensions. Also determining, identifying interestingness in a subset of time periods. The tasks carried out in E-discovery and investigations with respect to E-mail communication data are: Is there any interesting trend/pattern before and after Oct 2001? how many messages were recorded and between whom during a particular time?	1. Discover and characterise time periods of interest 2. Discover and characterise individuals of interest 3. Discover and characterise threads/conversations of interest

	Support a multi-faceted decision - eg., We need to consider Oct 2001 as that period had more interesting strong relationships. Are E-mails on track for this quarter? Explain and understand change in email communication behaviour What kind of topics employees/employers talk about with various email contacts? Explore and understand what makes an individual interesting. Then determine, visualise and identify interestingness in a subset/cohort of individuals. Explore and understand threads and find what makes a conversation interesting. Classify the communication types based on the email patterns. Some interesting tasks that will be useful in the investigation cases: - Characterising clusters of communication and their changes over time and over multiple scales - Observing changes in communication groups - Groups might be different domains, - Visual depiction of communication threads over time - This relates to the depiction of individual email threads, e.g., who joins, who drops and when, etc. - Will need to eventually relate to the above - Validating/comparing the given structure of an organisation (schema)	

Techniques	Red Sift experts require Data-driven Documents (D3.js) for web-based data visualisations (basic charts such as bar graphs, line graphs, area graphs etc.). It is fine to start with paper sketches and quickly drag-and-drop visualisation tools for testing.	Design Process: Step 1: Paper Sketches Step 2: Existing Visualisation Tools Step 3: D3.js to build our own tool.
Case Studies & Datasets to use	The potential case studies could be Enron Case and Hillary Clinton Case. We can use their data for the analysis	Enron & Hillary Clinton Email Datasets & use case.

Design Themes (re-ordered)	Design Requirements (mapped one to one)
1. Explore temporal feature 2. Compare temporal characteristics 3. Understand activities in a temporal selection 4. Explore individuals feature 5. Compare individual characteristics 6. Understand activities in an individual selection 7. Explore thread feature 8. Compare thread characteristics 9. Understand activities in a thread selection 10. Classify thread communication types	1. Investigate high-level temporal characteristics. 2. Compare multiple time periods / temporal dimensions. 3. Understand activities in a temporal selection 4. Investigate high-level individual characteristics. 5. Compare multiple individual connections 6. Understand activities of individuals 7. Investigate high-level thread characteristics 8. Compare multiple threads 9. Understand activities in a thread 10. Specify thread communication types

Design Themes & Analysis Goals Themes (re-ordered)	Analytical Tasks
Design Themes: 1. Explore temporal feature 2. Compare temporal characteristics 3. Understand activities in a temporal selection 4. Explore individuals	Temporal Analysis (Design Phase 1) 1. Explore E-mail communication patterns (activities) of time periods of interest that differ between the following and find if it is interesting from different perspectives: years, months, days, hours, weekends and weekdays, mornings and nights. 2. Identify interestingness in the temporal gaps using a subset of time periods. Therefore, if the analysts/users know some events

feature 5. Compare individual characteristics 6. Understand activities in an individual selection 7. Explore thread feature 8. Compare thread characteristics 9. Understand activities in a thread selection 10. Classify thread communication types Analysis Goals Themes: 1. Discover and characterise time periods of interest 2. Discover and characterise individuals of interest 3. Discover and characterise threads/conversations of interest	they can easily relate those to them (e.g. weekends, holidays, trips, etc.). 3. Understand and investigate the changes in the volume of emails over time and also assess whether the changes are indeed unusual. Individuals Analysis (Design Phase 2): 4. Explore E-mail communication patterns (activities) of individual(s) of interest with others over time and find if it is interesting from different perspectives: sent, received, sent and received. 5. Identify interestingness in the communication (contact/relationship) of individual(s) of interest (using a subset/cohort of individuals). 6. Understand and investigate the changes in the communication of individuals and also assess whether the changes are indeed unusual. Thread Analysis (Design Phase 3): 7. Explore E-mail communication patterns (activities) of threads/conversations of interest and find if it is interesting from different perspectives: based on time, based on individuals (senders/receivers, inclusion/exclusion, active/passive), based on thread types. 8. Identify interestingness in the thread(s) of interest (conversations). 9. Understand and investigate the changes in the conversation/thread characteristics and also assess whether the changes are indeed unusual. 10. Compare multiple threads to understand individual behaviour.

Final Table (mapping of Challenges, Design Requirements, Analysis Goals and Tasks)

Challenges (C)	Design Requirements (R)	Analysis Goals (AG)	Tasks (T)
C1. Finding interesting subsets within the large volume of data C2. Complex and dynamic nature of communication patterns C3. Open-ended	R1. Explore temporal feature R2. Compare temporal characteristics R3. Understand activities in a temporal selection	AG1. Discover and characterise time periods of interest	T1. Explore E-mail communication patterns (activities) of time periods of interest that differ between the following and find if it is interesting from different

data exploration to find interesting communication patterns.			perspectives: years, months, days, hours, weekends and weekdays, mornings and nights. T2. Identify interestingness in the temporal gaps using a subset of time periods. Therefore, if the analysts/users know some events they can easily relate those to them (e.g. weekends, holidays, trips, etc.). T3. Understand and investigate the changes in the volume of emails over time and also assess whether the changes are indeed unusual.
C1. Finding interesting subsets within the large volume of data C2. Complex and dynamic nature of communication patterns C3. Open-ended data exploration to find interesting communication patterns.	R4. Explore individuals feature R5. Compare individual characteristics R6. Understand activities in an individual selection	AG2. Discover and characterise individuals of interest	T4. Explore E-mail communication patterns (activities) of individual(s) of interest with others over time and find if it is interesting from different perspectives: sent, received, sent and received. T5. Identify interestingness in the communication (contact/relationship) of individual(s) of interest (using a subset/cohort of individuals). T6. Understand and investigate the changes in the

			communication of individuals and also assess whether the changes are indeed unusual.
C1. Finding interesting subsets within the large volume of data C2. Complex and dynamic nature of communication patterns C3. Open-ended data exploration to find interesting communication patterns.	R7. Explore thread feature R8. Compare thread characteristics R9. Understand activities in a thread selection R10. Classify thread communication types	AG3. Discover and characterise threads/conversations of interest	T7. Explore E-mail communication patterns (activities) of threads/conversations of interest and find if it is interesting from different perspectives: based on time, based on individuals (senders/receivers, inclusion/exclusion, active/passive), based on thread types. T8. Identify interestingness in the thread(s) of interest (conversations). T9. Understand and investigate the changes in the conversation/thread characteristics and also assess whether the changes are indeed unusual. T10. Compare multiple threads to understand individual behaviour.

Design & Validation

- We conducted several meetings/interviews (unstructured) with the Red Sift analysts to understand designs (improve) and validate constantly.
- We were making regular notes in our diary book. Later it was transcribed straight after the interview to consider any clarification. This process was carried out on Google Documents (by sharing it with my supervisors).
- We focussed on the coding and thematic analysis to identify themes.
- The methodology is described in Chapter 2.
- The Design & Validation is discussed in Chapter 5.

Encoding Meaning

ABCDEFGHIJK – Visualisation Techniques

ABCDEFGHIJK – Temporal Analysis (Features / Design / Implementation)

ABCDEFGHIJK – Individual Analysis (Features / Design / Implementation)

ABCDEFGHIJK – Thread Analysis (Features / Design / Implementation)

ABCDEFGHIJK – Guiding Observations

Phase 1 (June 2016 - Dec 2016)

How to identify the role of a particular person or a group in email discussions over time and how to detect the events?

High-level tasks to design for:

- Analyst interface
  - Multiple aggregation levels
    - Aggregate by individual
    - Aggregate by department
    - Aggregate by temporal scales
    - Aggregate by threads
  - Comparison
    - Mechanics of comparing equal/unequal sets (data cubes?)
    - Identifying differences
    - Compare various facets in the data

Agree on a case-study as a motivation - Decided to go with Enron data
Who is the user (analyst) - Red Sift experts
What are we trying to find?

- interesting patterns
- trends, changes

How to identify and detect changes?
What are the various approaches/techniques used in investigating this case-study?
Which are the vis tools implemented for investigating this case-study? Any success? Identify the strengths and weaknesses.
Can we map them to the high-level tasks?
In the paper prototypes, consider bar graphs and heat matrices.
Multi-granularity in the Tableau design

Consistent **feature generation (e.g., aggregation, filtering)/interaction/visualisation** for **time/individual/feature** operations

- Defining the norm (another feature?), there can be several norms and we need to account for that
- **Normalisation according to norm/baseline** across all data facets
- Flexible **comparison framework** to handle multiple different **granularities of time/individual/feature**
- Generated features as **externalised knowledge artefacts** which can:
  - be applied **across datasets**
  - be **transferred** to different analytical tasks
  - be instrumental as **evidence**
  - serve as **starting templates** for new analysis rounds
- Discuss the E-discovery expert workshop insights.
- To discuss further on the paper prototypes designed (aggregation and comparison: drill-down approach). Since the experts are in strong favour of time-series and bar graphs for email investigation purposes, how do we go about getting the first version of the functioning prototype?
- How do we define normal behaviour (behaviours in departments/individuals)?
- Understand deeply what an E-discovery analyst will be looking for?

(a) What will be an interesting behaviour an analyst will be looking for?
(b) What kind of patterns an analyst will be looking for? And how we can support it?
(c) How would an analyst define a particular contact?
(d) How the investigation can be presented to the court?
Discuss progress on mock ups for discussion with the experts
Agree on a date to have initial version that we can review internally and iterate
Summarise on the questions identified from the previous meetings.
Discuss on the multi-modal and multi-level approach.
Discuss the first version of the D3 visuals generated using the real data.

Use basic visualisations such as bars, graphs, matrices, scatter and other charts to display information.

Aggregate by temporal scales, compare different timescales which can include individuals too (sent & received).

Aggregate by individual
Aggregate by department/organisation roles
Compare different individuals (based on sent, received and combination of both)
Consider time, individual, engagement and context.

Improve comparison of two or more subsets of data. Some of the current techniques/approaches does not aid in supporting various features in comparing subsets of multi-faceted data.
The E-discovery analysts have difficulty in defining anomalies/abnormalities.
We might need a tool with better exploration facility (better filters). The E-discovery analysts have difficulty in exploring large datasets and they have become a big concern due to navigation issue, especially for communication data. In our case, email communication data.
We can consider the possible Exploration tasks

- Which months, days and hours were the busiest (emails were sent)?
- What happened to the email communication of individual 'A between 2000 and 2001?
- Are emails being sent on the weekend?
- How does time-of-day correlate with emails being sent?
- What happened on October 21 2001?

The time can be broken down into years, months, days, weeks and hours.
Consider time, individual, engagement and context.
For discovering temporal characteristics
1. Main Temporal View
   - Years
   - Months
   - Days
   - Days of the Week
   - Hours
2. Bar View (to support in navigation)
   - Years
   - Months
   - Days
   - Days of the Week
   - Hours
3. Individuals View

- Senders
- Receivers

4. Content to be displayed

The expert (E1) “You can do self validation to check if the solution is working and find anything you can interesting based on the Enron case (from literature).”

The expert (E1) “Doing this saves time in conducting empirical/user studies. So, personally validate the tool, report on findings, then quickly walkthrough the case with us. If it is working good, then we can deploy a solution immediately to check for engineering issues.”

The expert (E1) says
Good use of multi-facetedness for exploration and multi-granular views
Good use of small multiples to visualise years, months, days, days of the week.
This helps to identify areas for further analysis, such as peak periods of activity (patterns/trends) and temporal gaps. Also, the supporting view (bar charts) helps in filtering and comparing different subsets of data.

The expert (E1) came up with questions:
``Can you select Richard Shapiro and find his connections?"
``Can you randomly search for an individual in the organisation and find to whom most messages were sent and received?".

We might need to merge the current data with the organisation roles or designations. Do you think we have something existing?

The tool seems to be useful and this can be connected with a GMAIL account so we can test for our business.

Phase 2 (April 2017 - Dec 2017)

For thread analysis:

- **Metrics** that one can use to make these views much more interesting
    - **CC--trails** for individuals
    - Weekend emails
    - Thread-size
- Lawyers want to make sure that they do not miss anything
- Combining the changes in the email trends with external data sources — Enron stock values
- Try out the “bespoke” metrics and observe patterns

- Bespoke visualisation for a single thread

First version of the bar-graph interaction using Crossfilter was demonstrated and then the technical feedback was noted.
Some of the visualisation requirements could be better aesthetics, less clutter, informative, better interaction (filtering capabilities) & usability.
The prototype must be able to address some of the questions noted in the meetings & consider the visualisation requirements.
Small multiples approach can be used for comparing temporal relationships. Small multiples can be a good approach for comparing individual's relationships too. The small multiples can help in understanding thread features.

Suggestions from the expert:
1. Merge sent and received messages in one visualisation, in the high fidelity prototype, will help in understanding the activities and communication patterns better.
2. Consider the departments/groups in an organisation will help in understanding the activities and communication patterns better.
3. Build an overview with aggregated statistics will help in refining and finding interesting individuals.
4. Consider a view that will help in understanding the volume of emails sent and/or received.

For discovering individual characteristics
1. Main Individual's View
    - Senders
    - Receivers
    - Both together
2. Another View
    - Sender breakdown info
    - Receiver breakdown info
3. Statistics View
4. Content to be displayed

The suggestions from the expert (E2) during the informal feedback on medium fidelity prototypes are,

- Merge sent and received messages in one visualisation, in the high fidelity prototype, so it will help in understanding the activities and communication patterns better.
- Consider the departments/groups in an organisation that will help in understanding the activities and communication patterns better.
- Build an overview with aggregated statistics that will help in refining and finding interesting individuals.
- Consider a supporting view that will help in understanding the volume of emails sent and/or received.

5. Include Organisation roles view
6. Different Email domains view (private such as gmail, yahoo etc and organisation like Enron)

Work closely with the engineers with the industrial partner to develop the prototypes further
Develop a functional prototype to utilise in targeted workshops to gather further structured feedback
Develop a number of case-studies to demonstrate the effectiveness of the initial designs
Iteratively refine and develop the visualisation and interaction designs
Prepare a set of questions that can be addressed using our prototypes.
Design a use case story (in a narrative way) to validate our approach/technique.
Discuss the R analysis - Used UpSetR, a more scalable alternative to Venn and Euler Diagrams for visualising intersecting sets using a novel matrix design, along with visualisations of several common set, element and attribute related tasks.
Demonstrate the real case study (Enron Fraud Case) that has temporal, individuals and context using D3.js, DC.js and Crossfilter (for interaction).

The expert (E1) again liked the idea of the multi-faceted exploration and multi-granularity approach. The use of bubble matrix to visualise relationships between the selected individual and their connections is helpful.
This can help analysts to identify interesting points and seamlessly switch between the different levels of main view and the supporting views.
The aggregated statistics view helps. The representation of organisation roles adds value to the individual analysis.

The expert (E1) came up with questions such as
``Can you find who are the senders and receivers in a thread?"
``Can you find which individuals have been excluded and included back? "
The tool can help in visualising individual connections and this can be tested to see our business connections over a period of time.

Phase 3 (April 2018 - Dec 2018)

Discussion on Thread Analysis

1. What are the different communication patterns that can be observed in the E-mail communication within an organisation?
2. What makes a thread relevant and/or interesting?
3. The core objective is to find different patterns in a thread
    - Classification rules for communication within threads:

        - **Binary Info:** x = 1 sender , y = 1 recipient
        - **Tertiary Info:** x = 1 sender , y = 2 recipients
        - **Denary Info:** x = 1 sender , 10 > y >= 3 recipients
        - **Quintary Announcement:** x = 1 sender , 50 > y >= 11 recipients
        - **Mass Announcement:** x = 1 sender, y >= 51 recipients
    - Criteria for classifying threads:
        - # senders / # receivers
        - # active / # passive
        - # inclusion / # exclusion
        - # unidirectional / # bidirectional
4. We might need an Overview and a detailed thread view
5. We discuss prototype progress (both Tableau & D3)
6. Discuss hard thresholds set and computation

Aggregate by threads
Compare different threads
Consider types of correspondence (cc, bcc)
Consider types of engagement
Find the communication types

Generate statistics from the threads (thread metrics), some ideas:

- Pace of interaction
- # of people (active/passive)
- # of people (same organisation/different/personal)
- # of inclusion/exclusion
- sender diversity (# of unique senders / # of all involved)
- thread length (in terms of duration (e.g., 5 days?) / in terms of message count)

Look for topologies/patterns of communication

- Broadcasting (announcement)
- Information
- Ping-pong
- Forward-sequences
- Loop
- Short discussions
- Bursty/Long discussions

For discovering thread characteristics

- Threads Overview (Dashboard)
    - Thread sizes (length)

    - No. of active individuals
    - All threads (thread versus time + only active individuals + selection of time-period)
    - ThreadStats View (Email replies, Entities of groups, etc)
- Single Thread View
    - Each individual thread to be viewed
    - Identify active & passive individuals
    - Identify inclusion & exclusion of individuals
    - Sort by time/chronology & engagement
    - CC/BCC--trails for individuals
    - “Visual Signatures” – compressed form of one full-thread.
    - Try out the “bespoke” metrics and observe patterns
- Thread Features View (with metrics)
    - Pace of interaction
    - Identify active & passive individuals
    - Identify inclusion & exclusion of individuals
    - Sort by time/chronology & engagement
    - # of people (active/passive)
    - # of people (same organisation/different/personal)
    - # of inclusion/exclusion
    - sender diversity (# of unique senders / # of all involved)
    - thread length (in terms of duration (e.g., 5 days?) / in terms of message count)
- Content View (complete message)

- How would you characterise individuals by observing their communication behaviour?
    - Type of threads they are involved in
        - Type of communication
        - Type of engagement
        - Type of correspondence
        - Type of individuals
    - Frequency and the temporal distribution of emails
    - Textual context
- Implicitly show some patterns or inclusion/exclusion
- Heuristics with threads
- cc & bcc to be expressed in the existing prototype.
- highlight when an individual is added back (at which point in time, the reply/email was).
- collapse redundancies (compress same patterns).

- we need more metrics to improve the visualisation (one message/total. n. of messages).
- How to identify branching?
- Select active & passive communicators.
- Select included & excluded individuals
- include different types of communication
- Comparison of multiple threads

The expert (E1) questions on using topic analysis
``Can you find what kind of topics the senders and receivers discussed in a thread?"
``Can you find which individuals in a thread discussed a topic on California Energy Deal? ".

The tool looks overall good and promising for investigation.

Encodings	Coded Lines	Extracted Themes
Visualisation Techniques	Small multiples approach can be used for comparing temporal relationships. Small multiples can be a good approach for comparing individual's relationships too. The small multiples can help in understanding thread features. To discuss further on the paper prototypes designed (aggregation and comparison: drill-down approach). The experts are in strong favour of time-series and bar graphs Discuss on the multi-modal and multi-level approach. Discuss the first version of the D3 visuals generated using the real data. Use basic visualisations such as bars, graphs,	Use basic visualisation techniques such as bar graphs, line charts, matrix charts, scatter plots. Use Small Multiples Approach Need aggregation and comparison Use Drill-down approach Use Multi-modal approach Use Multi-level approach Use D3 for building prototypes Feature Engineering for deriving thread features Statistical Computation with visual analysis Active Learning Bespoke metrics & visulisation

	matrices, scatter and other charts to display information. Consistent feature generation (e.g., aggregation, filtering)/interaction/visualisation for time/individual/feature operations	
Features / Design / Implementation (Phase 1)	In the paper prototypes, consider bar graphs and heat matrices. Multi-granularity in the Tableau design Aggregate by temporal scales, compare different timescales which can include individuals too (sent & received). The time can be broken down into years, months, days, weeks and hours. Consider time, individual, engagement and context. 1. Main Temporal View ● Years ● Months ● Days ● Days of the Week ● Hours 2. Bar View (to support in navigation) ● Years ● Months ● Days ● Days of the Week ● Hours	Question: To what extent visualisation can support analysts in discovering interesting temporal information in the E-mail communication data? Paper designs Tableau designs D3 designs Main view with heat matrices (temporal) Supporting views with heat matrices (individuals) Validation

	3. Individuals View ● Senders ● Receivers 4. Content to be displayed The expert (E1) says Good use of multi-facetedness for exploration and multi-granular views Good use of small multiples to visualise years, months, days, days of the week. This helps to identify areas for further analysis, such as peak periods of activity (patterns/trends) and temporal gaps. Also, the supporting view (bar charts) helps in filtering and comparing different subsets of data. The expert (E1) came up with questions: ``Can you select Richard Shapiro and find his connections?" ``Can you randomly search for an individual in the organisation and find to whom most messages were sent and received?". We might need to merge the current data with the organisation roles or designations. Do you think we have something existing?	

	The tool seems to be useful and this can be connected with a GMAIL account so we can test for our business.	
Features / Design / Implementation (Phase 2)	Aggregate by individual Aggregate by department/organisation roles Compare different individuals (based on sent, received and combination of both) Consider time, individual, engagement and context. Suggestions from the expert: 1. Merge sent and received messages in one visualisation, in the high fidelity prototype, will help in understanding the activities and communication patterns better. 2. Consider the departments/groups in an organisation will help in understanding the activities and communication patterns better. 3. Build an overview with aggregated statistics will help in refining and finding interesting individuals. 4. Consider a view that will help in understanding the volume of emails sent and/or received. 1. Main Individual's View	Question: To what extent visualisation can support analysts in discovering interesting individuals with their designations (organisation roles) in the E-mail communication data? Paper designs Tableau designs D3 designs Main view with bubble matrices (sent & received) Supporting views to provide more details of the sent and received Additional views to support organisational roles An individual overview - an overlay with aggregated statistics. Validation

	● Senders ● Receivers ● Both together 2. Another View ● Sender breakdown info ● Receiver breakdown info 3. Statistics View 4. Content to be displayed The suggestions from the expert (E2) during the informal feedback on medium fidelity prototypes are, ● Merge sent and received messages in one visualisation, in the high fidelity prototype, so it will help in understanding the activities and communication patterns better. ● Consider the departments/groups in an organisation that will help in understanding the activities and communication patterns better. ● Build an overview with aggregated statistics that will help in refining and finding interesting individuals.	

	• Consider a supporting view that will help in understanding the volume of emails sent and/or received. 5. Include Organisation roles view 6. Different Email domains view (private such as gmail, yahoo etc and organisation like Enron) The expert (E1) again liked the idea of the multi-faceted exploration and multi-granularity approach. The use of bubble matrix to visualise relationships between the selected individual and their connections is helpful. This can help analysts to identify interesting points and seamlessly switch between the different levels of main view and the supporting views. The aggregated statistics view helps. The representation of organisation roles adds value to the individual analysis. The expert (E1) came up with questions such as ``Can you find who are the senders and receivers in a thread?"	

	``Can you find which individuals have been excluded and included back? " The tool can help in visualising individual connections and this can be tested to see our business connections over a period of time.	
Features / Design / Implementation (Phase 3)	Aggregate by threads Compare different threads Consider types of correspondence (cc, bcc) Consider types of engagement Find the communication types Thread Metrics: Generate statistics from the threads (thread metrics), some ideas: - Pace of interaction - # of people (active/passive) - # of people (same organisation/different/personal) - # of inclusion/exclusion - sender diversity (# of unique senders / # of all involved) - thread length (in terms of duration (e.g., 5 days?) / in terms of message count) 1. Threads Overview (Dashboard)	Question: To what extent visualisations can support in discovering interesting individual behaviour (conversations) in the E-mail communication data? Paper designs Tableau designs D3 designs Single thread view Multiple thread view Thread features view Projection view Content view Validation

	○ Thread sizes (length) ○ No. of active individuals ○ All threads (thread versus time + only active individuals + selection of time-period) ○ ThreadStats View (Email replies, Entities of groups, etc) 2. Single Thread View ○ Each individual thread to be viewed ○ Identify active & passive individuals ○ Identify inclusion & exclusion of individuals ○ Sort by time/chronology & engagement ○ CC/BCC--trails for individuals ○ “Visual Signatures” – compressed	

	form of one full-thread. ○ Try out the “bespoke” metrics and observe patterns 3. Thread Features View (with metrics) ○ Pace of interaction ○ Identify active & passive individuals ○ Identify inclusion & exclusion of individuals ○ Sort by time/chronology & engagement ○ # of people (active/passive) ○ # of people (same organisation/ different/personal) ○ # of inclusion/exclusion ○ sender diversity (# of unique senders / # of all involved)	

	○ thread length (in terms of duration (e.g., 5 days?) / in terms of message count) 4. Content View (complete message) The expert (E1) questions on using topic analysis ``Can you find what kind of topics the senders and receivers discussed in a thread?" ``Can you find which individuals in a thread discussed a topic on California Energy Deal? ". The tool looks overall good and promising for investigation.	
Guiding Observations	Decided to go with Enron data Red Sift experts will be the users/analysts We will try to find interesting patterns trends, changes Also identify and detect changes Design a use case story (in a narrative way) to validate our approach/technique. Demonstrate the real case study (Enron Fraud Case) that has temporal, individuals and context using D3.js, DC.js and Crossfilter (for interaction).	Enron data (real) Red Sift experts (real) Find interestingness in the data Need to consider use cases. Use real case study (Enron Fraud Case) Better exploration facility for finding anomalies.

	Improve comparison of two or more subsets of data. The E-discovery analysts have difficulty in defining anomalies/abnormalities. We might need a tool with better exploration facility (better filters). We can consider the possible Exploration tasks	

Evaluation

- We conducted an empirical study with the Red Sift analysts to understand the usefulness of our solution.
- We were making notes in our diary book. Later it was transcribed straight after the interview to consider any clarification. This process was carried out on Google Documents (by sharing it with my supervisors).
- We focussed on the coding and thematic analysis.
- The methodology is described in Chapter 2.
- The Evaluation is discussed in Chapter 5.

Coding:
Workflow of the exploration
Use of Visualisation Views, Visualisation Information & Interaction
Use of Textual View and the Textual Information (Email Messages View)

Specifying the Characteristics of Threads
Use of Labelling
Number of instances of threads analyst considered before deciding on a label

Use of Active Learning
Response to the labels returned from the model
Refinement of labels/categories

Dec 2018

Session#1 — Analyst-1: A1

Analyst said “Interested to start with “Feature Projection” and then the thread exploration to get a sense of what the threads are about and the participants in threads”. The study is broken down into steps:

Workflow: Step-by-step action followed:

Step 1: Started with the exploration of the threads in the “Feature Projection”. After some initial exploration, by selecting various clusters, he labelled it as “Discussions”, “Broadcast” and “Engaged”. I call it as ``Discussion”, if the messages are being exchanged continuously for a long period of time, “Broadcast” - if the message is sent by one person to many individuals in the company without much of replies. “Engaged”, if all the recipients in the email were actively involved in the email thread.

Step 2: Always started with the “Feature Projection”, high-level investigation on the “Thread features” and spent some time on the “Thread Overview” to understand the thread comparisons. Selected each of the threads in the “Feature Projection” and labelled them based on the “Thread Messages” and the ‘Messages”.

Step 3: Later selected a cluster of threads in the “Feature Projection” and labelled them in one go. However, made a good use of the “Thread Overview”, and “Thread Messages” to label.

Step 4: Clicked on the update to see all the threads being classified.

Step 5: worked on only a few recommended samples and continued to re-investigate on the other threads if they are classified properly.

Step 6: Selected “High engagement” in the “Thread Features” over a period of time. He says “that’s what I wanted to see…….” verifies with the “feature projection”. Check both high & low engagement & check across other features as well.

Step 7: Selected “Interaction Pace”, selected both sparse and dense regions and checks in the thread overview and feature projection. Check across other features as well. He says he can see it in the Email messages.

Step 8: Moves on to “participant growth”, clicks on the high & low levels, re-checks the same in the feature projection and has a close look at the “thread overview”. Check across other features as well.

Step 9: Moves on to “Participant Growth Variation”. Selects more size variation, checks the thread overview -> Interprets and relates to the kind of conversation being shown on the “Thread Messages”. Nicely clustered in the projection view using the thread features. He says he can see it in the Email messages.

Step 10: Then selects “Sender Diversity”, is it the change/variation of roles of the participants in the thread….it is fairly homogeneous (from sending to receiving) at the top (high sender diversity), how diverse the roles of individuals (from cc to bcc) and at the lower (low sender diversity), there is no much diversity. This blends very well with the label “Broadcast” (again checks with thread overview and thread messages)

Step 11: Finally, selects “Sender Diversity Entropy”, the rate at which the events are happening. There is quite a variation in a thread selected as expected. Able to see the range. Checks with thread overview and thread messages, identifies the threads where the roles are changing rapidly. Identifies outliers. He says “It is nicely clustered in the “feature projection”…..

Other observations:

- In the thread comparisons -> able to identify activities in both the threads. Identified people-overlaps and time-overlaps in the threads.
- Bars in the thread are helpful to understand the engagement. Length of the bars and the colour of it are useful.
- Glimpse of how communicative they are.
- Keen to add one more class after the exploration and interpretation
- Relabelling was performed when an issue was spotted
- Long Discussion was labelled -> if the discussions are quite long over a period of time irespective of the number of people involved.
- The analyst did not make much use of the active learning and the recommended samples. The response to the labels returned from the model was not deeply investigated. Not many re-labelling was done.
- The different ordering in the thread details view (Time first vs. Engagement first) was used to understand the "*diversity*" proxy feature
- Identified underlying themes within one of the classes by observing the content and explained some of the behaviour, i.e., high engagement — Halloween vs. Thanksgiving [**]
- On ThreadOverview:
    - provides a "glimpse" on what could be the discussion inside look like
    - Seeing overlaps between individuals is helpful
    - A1 used it to observe multiple representatives from a class and get an overview

Positives & Limitations:

- Question raised – what if I select only one class like "long discussion", the tool must show all the selected class.
- Can one thread have multi-labels? (fuzzy logic). Wanted to do an additional class on time-based characteristics and wanted threads to have multiple labels (Wish: Fuzzy labelling would have helped)
- Sorting of the groups in the Thread Message view is interesting for finding chronology and engagement.
- Navigation & labelling is good.
- Legends in all the views to support navigation.
- Confusing which is "high engagement" and which "low engagement". Level of spectrum in the thread overview and provide a small guide.
- Interesting and never seen threads this way
- Great exploration and an investigation tool
- Front-to-back integration and connection is useful.
- Classification is quite useful. The clusters in the feature projection are informative
- Helps in navigating from high-level to low-level – a different perspective. Co-relation combined with multiple perspectives.

- People in the groups (same signature) makes it interesting.
- The case study used and the tasks used are quite relevant and important from organisation point of view.
- This current working model can be deployed in our Red Sift Platform.
- Was able to identify emails pertaining to festivals, parties and business discussions. The characteristics are same but different labels need to be formed.
- When more data points are selected/considered, the system can get complicated. That time it becomes difficult to comprehend, explore smoothly and classify.
- How about the scalability of it?
- It might be good to have a selection criteria.
- Individual arrows in the thread message view for guiding users.
- Can be personalised to different analysts, i.e., different analysts can build different classes
- The characterisation can be done both at a high-level and a low-level
- Large sets would be an issue, so sampling would be a good idea
- Ability to break-down clusters
- Legends to help navigate the new representations
- To test additional data types or tasks.

Session#2 — Analyst-2: A2

Analyst said “Interested to start with thread features to identify interestingness & relevance”. The study is broken down into steps:

Workflow: Step-by-step action followed:

Step 1: In thread features (in particular), interested to start with high “Participant Growth” because more and more people are being involved and topic of discussion might be of interest. Checked the “Thread Overview”, thread messages and found Sheila.Nancey as a person of interest. Identified the whole thread communication as “business chat” by checking the thread overview multiple times. Created the first model and labelled “business chat”. - > High-growth could be an indicator of business discussions which are of importance. Analyst works on the features rather than the clusterings on the “feature projection space”. He says “I understand what the features are and I can explain & relate to the features when constructing”. Analyst starts with and focuses on individual cases and tries to find representative threads and looks for expanding from them. He says “I don’t want an incorrect labelling to spread across the classification, so I mark only individual cases first and then expand checking carefully. If I marked groups, I would be making mistakes easily”.

Step 2: Moved back to thread features and selected high “Interaction Pace”. Explored threads in the “Thread Overview” and marked few threads as “business chat” but mentioned it as not so “interesting” based on the Email messages.

Step 3: Moved on to “Interaction Pace”, selected high interaction pace, followed same approach……thread overview, thread messages and marked some of the threads as “business chat” but mentioned it as not so “interesting” based on the Email messages.

Step 4: Now, selected high “Sender Diversity” -> thread overview -> thread messages and marked some of the threads as “business chat” but mentioned it as not so “interesting” based on the Email messages.

Step 5: Later, selected low “Participant Size variation” -> thread overview and identified some interesting threads but not particularly remarkable. Later, selected high “Participant Size Variation” and identified many “Farewell messages”, which is of not interest to him. Created a new label “Social” & marked the thread. Also, identified sales related emails and marked them as “SALES”. If the messages were related to farewell and parties, I call them “SOCIAL” (A single sender with people replying back only without much further engagement) and if the messages were related to some Enron sales & marketing, I will label them “SALES”.

Step 6: Finally, selected low “Sender Diversity Entropy” -> thread overview and identified some interesting threads about legal and felt they are relevant for deeper investigation. Labelled the threads as “Legal-trouble” -> if the messages had legal information about the Enron case. Also, identified “Pay roll” Enron and other announcement related emails and created a label as “Announcements” (if the messages were sent to a large group of people without any replies to it).

The analyst checked all the features several times to mark thread types. The feature projection was not used much by the analyst. The only time he used was to see the “recommended ones”. This actually helped in further investigation and marked each thread correctly.

Other Observations:

- Went through feature by feature, using the threadview for content explanations
- Used the recommended samples coming from Active Learning very closely, and even continued all were exhausted
- Uses *feature ranges* to look for interesting cases
- Drawn to the individual cases that are at the top of each feature as instances to investigate, being drawn to outliers [comment from CT: this is a limitation of the approach to comment on, people are drawn to the outliers but this is an easy signal to capture]
- Used the thread overview effectively
- Commented that usually the feature values are all over the place, high/low values in one do not always mean high/low values in another
- **On feature refinement (!!!)**: Spotted the issues with the “Engagement” feature when looking at a thread with low engagement. Spotted that due to the formula of this derived feature, the threads that contain many people are disadvantaged and those

that have less people are advantaged. Suggested that a non-linear function would work better and suggested a re-definition of the derivation

On the deployment and next steps:

- An alternative work-flow to filter out sessions
    - Usually, there is a multitude of uninteresting / irrelevant data records and they are the majority of the data, the ability to filter these out progressively will make the process
- **On legal case:**
- The key task is to identify the important threads. This approach helps by focusing on sub-groups which are likely to contain those cases that are interesting and through the informed filtering of those cases that are less interesting

Positives:

- The interface is good and helps in navigation & exploration.
- Easy to classify the threads based on the conversation or emails exchanged.
- The complete pipeline of the workflow is good but can be improved further.
- It is definitely useful for investigation and we will use it as a solution for our products planned.

Some drawbacks/limitations/improvements:

- Linearity & non-linearity of Engagement in threads. Unfavours large groups and favours smaller groups.
- He is fine with selecting “sparse data points” rather a big cluster of data in some of the thread features. They are quite difficult to understand & interpret, selecting a particular thread is a cluster is a challenge.
- Comments on the big clusters at the bottom of each feature is hard to work with
- Too much of selection & deselection is a concern.
- If an analyst selects on one of the labels already created/labelled, all the marked threads must be shown in the “Thread Overview”.
- The interaction could be fast when moving between “thread overview” to “thread messages” & “Message”, that when the mouse is hovered across all the threads in the “thread overview”.
- For Verification, the feedback loop can be improved -> the labelling process can be improved.
- Multi-labelling for each thread could be considered as well (fuzzy labelling would help)

- For example, if an analyst clicks on one of the labels created (like legal-trouble), he must have all the bunch of threads to be downloaded to be read (put them in the workflow).
- It is useful to create labels such as “interesting” and “non-interesting” and delete all the “non-interesting” ones, this will help us reduce the number of emails to be reviewed/investigated. This will actually help in removing many of the clusters in the thread features.
- High signal -> low noise , this will help in a good representation of the threads.
- Develop Progress Indicators (visual indicators) - how many labelled, not being labelled and being removed -> indication of progress.
- Ability to choose the labelled threads so far. If a label is selected, all the threads related to the label must be selected.
- Ability to choose the clusters as a whole

Session#2 — Analyst-1: A1 (again)

We considered the Analyst (A1) again to build a new model with a new dataset and evaluate the model built by another Analyst (A2). This was done so that there is consistency in modelling and utilising the tool.

Again, he started with the “exploration-first” approach, so he understands the data better.

He created classes/labels such as high-engagement & low-engagement to validate one of the Thread features (pseudo-features), “Engagement”, derived by us. Immediately, he started investigating the “projection”, he says “recommendation is a debatable and we mark them based on our interest and content”. In the middle of the feature projection, it is a 50-50 for high & low engagement. Based on his analysis, he feels some of the threads are interesting and labels the mid-section in the projection as “mid-level engagement”. This helps to see the “group discussions”, this is updated and could see all the group discussions.

Others:
Low & high engagement does not change in the projection, middle section is the mixture of the two labels (low & high).

Also, he worked on the model built by Analyst (A2) to classify the email threads manually. Feels it is a good way of evaluating.

Coding	Themes
Workflow of the exploration Use of Visualisation Views, Visualisation Information & Interaction	Pattern Discovery

Use of Textual View and the Textual Information (Email Messages View)	
Specifying the Characteristics of Threads Use of Labelling Number of instances of threads analyst considered before deciding on a label	Pattern Specification
Use of Active Learning Response to the labels returned from the model Refinement of labels/categories	Pattern Modelling

Encodings	Coded Lines	Extracted Themes
Workflow of the exploration	*Analyst 1 Workflow: Step-by-step action followed: Step 1 to Step 11* *Analyst 2 Workflow: Step-by-step action followed: Step 1 to Step 6*	Work Strategy of Analyst 1 Work Strategy of Analyst 2
Use of Visualisation Views, Visualisation Information & Interaction	Thread Exploration High-level investigation on the “Thread features” and spent some time on the “Thread Overview” to understand the thread comparisons. made a good use of the “Thread Overview”, and “Thread Messages” to label. Selected “High engagement” in the “Thread Features” over a period of time. Check both high & low engagement & check across other features as well. Selected “Interaction Pace”, selected both sparse and dense regions and checks in the thread overview and	Use of multi-faceted exploration Use of multi-granular approach.

	feature projection. Check across other features as well. Moves on to “participant growth”, clicks on the high & low levels, re-checks the same in the feature projection and has a close look at the “thread overview”. Check across other features as well. Moves on to “Participant Growth Variation”. Selects more size variation, checks the thread overview -> Interprets and relates to the kind of conversation being shown on the “Thread Messages”. Nicely clustered in the projection view using the thread features. Then selects “Sender Diversity” it is fairly homogeneous (from sending to receiving) at the top (high sender diversity), how diverse the roles of individuals (from cc to bcc) and at the lower (low sender diversity), there is no much diversity. Finally, selects “Sender Diversity Entropy”, the rate at which the events are happening. There is quite a variation in a thread selected as expected. Able to see the range. Checks with thread overview and thread messages, identifies the threads where the roles are changing rapidly. Identifies outliers. He says “It is nicely clustered in the “feature projection”…..	

	Sorting of the groups in the Thread Message view is interesting for finding chronology and engagement. Navigation & labelling is good. Legends in all the views to support navigation. Confusing which is “high engagement” and which “low engagement”. Level of spectrum in the thread overview and provide a small guide. Uses *feature ranges* to look for interesting cases Drawn to the individual cases that are at the top of each feature as instances to investigate, being drawn to outliers [comment from CT: this is a limitation of the approach to comment on, people are drawn to the outliers but this is an easy signal to capture] Used the thread overview effectively Commented that usually the feature values are all over the place, high/low values in one do not always mean high/low values in another	

Use of Textual View and the Textual Information (Email Messages View)	can see it in the Email messages Identified underlying themes within one of the classes by observing the content and explained some of the behaviour, i.e., high engagement — Halloween vs. Thanksgiving [**].	Message Box (content view) is helping to find interesting information.
Specifying the Characteristics of Threads	I call it as ``Discussion", if the messages are being exchanged continuously for a long period of time, "Broadcast" - if the message is sent by one person to many individuals in the company without much of replies. "Engaged", if all the recipients in the email were actively involved in the email thread. If the discussions are quite long over a period of time irrespective of the number of people involved. High-growth could be an indicator of business discussions which are of importance. If the messages were related to farewell and parties, I call them "SOCIAL" (A single sender with people replying back only without much further engagement) and if the messages were related to some Enron sales & marketing, I will label them "SALES". if the messages had legal information about the Enron case.	Email Classification

	if the messages were sent to a large group of people without any replies to it).	
Use of Labelling	labelled it as “Discussions”, “Broadcast” and “Engaged”. This blends very well with the label “Broadcast” (again checks with thread overview and thread messages) Relabelling was performed when an issue was spotted. Long Discussion was labelled Question raised – what if I select only one class like “long discussion”, the tool must show all the selected class. Can one thread have multi-labels? (fuzzy logic). Wanted to do an additional class on time-based characteristics and wanted threads to have multiple labels (Wish: Fuzzy labelling would have helped) Created the first model and labelled “business chat”. Created a new label “Social” & marked the thread. Also, identified sales related emails and marked them as “SALES”. Labelled the threads as “Legal-trouble” created a label as “Announcements”	Labelling the communication types

Number of instances of threads analyst considered before deciding on a label	Identified the whole thread communication as “business chat” by checking the thread overview multiple times. The analyst checked all the features several times to mark thread types.	Decision-making based on features and visualisations
Use of Active Learning	start with “Feature Projection” Always started with the “Feature Projection”, Selected each of the threads in the “Feature Projection” selected a cluster of threads in the “Feature Projection” and labelled them in one go. Clicked on the update to see all the threads being classified. worked on only a few recommended samples and continued to re-investigate on the other threads if they are classified properly. Classification is quite useful. The clusters in the feature projection are informative Helps in navigating from high-level to low-level – a different perspective. Co-relation combined with multiple perspectives. The feature projection was not used much by the analyst. The only time he used was to see the “recommended ones”. This actually helped in further investigation and marked each thread correctly.	Active Learning (to classify for all the data)

	Used the recommended samples coming from Active Learning very closely, and even continued all were exhausted	
Response to the labels returned from the model	The analyst did not make much use of the active learning and the recommended samples. The response to the labels returned from the model was not deeply investigated. On feature refinement	
Refinement of labels/categories	Not many re-labelling was done. Spotted the issues with the “Engagement” feature when looking at a thread with low engagement. Spotted that due to the formula of this derived feature, the threads that contain many people are disadvantaged and those that have less people are advantaged. Suggested that a non-linear function would work better and suggested a re-definition of the derivation.	

A.8 Work Plan

Detailed Work Plan for PhD Completion

Visual Analysis of E-mail Communication Data within Digital Forensics and E-discovery

The aim of the research is to design and develop interactive visual solutions to explore and find/discover relevant/interesting information in the given communicated E-mails from an investigation perspective to support organisations specialising in Digital Forensics and E-discovery.

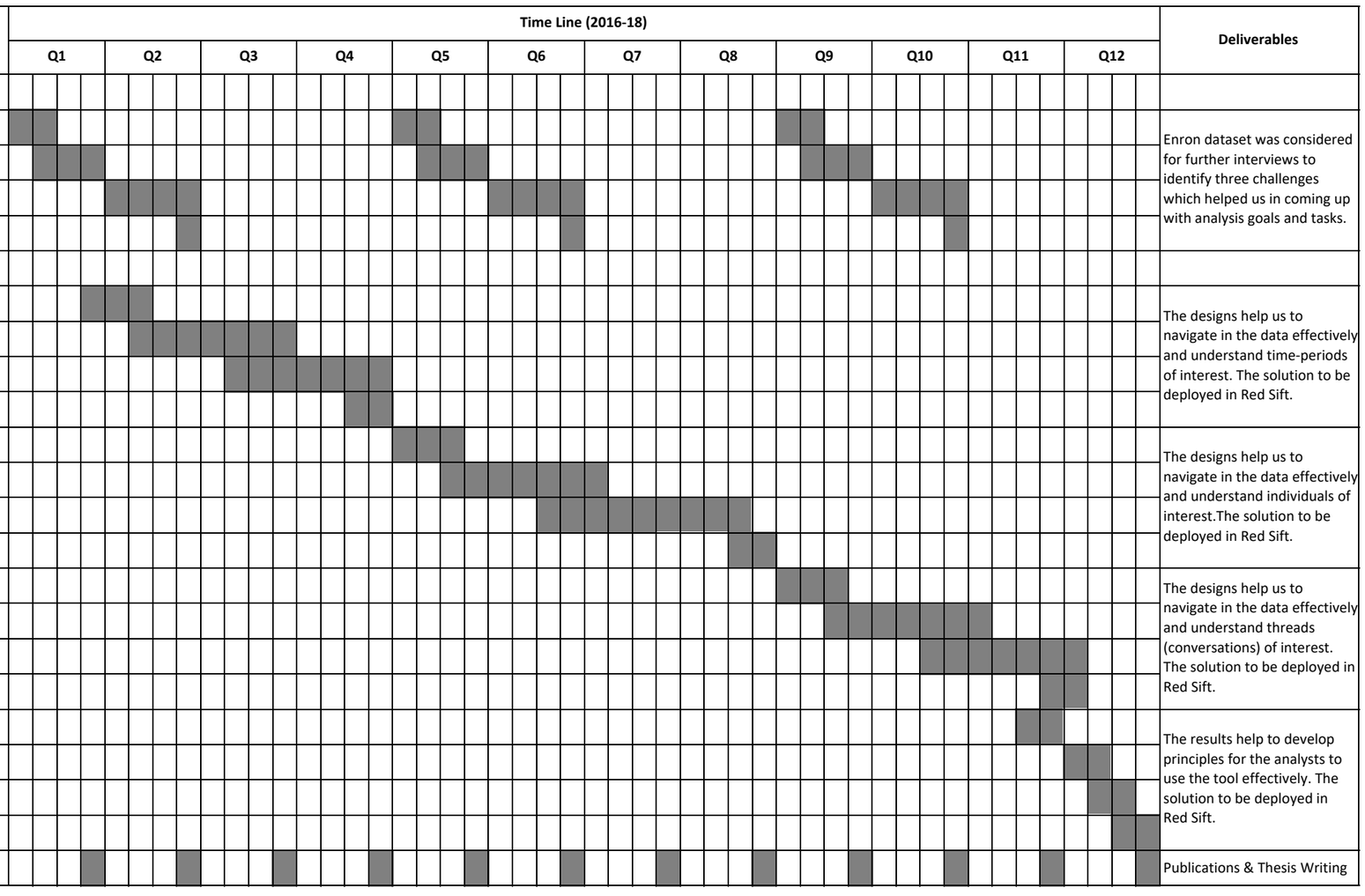
The Researcher

Mithileysh Sathiyanaryanan

The Supervisors

Dr Cagatay Turkay and Prof Jason Dykes

Steps	Tasks	Objectives	Activities	Time Line (2016-18)	Deliverables
1	**Domain Characterisation**				
a	Initial Interviews, Discussion Sessions, Consultations, Case Studies and Datasets	To understand the E-discovery challenges, tasks and requirements. To investigate on the case studies and datasets related to E-discovery.	Understand the exisiting challenges		Enron dataset was considered for further interviews to identify three challenges which helped us in coming up with analysis goals and tasks.
			Introduce Case studies to the experts		
			Develop Analysis Goals & Tasks		
			Validation		
2	**Develop Solutions based on Requirements & Tasks**				
2a	Discovering and characterising time period(s) of interest (Pattern Discovery)	To develop design solutions to explore, understand and identify changes, similarities (commonalities), differences (variations) to find interestingness.	Develop Low-fidelty prototypes		The designs help us to navigate in the data effectively and understand time-periods of interest. The solution to be deployed in Red Sift.
			Develop Medium-fidelty prototypes		
			Develop High-fidelty prototypes		
			Validation & Deployment		
2b	Discovering and characterising individual(s) of interest (Pattern Discovery)	To develop design solutions to explore, understand and identify changes, similarities (commonalities), differences (variations) to find interestingness.	Develop Low-fidelty prototypes		The designs help us to navigate in the data effectively and understand individuals of interest.The solution to be deployed in Red Sift.
			Develop Medium-fidelty prototypes		
			Develop High-fidelty prototypes		
			Validation & Deployment		
2c	Discovering and characterising thread(s)/conversation(s) of interest (Pattern Discovery)	To develop design solutions to explore, understand and identify changes, similarities (commonalities), differences (variations) to find interestingness.	Develop Low-fidelty prototypes		The designs help us to navigate in the data effectively and understand threads (conversations) of interest. The solution to be deployed in Red Sift.
			Develop Medium-fidelty prototypes		
			Develop High-fidelty prototypes		
			Validation & Deployment		
2d	Empirical Evaluation	Express positive findings and recommendations based on the evaluation that can help analyst to explore & identify various interesting information relevant to the investigation.	Designing the experiments		The results help to develop principles for the analysts to use the tool effectively. The solution to be deployed in Red Sift.
			Conduct the experiments		
			Analyse the results		
			Validation & Deployment		
3	**Reflection on the research work**				Publications & Thesis Writing



A.9 Risk Analysis

The interactive visualisation will help in understanding various stages we undertook in this research. Link: `https://mithileysh.github.io/Mits-PhD-Timeline/`

Below are the risk analysis considered before the start of this project.

Description	Probability	Impact	Action
Domain Characterisation Issues	3	7	The domain problems, challenges and issues to be discussed with the domain experts on a regular basis to mitigate any circumstances.
Design Issues	4	7	The design principles and implications to be discussed with the supervisor and the giCentre team on a regular basis to mitigate any circumstances.
Deployment Issues	4	2	The engineering issues to be taken care by the partner company.
Laptop/Hardware Issues	3	4	It is important to back up all the documents and codes in the online drives or pen drives.
Time Management Issues	2	3	Sticking to the work plan is very important.
Health Issues	5	5	It is always good to make use of the weekends, holidays and annual leaves and plan things well ahead.
Moving away from the scope of the project	4	4	Regular meetings with the supervisors and collaborators will ensure the work is within the scope of the project.
Supervisor not able to supervise or relocated to a different place	2	6	The second supervisor to take the lead.

Values	Risk Type
1-3	Low Risk
4-6	Medium Risk
7-9	High Risk

Figure A.5: Risk Analysis and management was carried to anticipate the risks/failures and manage the activities efficiently by avoiding any delays in the project.