Challenges and Opportunities in using Analytics Combined with Visualisation Techniques for Finding Anomalies in Digital Communications

Mithileysh Sathiyanarayanan
City, University of London, UK
Email: Mithileysh.Sathiyanarayanan@city.ac.uk

Cagatay Turkay
City, University of London, UK
Email: Cagatay.Turkay.1@city.ac.uk

***Abstract*—Digital communication has changed human life since the invention of the internet. The growth of E-mail, social websites and other interpersonal communication systems in turn have brought rapid development in especially the key technological area of data analytics. Using advanced forms of analytics helps the examination of data and better informs investigative sense-making and decision-making of all kinds. The legal process called Electronic discovery (E-discovery) is used for investigating various events in the digital communication world, for the purpose of producing/obtaining evidence (such as evidence in the form of emails used in the Enron fraud case). Investigating digital communications collected over a period of time, manually, is a strenuous process, time consuming, expensive and not very effective. More recently, within E-discovery there has been development of analytics known in the legal community as "Technology assisted review" (TAR). TAR is a technology-driven assistant in E-discovery for identifying relevance in the documents/data which saves time and improves efficiency in investigation. At the same time, the efficacy of visualisation tools currently available in the market is increasing, where such tools depend on a combination of simple keyword searches and more complex representations (e.g. network graphs). Also in E-discovery, early case assessment is a process of estimating risk (cost and time) to prosecute or defend a legal case based on an early review of potentially relevant electronically stored information (ESI). Legal firms largely determine the duration of the E-discovery process and charge companies based on the volume of information collected and reviewed after an automated search, where ESI may then be manually reviewed intensely to determine relevance and privilege. This results in significant costs for the company or in a number of cases settlement because a party cannot afford to continue with the lawsuit due to E-discovery costs.**

This paper examines some of the opportunities and challenges in searching digital communication data for E-discovery and investigations, and will explore how analytics coupled with visualisation techniques may lend support and guidance in these efforts. Addressing these combined techniques may yet yield improved data collection, analysis and understanding of how analysts/lawyers can work together using visualisations. In particular, we attempt to address two challenges: (i) improving comparison of subsets of data, and (ii) identifying anomalies (including sensitivities) in email communications.

***Index Terms*—E-discovery; Visualisation; Decision support; Technology Assisted Review**

I. Introduction

With the continuing growth and adoption of digital communications, investigating huge volumes of data is becoming an increasingly complex challenge in both E-discovery and investigations. Electronic Discovery (E-discovery) [1] is a domain where electronic/digital communication data is sought, located, secured, and searched with an intent of using it as evidence in a civil or criminal legal case. With the amount of "Big data" generated by and stored within organisations today, legal departments and their external counsel often struggle to quickly get a handle on the data needed for a given investigation or litigation. Increasingly, to review that data in a strategic manner means using some forms of analytics while as part of early case assessment [2].

Electronically stored information (ESI), for the purpose of the US Federal Rules of Civil Procedure (FRCP) is information created, manipulated, communicated, stored, and best utilised in digital form. Rules for litigation (in both federal and state courts) allow for an expansive approach to what may be discovered during the fact-finding stage of civil litigation. ESI includes writings, drawings, graphs, charts, images, audios, videos, documents and other data compilations stored in an electronic medium. From Fig. 1, we understand organisational communications (e.g., E-mail) and social media communications (e.g., Facebook, Twitter etc.) are a subset of digital communication which is in turn a subset of all ESI. In this work, our focus is on digital communication data.

The Electronic Discovery Reference Model (EDRM) [3] represents a conceptual view of the E-discovery process and is useful also in setting out workflows for related areas, including compliance, business intelligence and investigations. Using this model, the principles of E-discovery can be applied iteratively to get a more precise set of results. Using this model, our aim here is to work in the direction of visualisations and presentations.

Technology assisted Review (TAR) is a technology-driven, proactive process that in recent years has proven to be a breakthrough for the legal process, in terms of providing for both time-and cost-savings and in assisting counsel and clients in making better-informed decisions. In recent studies (Grossman & Cormack 2012), technology-assisted review has been shown to yield more effective results than exhaustive manual (human) review, with much lower effort. However, the efficacy of visualisation tools currently available in the market, based on simple keyword searches even if coupled

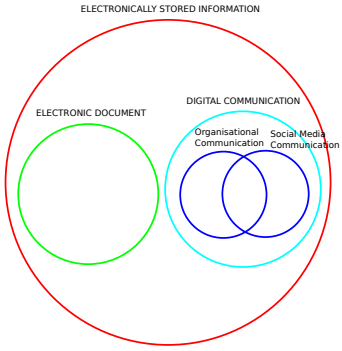

Fig. 1. Automated Classification of ESI Using A Combination of Methods

with more and complex representations (e.g. network graphs), are not particularly as efficient and effective as they could be in comparing subsets of data including identifying anomalies at one go.

Visualisation tools enable legal teams to create an index of search terms, visually review the documents/data containing those terms, and determine quickly whether those terms are of relevance. Also, visualisation tools help analysts see a visual representation of relationships between subject parties and the people they communicate with via E-mail or social media or interpersonal communication systems. These kind of reviews and analyses hold the potential to provide significant new opportunities for attorneys in law firms and in corporate legal departments. The visualisation tools currently available on the market are based on simple keyword searches. Legal firms mainly charge companies based on the volume of information involved in the keyword search, the output of which is then manually reviewed intensely for relevance and privilege [4]. The recent review report by the UK Home Office [4] states, there are no E-discovery tools that have the ability to display temporal or spatial information in an innovative way. But for E-discovery compliance, experts regularly need to investigate "samples of emails", and it is important for them to select a representative sample (or, alternatively, an interesting one containing anomalies or sensitivities however defined). However in a data context such as email (one that is multi-modal and dynamic), the definition of "interesting" is vague and the information obtained is multi-faceted. Hence there is a need for visualisation- empowered solutions to support the analysts with this particular task.

The avalanche of E-mail data is expanding exponentially with different degrees of variety and complexity. With E-mail traffic continuing to grow at 5% [5] a year, in the business context more companies either are or should be requiring time-saving and cost-effective solutions in connection with anticipated E-discovery. As the E-mail data keeps increasing exponentially, understanding the meaning contained in the data grows more complex, tedious, time-consuming and expensive [4]. To reduce the aforementioned and maintain high quality in the E-discovery process, an advanced, powerful and effective analytic tool is in need that could visually compare two or more subsets of email data to understand what constitutes "interestingness" and "relevance". A deeper focus on analytics will help legal teams develop more insightful strategies, one of which would be to combine keyword and context searches with visual representations about key players and their relationships [4]. In E-discovery, retrieving key information is important, as data can be noisy and diverse in nature and origin. As the amount of digital information to investigate is continually growing, the current visualisation tools are too often complex and cumbersome in nature to carry out the process. We make these observations as a starting point. To visualise subsets of data, we need effective filtering techniques to roll-up/drill-down the data. E-discovery tools such as Brainspace Discovery5^{TM}, Jigsaw [6], Concordance by LexisNexis [7] and/or IN-SPIRE [8] to analyse unstructured data. However, a visualisation tool such as Brainspace Discovery5^{TM} does not perform individual document reviews, but is an effective culling tool that effectively produces visual representations of document categories for prioritization purposes.

Visual Analytic tools specific to E-discovery: From the discussion sessions with the experts, we tried to scope digital communication data to E-mail communication as most of their E-discovery investigations are related to E-mails. The experts/analysts use E-discovery tools such as Brainspace Discovery5^{TM} [9], Jigsaw [6], Concordance by LexisNexis [7], IN-SPIRE [8], Radiance [10], Zovy Advanced E-discovery (AeD) [11], DocuBurst [12] to analyse unstructured data. There are many E-mail visualisations developed by various researchers, some of the well-known ones are ContactMap [13], ConversationMap [14], E-mailMap [15], E-mail time [16], EzMail [17], re-mail [18], The-mail, See-mail. From our analysis, considering the tools for the investigation domain, many of the visual analytic tools have been used by many analysts; however, in our experience it has its own drawbacks, in that the tool must be used in combination with other tools to carry out investigations. Also, other drawbacks, such as inconsistency, complexity, and a not very powerful toolset for E-discovery and investigations. The other limitations are as follows: (i) arduous to compare two or more subsets of data. (ii) strenuous to detect anomalies and changes in data. (iii) onerous to explore data. (iv) Unfavorable interaction facility and (v) Unfavorable multi-faceted data analysis. These limitations will be addressed in our work.

In preparation for the DESI VII workshop, we engaged in discussions with several legal experts. This helped us examine the opportunities and challenges in using analytics against digital communication data in E-discovery. In particular, we were able to identify specific problems that require visualisation support and guidance. Addressing these problems will yield improved data collection, analysis and understanding of how analysts/lawyers can work together using visualisations.

Discussion with Legal Experts: The first discussion session was with a legal team of six solicitors in Bangalore, India. The key questions that we asked were:

Q1: What are the challenges of E-discovery and investigations

with respect to visualisation?
Q2: How do you investigate the key time-frame, key words and key individuals/players involved?
Q3. How can visualisation inform unexpected behaviours? Do you think there are useful tools for investigations?
The second discussion session was with an intelligence analyst who works at the cyber investigation department in Bangalore. Similarly, the key questions that we discussed were:
Q1. How to identify a "normal" subset in a given entire dataset?
Q2: How to identify "interesting" and "relevant" subsets in a given entire dataset?
Q3: How do you categorize normal and abnormal digital communication data? In fact, how do you characterize suspicious behaviours?
The discussion with the experts helped us better understand the opportunities and challenges with respect to searching and filtering digital communication data. It also helped to identify recurring problems that in our view require or would be greated helped by visualisation support and guidance.

II. Challenges

In this section, we discuss the challenges in E-discovery and investigations with respect to digital communication data, and How visualisation tools can be of assistance. Our goal is to set out the relevant issues so that analysts can showcase complex data-driven results in a compelling, interactive, and easily understandable visual format.

A. Challenges in E-discovery Investigations

Many corporate legal teams and other E-discovery experts are looking for solutions that address various challenges. Some of the challenges we identified are [19]:

1) **Remaining Competitive:** the ability to narrow-down (drill-down) the data to be investigated so as to enable a more efficient and competitive E-discovery process. A visual analytic tool should be able to sift out as much as 90% of non-relevant data during E-discovery, including early case assessment (ECA), and the tool should help in acquiring the most relevant data to be investigated to achieve faster results.
2) **Fluctuating Costs:** there are unpredictable, fluctuating costs involved in the E-discovery process (in simple words, the process can be very expensive). To reduce cost, an effective tool must be able to narrow-down a large data into a smaller data set for further manual investigation and review.
3) **Maintaining Quality:** to maintain high quality in the E-discovery process, advanced and powerful tools are needed for filtering, grouping, aggregating, processing and reviewing ESI.
4) **Searching Information:** the effectiveness of keyword, context and connections searches can be improved if they involve a combination of testing, sampling and iterative feedback with a visual representation of key words in documents used by key players, as well as relationship data in their communications. Visualisation tools can elevate the user experience and help analysts understand knowledge/information association and representation.
5) **Motivating Legal Community:** some digital communication data is at such a significant level of complexity that it cannot be managed with conventional analytic approaches. Technological advancements increasingly drive strategic decisions, and a deep focus on analytics will help legal teams develop more insightful strategies. Smaller data sets (subsets from the original data) coupled with more information about those subsets will help legal teams to make best decisions in the case. The strategies will help to yield time (faster/quicker analysis), save cost and reduce flaws. Current TAR methods (also known as predictive coding, intelligent review and computer assisted review) are complex and may yield inconsistent results in determining relevance. Search metrics can identify key terms and concepts and eliminate inefficiencies. Effective comparison strategies remain a consistent predictor for successfully engaging with digital communication data.
6) **Utilising Strategies:** digital communication data is on a significant level of complexity that cannot be managed with conventional analytic approaches. Technological advancements increasingly drive strategic decisions and a deep focus on analytics will help legal teams develop more insightful strategies. The smaller the data (subsets from the original data) and more information will help legal teams to make best decisions in the case. The strategies will help to yield time (faster/quicker analysis), save cost and reduce magnifying of flaws. The current TAR (also known as predictive coding, intelligent review and computer assisted review) methods are complex and inconsistent in determining relevance. Search metrics identifying key terms and concepts and eliminate inefficiencies. Effective comparison strategies remains a consistent predictor for successfully engaging with digital communication data.
7) **Adopting Tools:** Ideally, the tools used will work in the way the human brain works , so as to be tightly integrated into the workflow of E-discovery analysts. A tool must be easy to use to help to improve speed, performance and scalability. The tool should also easily represent important, interesting and/or relevant data visually, and interpret that data with ease, such that it improves the analytical reasoning and tech-savvy-ness of lawyers and judges.

B. Challenges in Digital Communication Data

The world of digital communication data is basically divided into entities and relationships (often called relations), where a group of things will be considered as a single entity and the relationships form the structures that relate the entities. Data attributes can be in the form of different levels of measurement: nominal, ordinal, interval, and ratio scales. Based on the different types of data, the bulk of

digital communication data challenges are being addressed by industries – but less in interpreting data than in presenting it in valuable and meaningful ways. The seven challenges we identified [20],[21]:

1) **Data Exponentiality:** digital communication data is exploding at a faster rate than expected. As the data keeps increasing exponentially on a daily basis, there is no specific way to reduce the speed of searching the data being electronically stored – managing large volumes of data is tedious. The challenge is how to deal with the size of data and how to transform the data into a form suitable for analysis. The other ways to investigate large volumes of data are to extract specific data (logically prioritised) and use effective analytics, which will make the E-discovery process more efficient.
2) **Data Multiplicity:** communication data is exploding at various sources, and aligning multiple datasets from various sources for decision-making can be tedious and complex. The challenge is how to handle multiplicity of types, sources, and formats such as text, sensor data, audio, video, images, graphs, files and many more in a way that does not treat each category as a "data silo". Also, there must be a smooth transition between structures, including semi-structured and unstructured data and complex data, which will make the decision-making in E-discovery more efficient.
3) **Data Exactitude:** to understand whether the communication data we analysed is good and/or accurate, we need to cope with uncertainty and imprecision. The other important challenge to tackle as to understand that extracting meaning contained in huge data sets is a complex, tedious, time-consuming and expensive process. As stated, an efficient tool must operate the way human brain works, which will help E-discovery analysts to understand the meaning contained in the data sets in front of them and more quickly recognize the important story of the case.
4) **Data Dogmatism:** there is always the tendency to lay down principles or guidance in analysis as undeniably true, without consideration of case studies/evidence or the opinions of domain experts. Analysis of digital communication data with recommendation from experts can offer quite remarkable insights for E-discovery teams.
5) **Data Streaming:** digital communication data comes in streams, and our task is to find interesting facts from streams in the real time. Most of the tools developed so far are useful for investigating archived data (also called as historical data). There is a big need for developing effective tools for investigating live-streaming data (also called real-time, fast streaming data), and enabling real-time data analysis. There is another issue in archived data called as "data accessibility" which needs to be addressed. Data accessibility is the process of collecting, storing, retrieving, and/or acting on data in a database/repository. These will make the sense-making and decision-making in E-discovery more efficient.
6) **Data Presentation:** there is a constant 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 it to judges for verdict and/or to share their investigated visuals with their colleagues to enable them to continue further investigation. Displaying complex analytics/visuals on a mobile device or non-mobile devices is a challenge. Reducing the cognitive complexity of data through visuals will help judges. Also, supporting real-time creation of dynamic, interactive presentations/reports allows analysts and E-discovery/legal teams to share and collaborate securely.

C. Challenges in Data Analytics

The challenges are with the five types of data analytics [22]:

1) **Prescriptive Analytics:** this type of analysis reveals what actions should be taken. This is the most valuable kind of analysis and usually results in rules and recommendations for next steps. Prescriptive analytics will give laser-like insights, but still has some challenges to tackle. Two of these challenges are to improve data integration and speed. Also, data integration and analytics needs to be a continuous process.
2) **Predictive Analytics:** an analysis of likely scenarios of what might happen in the future based on past patterns. The deliverable is usually some form of predictive forecast. The challenge is to improve harnessing of data and also improve automation and machine-based innovation for better decision-making.
3) **Discovery Analytics:** this analysis is the action of discovering insights. The challenges are to improve discovering hidden insights and improve decisions, by enriching information for decision makers. Also, as noted, data discovery that combines machine learning with visualisation improves the process.
4) **Diagnostic Analytics:** this analysis looks at past performance to determine what happened and why. The result of the analysis is often an analytic dashboard and they are used to determine why something went wrong or what happened.
5) **Descriptive Analytics:** this is also called "data mining", and helps to understand what is happening based on incoming data. To mine the analytics, we typically use a real-time dashboard and/or email reports. The challenge is to understand the data, improve the data quality and make analysts more data savvy. One of the questions here is "how to verify and normalize the data as quickly as our system can deliver and parse it?".

D. Challenges in Text Analytics

In text analytics, we have many challenges in the areas below, which will be discussed in detail in the extended version of the paper. The challenges include: text identification, text mining, text categorisation, text clustering, search access, entity/relation modelling, link analysis, sentiment analysis,

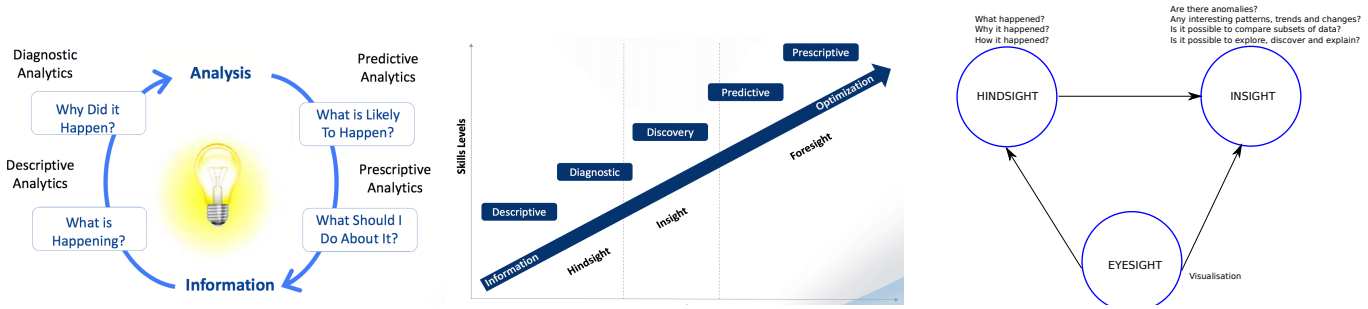

Fig. 2. (a) How analytic excellence leads to better decisions (b) Information to Optimisation (c) How our eyesight can be of help? Source: [22]

summarisation and visualisation. One important challenge is to converge traditional relational databases and digital communication data with the right technology that will improve the convergent model.

E. Challenges in Humanised Analytics

Humanised forms of digital communication data cannot be fully automated, and handling the entire processing task over to smart machines with techniques is also fallacious. Since data points are provided by humans (analysts) and they are involved in the processing of data, having a right balance between machines and the human element for the analytics process is the key to get optimal results.

F. Challenges in Visual Analytics

The core disciplines of visual analytics are visualisation, data management, data mining, data analysis (visual analysis), perception and cognition. In this paper, our focus is on visualisation, as research in this direction is currently emerging and analysts are in a big need of simple and effective visuals to identify anomalies and hidden patterns.

Challenges in Visualisation

Visualisation is becoming an increasingly important component of analytics in the age of digital communication platform. Visualisation for digital communication data is still emerging, tracking the revolution in the communication world as the "information overload problem" continues to grow. A danger is in getting lost in data, which may be due to irrelevant choice in designs, tasks, strategies or charts to present the information. Over the years the visualisation community has developed a wide range of visualisation techniques to analyse digital communication data. Most of the research and development carried out in the communication data visualisation area is of academic interest rather than for applicability in legal practice. Nevertheless, we believe there is a big need to have different approaches in understanding data systems, and the user/analyst/decision-makers needs to create effective visualisations to assist in this understanding. Though there are many ways developed to visualise digital communication data, not many are very efficient and analyst-friendly when data is considered over time. They work well for visualising a small set of data, but not a huge chunks of data over time, such as data from live-streaming. Most of the digital communication visualisations are complex, cluttered and difficult to understand in terms of events and various relationships. There is a need to understand how users/analysts interact with data, how they perceive it visually and non-visually, and how their mind works when searching for both known and unknown information. Since visualisation is the medium of a semi-automated analytical process, humans and machines must in the future coordinate using their respective distinct capabilities to achieve effective results. To sum up, from the preliminary discussions with legal experts and a reading of the literature [23], we see the overall challenges as expressed below:

1) **Improve search facility and ease of use/interpretation:** with the ever increasing amounts and complexity of data, the tools currently available on the market are complex and are based on simple keyword search, which is strenuous. The challenge is to improve visualisation support with analytical abilities.
2) **Improve investigating live-streaming data:** most of the visualisations developed so far are useful for investigating archived data (also called historical data). There is a further big need for developing visualisations for investigating live-streaming data (also called as real-time fast streaming data).
3) **Improve investigating dynamic data:** there are very few visualisation techniques which visualise and analyse dynamic change for communication data; when observed over time they are not very feasible for a real-world application. The challenge here is to understand dynamic communication changes using a simple and effective visualisation.
4) **Improve data input and visibility:** there are many tools for investigation but one of the issues is with inputting data for analysis. Legal experts are not tech-savvy to handle this issue, so there is a need for making abstract concepts and relations within data visible, then create a smooth flow between data input and carrying out analysis.
5) **Improve comparing of two or more subsets of data:** currently, in many instances a manual sampling is used for comparison that enables E-discovery analysts work-

ing on subsets of data to spot similarities and/or differences, including stratified features/attributes based on reports or clues. The iterative process of sampling data and comparing is strenuous and not always integrated sufficiently to help in identifying important differences in subsets. Also, some of the current techniques/approaches do not aid in supporting various features in comparing subsets of multi-faceted data. But representations must be effective for displaying multiple relationships and comparisons when placed close together or side by side (in an integrated format). Where effective, these representations improve the comparing of two or more subsets of data over time to identify similarities and differences.

6) **Improve detecting anomalies, changes and correlation:** E-discovery analysts have difficulty in defining anomalies or abnormalities in data .In fact, "anomalous behaviour" is hard to define and we need a more robust model of normality to be define what is "normal" as well as to 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 adapt into real-world applications due to their cumbersome approach, especially when considering multi-faceted communication data over time. Representations must be simple and efficient to identify and detect anomalous behaviours in data over time. As mentioned earlier, E-discovery analysts still use manual sampling to work on subsets of data for various reasons. One of the problems 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 viewed as not being adequately addressed in current tools.
7) **Improve guided open-ended data exploration:** E-discovery analysts have difficulty in exploring large datasets and they have become a big concern due to navigation issue, especially for communication data. The exploration of the email corpus must be beyond target search, i.e., supporting visual querying along temporal, connections, context and conceptual dimensions. So, the challenge here is develop an interactive visualisation tool with exploratory guidance that will help in navigating smoothly across various dimensions and also aid in suspension (pause and resume while exploring).
8) **Improve effusiveness in data explanation:** explanatory visualisations, also called "Visual Storytelling", have gained prominence in recent years with the rise of big data. Various studies have made it easier for humans to understand information integrated into visual stories than into many scattered visuals, as visual stories are more compelling. There are not many visualisations for digital communication that can provide a story or hierarchial information that happened in a specific year and in an effusive way (a way that quickly explains to gain insights). The challenge is to develop an interactive visualisation tool that will produce a visual story to efficiently convey information and foster better understanding of the chosen scenario.
9) **Improve simplicity in visual presentation and collaboration:** 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.
10) **Improve analytical reasoning approaches/techniques:** many visualisation tools in the market still lack analytical reasoning approaches, which makes some lawyers feel the tools are cumbersome to use. With more effective analytical strategies and techniques built in, lawyers amd analysts will feel better about the convenience of the tools to help in carrying out E-discovery and investigations.

Most E-discovery jurisprudence confirms the effectiveness of a process that involves a combination of testing, sampling and iterative feedback. For, E-discovery and investigations to be a fast-iterative process, we need effective visualisations that aid comparison and detection of anomalies. With the proper use of visualisations, workflows can often be streamlined to eliminate the long and multiple review processes that generally needs manual reviews. So, in simple words, the challenge is to develop a simple, effective, efficient and analyst-friendly visualisation tool which will be tangible and feasible to explore, understand dynamics, and which will aid in comparing and identifying anomalous behaviours to create a complete visual story in a set of digital communications.

III. Design Problems

In developing visualisations for various different domains one generally has multiple design problems that range from ones with low-level impact to high-level impact. One problem that has taken a toll on some E-discovery analysts is the decoding problem for comparing two or more units effectively with temporal and context connections, or their combination. We surveyed the existing techniques in the small multiples and aim to fill the gaps that have been left unexplored and unpublished so far, related to determining baselines, defining normal and identifying anomalies. At the outset, we took inspiration from the work by Dasgupta et al. [24] that classifies design problems into encoding, decoding and other problems. Our focus is mainly on the encoding techniques and reducing decoding problems, as explained below:

Encoding Problems. From the literature, we understand that appropriateness (also called "selection") is quite important. This allows analysts to characterize the level of data, task or charts to be used effectively for a specific purpose and aid in decision-making process. We classified the selection approach

into chart, data and task appropriateness. Interestingly, they are all inter-related. As a starting point, an analyst or a domain expert needs to understand the type of data he/she is working with. The next step is to understand the tasks needed and what kind of tasks can be performed. Once the data and tasks are decided, choosing an appropriate chart can make a big difference in the decision-making process [24]. As discussed by Dasgupta et al. [24], the two causes for the appropriateness problem are mismatch and configuration that can be in the data, tasks, or charts. As encoding problems mostly depend on the choices made by the designer and very little by the domain experts, we are not explicitly addressing the encoding problems – as the subject needs a more grounded theory approach and evaluation.

Decoding Problems. From the literature, we understand decoding problems mainly occur due to the elements of perception and cognition in the visualisation. It might be perceptually confusing or visuals might be too distorting or might be too complex. Often, the intended information will not be conveyed clearly enough. To address "How to effectively decode the visual explicit encoding?", we need to understand comparison complexity and color clutter [24]. To effectively compare and visualise multiple variables (decode), comparison complexity needs to be minimised. The comparison complexity is mainly due to lack of explicit encoding. To visualise multiple variables, small multiples are the best choice, but for optimal results, position and sequence of the individual charts are quite important. For example, to identify differences in a chart, a random arrangement will not help in identifying the similarities/differences task. It might need a deep visual inspection to identify and extract some information. In our work, we are using the techniques of juxtaposition and explicit wrapping. and we are discussing the above mentioned problems in our solution and the implementation [24].

Other Problems. The most important criteria in visualisation representation is to represent the data correctly and accurately. However, there are problems other than encoding and decoding, such as ambiguity, inconsistency and unascertainness [24]. The design problems help us understand the design space that is needed to build a general framework and that can be adapted by any domain that needs to investigate on temporal, connections and/or context.

IV. Design Requirements

Design requirements can also be called characteristics (principles/qualities) of good design (visualisation). In simple words, design is the process of visual communication, systematic design and problem-solving to achieve certain objectives. It is also visual communication and the aesthetic expression of concepts and ideas using various design elements and tools. The principles of a good design must describe the ways that designers use the elements of visualisations in a design, that is, a balance is the distribution of the visual weight of objects, colors, texture, and space. Since the visualisation is aimed at designed especially for legal and E-discovery analysts, several design requirements are identified and discussed in detail based on the discussion with the legal experts.

1) **Affordances:** they are clues about how visualisation must be used, typically provided by the visualisation itself or its context or the domain area used. Example, whether to use as overview, drill-down (filtering, sorting, selecting) or deep analysis. So, visualisations must be presented with some visual cues about what to consider and what not to (temporal cueing).
2) **Aesthetics:** visualisations must be aesthetically pleasing (visually appealing), simplified and systematically organised such that it can be perceived quickly (visual perception), showing attention to details that helps in reading, thinking (cognition) and problem solving. This will help to maximise the visualisation's impact and memorability.
3) **Accessibility:** visualisations must be simple, straightforward to understand, and easy to deploy and use. They must be accessible to modify easily when necessary.
4) **Applicability:** visualisations developed must be helpful in gaining insights and can be used in various domains or areas.
5) **Acceptability:** visualisations must be immediately accepted and used over time.
6) **Ambiguity:** visualisations must be easy to predict trends/future (to some extent) and must not distract or mislead analysts.
7) **Adoptability:** can also be called adjustability or flexibility or feasibility. Visualisations must be able to get adapted to various data formats being used.
8) **Clutter:** visualisations must be free from clutter such that it will help in readability and improves users attention.
9) **Desirability:** visualisations must not only be easy to use but also pleasurable to use, easy to understand, explore, search, find, detect, analyse and get insights in a timely way as the data keeps growing. This characteristic can also be called as "necessary".
10) **Informative:** visualisations must understand the importance of context and include information about who, what, when, where and how of the data. Visualisations must be a means to discover and understand investigative stories, and then to present them to others. Visualisations must be intuitive and easy to interpret.
11) **Interaction:** visualisations must have interactions that are simple, powerful and analyst-friendly to explore data at different levels of granularity. Animations and 3D can be avoided to a maximum extent.
12) **Scalability:** visualisations must be scalable as data size increases.
13) **Selectability:** Determine the appropriate type of graph for data data selection, graph selection.
14) **Usefulness:** visualisations must help users to make relevant and right decisions by viewing all the information they need in one place. Good design uses visual

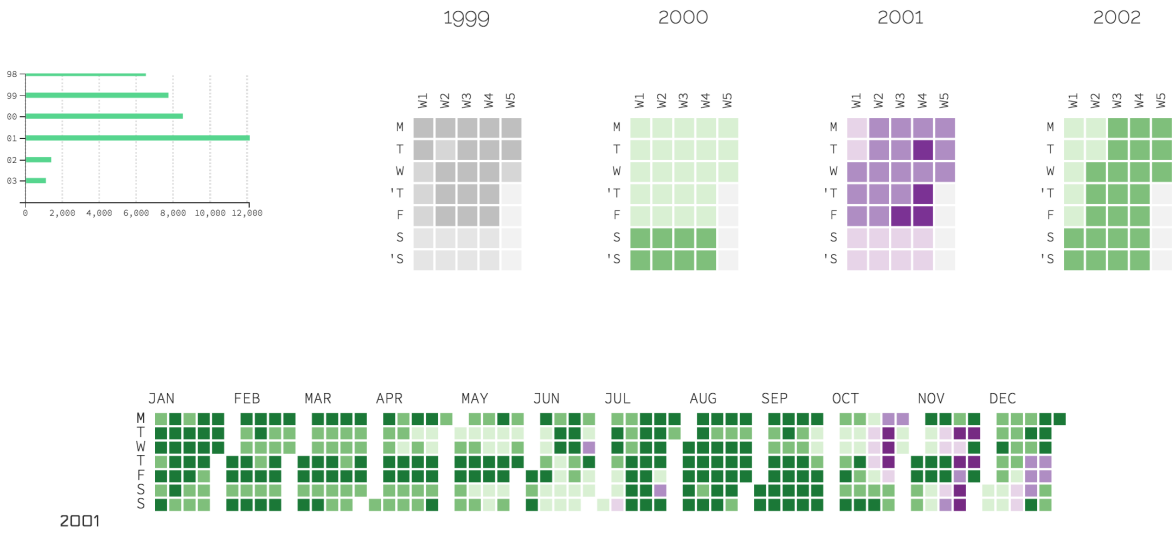

Fig. 3. Overall E-mail Behaviour: total number of emails sent in all the years is shown in the bar graph. The emails over a period of four years is investigated and the year 2001 is interesting as we can spot anomalies (purple color). The specific calendar chart (year 2001) is compared on a monthly basis, where months Sep-Dec is of interest as anomalies are spotted.

hierarchy to ensure that the most important information is lightened or made larger to read first, and others softened. This will help in telling a story. Visualisations must be useful and effective for displaying patterns, relationships and comparisons.

15) **Usability:** visualisations must help to accomplish user goals quickly and easily, ideally in order that large amounts of search effort on the collection as a whole can be avoided.

V. Demonstration of Visualisation for a Specific E-discovery Problem

From the above discussions, we tried to address two three challenges: improve comparison of subsets of data, and identify sensitivity/anomalies in email communication

Datasets: We identified few available E-mail corpora for studies related to E-mail and they are: British Columbia Conversation Corpus (BC3) [25], The World Wide Web Consortium Corpus (W3C) [26], The CSIRO Corpus [27], PW Calo Corpus, The Enron Corpus (EC) [28], Enron Sent Corpus [29], Hillary Clinton Email Dataset [30], European Union E-mail Communication Network [31], Attachment Prediction Dataset [32], Person Name Annotations [33], Conversation Threads, Multi-Lingual Conversations, Communication Network [34], Avocado E-mail Dataset [35], Jeb Bush Emails [36][37], Customer Interaction Data of German Emails and Online Requests [38], Spam email datasets [39], ECUE Spam E-mail Datasets [40].

For implementation purpose, certain criteria were considered: an E-mail corpus must have a rich collection of E-mails, must be real-one, publicly available to access, useful for investigation purpose, must contain features such as temporal, connections and context. In the survey, only two datasets with case studies, Enron [41] and Hillary Clinton [30] dataset, matched the criteria and hence the reason for using the datasets for implementing the framework and addressing the investigation tasks.

Case Study: The Enron [41] controversy is a well-known case in the E-discovery field, and it serves as a real-world benchmark given that as the nature of Enron's email data was private and unstructured. Enron produced fake profit reports and company's accounts which led to bankruptcy. Most of the top executives were involved in the scandal, as they sold their company stock prior to the company's downfall. Enron email is available for the public to access. In our work, we value the Enron data as a valuable test case, as it contains real-world distributions, challenges and tasks with respect to various features and noise. The Enron email archive contains more than 200,000 emails from the year 1999 to 2002. There are many missing individuals and emails in the original dataset, for unknown or possibly sensitive reasons. In this paper, a reference unit for comparison was defined, then investigative units were computed and visually encoded. Considering various investigative tasks, we developed comparative designs for the three main features (temporal, connections and context) using the Enron data. In general, a user can choose the particular unit(s) to be compared with the investigative units.

Analysis of the Email Communication: In the Enron email scandal, the question we had in our mind was "How to effectively use visualisations to identify normalities, similarities (commonalities), differences, abnormality and to make comparative decisions efficacious within subsets of data?". In previous investigations by researchers, there is little in the way of a deep analysis with respect to identifying (ab)normalities within time-frame, individuals and context, and then compar-

Fig. 4. Specific E-mail Behaviour: total number of emails sent in 2001 by the top individuals is shown in the bar graph. The specific months of interest for the selected individuals is compared with the average values for the year. The matrix relationship reveals Jeff had a strong relationship with James, Kay and Kaminski.

ing and identifying changes within the subset of data chosen. Our design solutions help us to achieve the above-mentioned improvements, and they are discussed below with visualisations. In this email analysis, we considered calendar charts for representing temporal behaviour, matrix charts for representing individuals' connections, and bar charts for representing context (influential words) used in the emails. We considered the small multiples approach for visualising the data in a way that increases understanding and identifying similarities/differences and anomalies. Based on the discussion with our legal experts, we considered these visual representations with simple interactions (hovering and selecting) as a suitable solution for the investigative domain problem, as the analysts are not tech-savvy. Based on the color brewer theory [42], we used a sequential single hue (light grey (low effect) to dark grey (high effect)) to indicate a fixed reference unit (i.e. in volume) and diverging colors (dark green (low effect) to dark purple (high effect)) to indicate investigative units (i.e. computed difference) and/or varying reference units. Since the comparison of means is a common analytical task with some limitations [43], we compute the differences between the means of two subsets of data, where the resulting measure is known as the effect size [44]. The Enron email analysis was developed using Data-driven documents (D3.js) [45] - shown in the Fig. 3,4,5,6.

VI. Conclusion and Future Work

This paper examined the opportunities and challenges facing digital communication data, E-discovery, and analytics, and identified problems that in our view require visualisation support and guidance. From the discussions, we tried to address the challenges of improving the comparison of subsets of

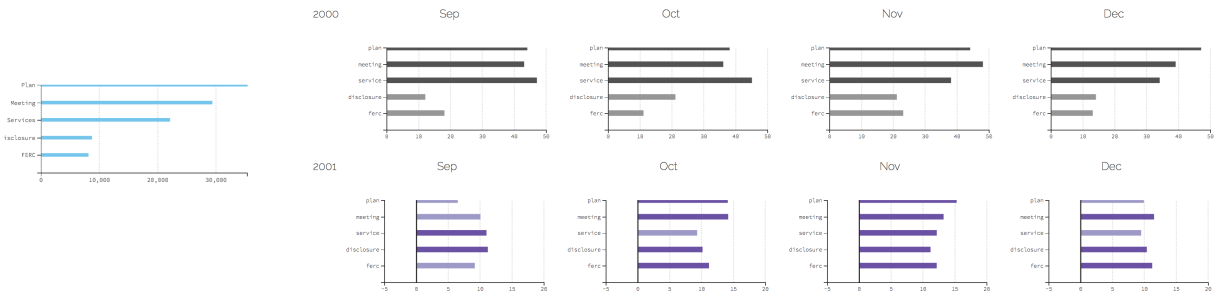

Fig. 5. Context Behaviour: Investigating on the words (context) used: the total number of words used in the year 2001 is shown in the bar graph. The months of interest (Sep-Oct) in 2001 is compared with the same months in 2000. All the selected words have been used many times in the months of interest.

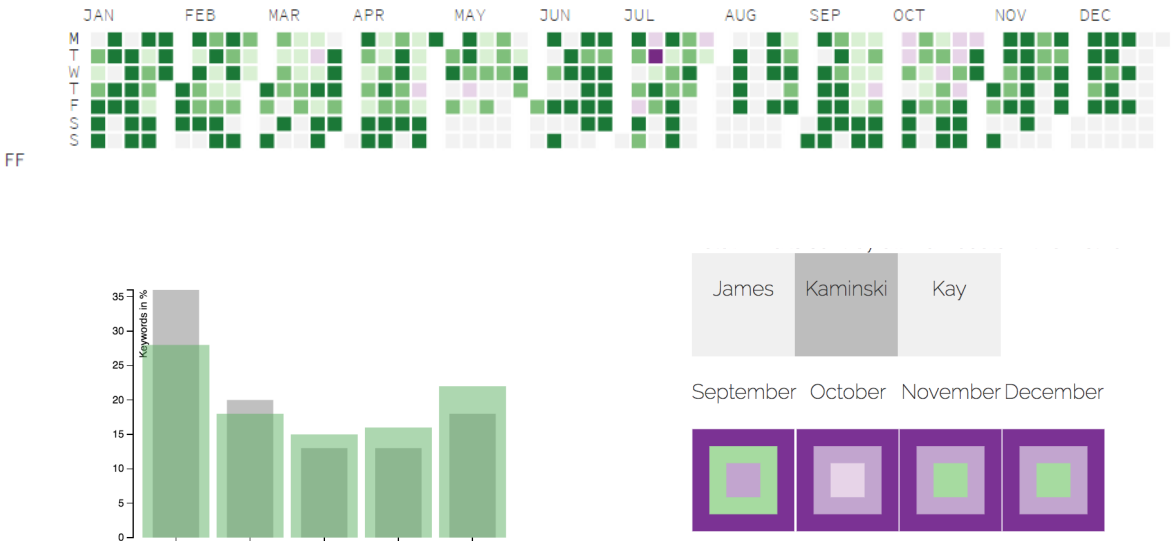

Fig. 6. Individual Behaviour: Jeff's email communication over the year of interest is monitored (Oct and Nov is of interest). His relationship with all the other three individuals (James, Kay and Kaminski) became very strong in the same months. Jeff used words such as services, disclosure and ferc the most when compared to the other individuals in the same month of interest.

data, and identifying sensitivity/anomalies in email communication [46][47][48][49][50]. As a future work, we aim to develop a simple, powerful, effective, efficient and analyst-friendly visualisation tool which will be have the feasibility to understand sensitivities and anomalous behaviours as well as what constitutes "normal", "pertinent", "interesting", and "relevant" data. We will attempt to have the tool understand the dynamic changes between two subsets and the underlying communication structures in email communication, which will help in E-discovery and investigations. We aim to deliver innovation on several fronts: Developing novel combinations of visual and algorithmic analysis: the complexity of the data requires us to not only utilise and improve state-of-the-art intelligent algorithms in data analysis but also calls for novel techniques where humans' cognitive capabilities are fostered. The potential of such novel combinations in information discovery and decision making within E-discovery domain has not in our view been adequately investigated; thus,it remains an innovation we want to exploit in this project. Text analytics such as automated Named Entity Recognition or Classification of Email categories will aid in providing valuable data pre-processing/analysis. Also, we will consider text visualisation in order to provide effective views for the processed data. The complete version of the tool will have user testing, using AmazonTurk to evaluate the visualisation design choices for some of the tasks, such as aggregation, comparison, etc. Our proposed methodologies will help analysts in their E-discovery and investigative tasks through interactive and visual analytics and promises to lead to faster and effective processes.

VII. Acknowledgments.

We would like to thank Rahul Powar and Randal Pinto of Red Sift Research for their insightful comments and enlightening us with their corporate knowledge and supporting us by funding. We would also like to thank people from the giCentre, City, University of London, UK and Jason Baron of Drinker Biddle & Reath LLP for their timely inputs and guidance.

REFERENCES

[1] E. Casey, *Handbook of digital forensics and investigation*. Academic Press, 2009.

[2] V. L. Lemieux and J. R. Baron, "Overcoming the digital tsunami in e-discovery: is visual analysis the answer?" *Canadian Journal of Law and Technology*, vol. 9, no. 1 & 2, 2011.

[3] G. Socha and T. Gelbmann, "The electronic discovery reference model (edrm)," http://edrm.net/, 2009.

[4] D. Lawton and R. Stacey and G. Dodd, "Uk home office," https://www.gov.uk/government/publications/ediscovery-in-digital-forensic-investigations, 2014.

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

[6] J. Stasko, C. Görg, and Z. Liu, "Jigsaw: supporting investigative analysis through interactive visualization," *Information visualization*, vol. 7, no. 2, pp. 118–132, 2008.

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

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

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

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

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

[12] C. Collins, S. Carpendale, and G. Penn, "Docuburst: Visualizing document content using language structure," in *Computer graphics forum*, vol. 28, no. 3. Wiley Online Library, 2009, pp. 1039–1046.

[13] S. Whittaker, Q. Jones, B. Nardi, M. Creech, L. Terveen, E. Isaacs, and J. Hainsworth, "Contactmap: Organizing communication in a social desktop," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 11, no. 4, pp. 445–471, 2004.

[14] W. Sack, "Conversation map: An interface for very large-scale conversations," *Journal of Management Information Systems*, vol. 17, no. 3, pp. 73–92, 2000.

[15] S. J. Luo, L. T. Huang, B. Y. Chen, and H. W. Shen, "Emailmap: Visualizing event evolution and contact interaction within email archives," in *Visualization Symposium (PacificVis), 2014 IEEE Pacific*. IEEE, 2014, pp. 320–324.

[16] M. E. Joorabchi, J.-D. Yim, and C. D. Shaw, "Emailtime: Visual analytics of emails," in *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*. IEEE, 2010, pp. 233–234.

[17] M. Samiei, J. Dill, and A. Kirkpatrick, "Ezmail: using information visualization techniques to help manage email," in *Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on*. IEEE, 2004, pp. 477–482.

[18] S. L. Rohall, D. Gruen, P. Moody, and S. Kellerman, "Email visualizations to aid communications," in *Proceedings of InfoVis 2001 The IEEE Symposium on Information Visualization, IEEE*, vol. 12, 2001, p. 15.

[19] https://www.globanet.com/sites/default/files/resources/Key%20Issues%20in%20eDiscovery%20-%20Globanet.pdf.

[20] http://www.watsonhelsby.co.uk/assets/files/Digital-Communications-Social%20Media-The-Challenges-facing-the-PR-industry.pdf.

[21] http://www.dbjournal.ro/archive/13/13_4.pdf.

[22] www.informationbuilders.co.uk/intl/co.uk/presentations/four_types_of_analytics.pdf.

[23] http://www.eis.mdx.ac.uk/vass/VASS2012/doc/pdf/VASS2012-3Sep-JasonBaron.pdf.

[24] A. Dasgupta, J. Poco, Y. Wei, R. Cook, E. Bertini, and C. T. Silva, "Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison," *IEEE transactions on visualization and computer graphics*, vol. 21, no. 9, pp. 996–1014, 2015.

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

[26] http://w3c.org.

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

[28] http://EnronData.org.

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

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

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

[32] http://www.cs.jhu.edu/m̃dredze/code.php.

[33] http://www.cs.cmu.edu/ẽinat/datasets.html.

[34] http://khorshid.ece.ut.ac.ir/m̃.dehghani/emaildataset.html.

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

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

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

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

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

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

[41] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Machine learning: ECML 2004*. Springer, 2004, pp. 217–226.

[42] C. A. Brewer, "Color use guidelines for mapping and visualization," *Visualization in modern cartography*, pp. 123–148, 1994.

[43] R. A. Johnson, D. W. Wichern *et al.*, *Applied multivariate statistical analysis*. Prentice hall Upper Saddle River, NJ, 2002, vol. 5, no. 8.

[44] S. Olejnik and J. Algina, "Measures of effect size for comparative studies: Applications, interpretations, and limitations," *Contemporary educational psychology*, vol. 25, no. 3, pp. 241–286, 2000.

[45] M. Bostoc, "Data-driven documents," https://d3js.org/, 2011.

[46] M. Sathiyanarayanan and C. Turkay, "Is multi-perspective visualisation recommended for e-discovery email investigations?" 2016.

[47] M. Sathiyanarayanan and C. Turkay, "Determining and visualising e-mail subsets to support e-discovery," 2016.

[48] M. Sathiyanarayanan, C. Turkay, and J. Dykes, "Visual comparison strategies for small multiples of digital communication data," 2017.

[49] M. Sathiyanarayanan and C. Turkay, "Improving visual investigation analysis of digital communication data within e-discovery," 2017.

[50] M. Sathiyanarayanan and T. Mulling, "Wellmatchedness in euler diagrams: An eye tracking study for informationvisualisation evaluation," in *Proceedings of the 8th Indian Conference on Human Computer Interaction*. ACM, 2016, pp. 70–74.