International Journal on Digital Libraries
https://doi.org/10.1007/s00799-024-00394-5

Towards privacy-aware exploration of archived personal emails

Zoe Bartliff[1] · **Yunhyong Kim**[1] · **Frank Hopfgartner**[2,3]

Received: 25 April 2023 / Revised: 12 January 2024 / Accepted: 14 January 2024

Abstract
This paper examines how privacy measures, such as anonymisation and aggregation processes for email collections, can affect the perceived usefulness of email visualisations for research, especially in the humanities and social sciences. The work is intended to inform archivists and data managers who are faced with the challenge of accessioning and reviewing increasingly sizeable and complex personal digital collections. The research in this paper provides a focused user study to investigate the usefulness of data visualisation as a mediator between privacy-aware management of data and maximisation of research value of data. The research is carried out with researchers and archivists with vested interest in using, making sense of, and/or archiving the data to derive meaningful results. Participants tend to perceive email visualisations as useful, with an average rating of 4.281 (out of 7) for all the visualisations in the study, with above average ratings for mountain graphs and word trees. The study shows that while participants voice a strong desire for information identifying individuals in email data, they perceive visualisations as almost equally useful for their research and/or work when aggregation is employed in addition to anonymisation.

Keywords Email visualisation · Privacy · Archives · Perceived usefulness · Research data · Data management

1 Introduction

Email has been referred to as 'the backbone of the internet', a 'virtual working environment' and the 'main means for distributed collaboration' ([1]). An email collection is an organically formed record that documents both important and everyday moments in an individual's life and work. The extent of information that can be extracted from such a dataset makes email collections a rich source for investigating patterns of human behaviour, relationships, and communications (cf. [2–7]). However, there are caveats to the valuable nature of this data, most notably the enduring ethical concerns provoked by facilitating access to such personal content. Email research often thrives on the details of individual lives and connections with others, information that can be deeply private, sensitive, and/or confidential in nature. The challenge has hitherto encouraged a caution-driven practice of closing or severely restricting access to collections.

Scholars and custodians of data alike have explored and implemented a great range of methods for accessing and exploring email collections (e.g. [2, 3, 8–10]), and yet the impact of these with regard to privacy preservation is not widely discussed nor, seemingly, understood ([11]). This partly reflects the complexity of thoughts surrounding privacy in itself (cf. [12], [13, 14]). Regardless, this disconnect amplifies continued uncertainty, resulting in a 'risk-adverse attitude' (cf. [15]) amongst custodians of data. Consequently, a great swathe of potential research data remains locked within closed or 'dark' archives ([3, 4, 15–17]). Whilst preventing access might be '[t]he most intuitive way to preserve privacy' ([18]), it also, in many ways, defeats the purpose of maintaining the records, particularly in instances where the relevance of the data might be time sensitive ([11, 15]).

Zoe Bartliff, Yunhyong Kim and Frank Hopfgartner have contributed equally to this work.

✉ Zoe Bartliff
zoe.bartliff@glasgow.ac.uk

✉ Yunhyong Kim
yunhyong.kim@glasgow.ac.uk

✉ Frank Hopfgartner
hopfgartner@uni-koblenz.de

[1] School of Humanities, University of Glasgow, 11 University Gardens, Glasgow G12 8QH, UK

[2] Institute for Web Science and Technologies, Universität Koblenz, Universitätsstraße 1, 56070 Koblenz, Germany

[3] Information School, University of Sheffield, 211 Portobello, Sheffield S1 4DP, UK

Published online: 21 February 2024

This is the second of three key challenges that Lise Jaillant identifies the archival sector to be facing along 'the path from the appraisal of records to their analysis'([16]).

Even in cases where an email archive is not 'dark', discoverability is a continued issue with a heavy reliance on search infrastructure and accurate metadata and cataloguing ([16]), as well as a demand on the end user to have 'a rough idea of the information they are trying to retrieve' ([19]). In response, data visualisation has been used in many facets of email research to support the holistic, creative, and perhaps even 'playful' (cf. [20]) exploration of email datasets. They have been shown to reveal patterns and insights that may otherwise be obscure to researchers (cf. [21–25]). The exploratory and browsing behaviour encouraged by visualisations (cf. [26, 27]) is of particular use for high volume data. They 'capitalise on the characteristics of digital sources' ([28]) facilitating a malleable perspective on a collection. In short, visualisations represent a method that may support both researchers and practitioners to engage usefully with email collections, irrespective of pre-existing data analysis skills.

Although many have noted the value visualisations have for research-enabling interface to email collections, the understanding of how the design of a visualisation interacts with, protects, or compromises privacy is understudied. Without an understanding of the impact of the visualisation on privacy, it is possible that the method of mediation might negatively impact upon access or open the data to reveal 'previously unknown patterns and relationships' ([11]) that might, contrary to intention, compromise privacy.

It is within this knowledge gap that the research presented within this paper sits. It presents findings from an empirical investigation on the potential for visualisations to facilitate access for users and provide a degree of protection for any personal or sensitive data contained within a dataset. Through these findings, this paper intends to promote a greater understanding of the relationship between privacy management strategies and the impact that these might have on the perceived usefulness of visualisations to users. This, in turn, might support both researchers and archivists 'to capitalise on the information available to them at the appropriate scale of privacy' ([11]), therefore mitigating the need to close email archives to adhere to the legal and ethical requirements of engaging with sensitive data. Should such an approach prove fruitful, it would fall within the calls for archivists and other custodians of knowledge to 'consider very different types of access' that more closely reflect user needs ([29]).

In the next section, we start by setting the scene to explain our approach to selecting and implementing visualisations in our user case study. This is followed by a detailed methodology of the user study (Sect. 3). Section 4 sets out the findings from the study which, in turn, is followed by a reflective discussion in Sect. 5, that considers the results of the study, their implications, and future work that might be conducted in this area.

2 Background

Our approach to the current study is developed through three steps. First, previous email research is reviewed, especially where data visualisation techniques have been employed and/or evaluated (Sect. 2.1). Second, ethical concerns for digital archives are also discussed, with a special focus on concerns associated with privacy and email collections (Sect. 2.2). Finally, in Sect. 2.3, we explain how we bridge these areas, to formulate our research questions and to select and generate our visualisations for our user case study.

2.1 Email visualisation

Research related to emails often poses questions concerned with understanding how people use email for communication and what this can reveal about them, their environment ([30–33]), and their social/professional network ([34–37]). Building indirectly on this understanding of email usage are studies aimed towards improving the efficiency and efficacy of communication workflows ([1, 38, 39]), and the filtering out of unwanted communication ([40–42]). Additionally, in the humanities, email data research naturally aligns with that of older forms of correspondence such as letters (cf. [43]), for example, the close reading of selected passages for qualitative analysis in the context of other events and achievements in their lives. Features such as the metadata found in email headers (e.g. time stamps, subject, who is sending and receiving) help broaden this context, to open up the researchers' gaze to a wider array of analysis than its technological predecessors might have allowed.

A systematic classification of email research ([11]) reveals two strands of thought (cf. Fig. 1)—one with the focus of enquiry on people (e.g. the patterns of relationships and social network analysis), and one which concentrates on the emails themselves and their usage (e.g. topic identification, content analysis, patterns of behaviour).

The use of specific types of visualisation has, on the whole, been agnostic of these branches of research ([11]), although there are exceptions to this with, for instance, studies focused on social network analysis prioritising network graphs (cf. [24, 44–53]). The great variety and adaptability of visualisations ensure that many common designs (e.g. bar charts [24, 49, 54–56], line graphs [2, 24–26, 49, 54, 55, 57, 58], scatter/bubble plots [46, 54, 57–59], pie charts [60]) can be adapted to diverse research objectives. There have been several, more specialised types of visualisations that were employed across the spectrum of research interests, such as timelines ([25, 57, 59, 61–63]), heatmaps ([64]) and icono-

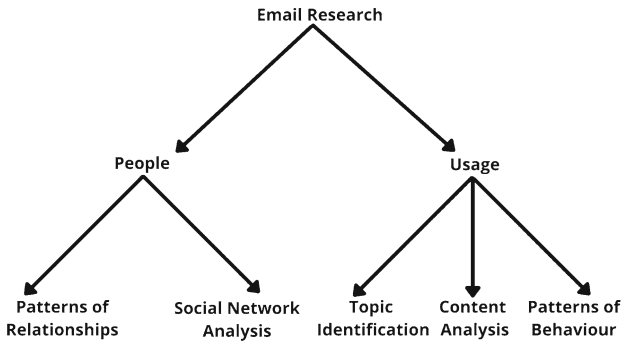


Fig. 1 A visual representation of the dyadic categorisation of email analysis and the more nuanced categories that sit within this. Source: [11]

graphic representations ([65–67]), and some studies even creatively combine visualisations in a hybrid approach (e.g. [22, 25, 45, 48, 48, 53, 63, 65, 66, 68]).

The literature shows that network graphs, of various types (e.g. random, force directed, tree), are notable as a mainstay of social network analysis research (cf. [24, 44–53]) with all 18 items reviewed in this area using this visualisation. The research for patterns of relationships employs a more varied selection with no particular preference: including widely popular visualisations such as scatter and/or bubble plots ([54, 59]) to newer visualisations such as mountain graphs ([57]). Bar charts are most regularly used in literature for studies investigating patterns of behaviour, although, as a mainstay of visualisation creation, they also appear in studies focused on other branches of investigation (cf. [24, 49, 54–56]). Email content analysis 'aims to help users navigate a collection and withdraw meaningful data whether as a search or summary mechanism' [11] and, as with many forms of textual analysis, the forms of visualisation used are quite broad (cf. [25, 52, 57, 57, 60–63, 69, 69, 70, 70]). The word tree visualisation is one of these (cf. Fig. 9), a type of visualisation that has proved useful for the early stages of textual exploration (cf. [71–74]) and, as such, will be employed in our study.

In exploring the 'state-of-the-art' approaches to visualisation design, [75] highlights several criteria that encompass successful visualisation. They indicate that data visualisations should be 'familiar', 'able to convert abstract information' in a way that 'preserves its underlying meaning but also provides insights to the user'. In each of the studies above, it is argued, if indirectly, that the method of visualisation utilised fulfils these criteria, therefore creating a useful interface for the potential users (cf. [2, 25, 26]). These studies, however, centre their focus on the particular features of the visualisation under investigation, rather than exploring the broader applications or benefits of the design outside of the stated purpose. Therefore, whilst the visualisations might be well suited to the task at hand and they might also fulfil the key criteria of good design established by [75] and other scholars, it is not possible to extrapolate meaningfully from these studies as to what might benefit the sector as a whole, particularly with reference to user needs.

2.2 Email collection ethics

Within the context of email collections, it is necessary to advance discussions of email visualisation beyond the immediate needs of the researcher to also address questions of ethical needs. Emails, in their raw form, not only contain information that can identify people by name, email address, and/or affiliation, but contain detailed information about locations, events, and relationships between people. Metadata alone can be used to infer identities, and sensitive and/or confidential information. For example, "e-mail headers reveal who is central to your professional, social and romantic life"([76]). The access to such collections creates opportunities for "private information within these collections to be disseminated widely and without consent" ([77]) even where it creates opportunities for much needed research ([78]).

Emails also often have a tendency to include information beyond that which is written or intended to be received by the email account owner, or worse, those who access it later. For example, emails have attachments which could, if distributed further, entail copyright infringement or communication of privileged, proprietary, or confidential information. In established archival practice, it is standard practice to consider materials of long deceased individuals of less risk of disclosure. Even when the primary owner of the email is deceased, content is directly associated with others who may still be living, potentially causing distress or issues of privacy. This challenge is compounded by the potential for the emails of others to get copied in as a thread and, sometimes, even sent to individuals who were not intended to have access. Effectively, when you archive emails in one person's personal archive, you are archiving other people's emails as well.[1] It has to be recognised that when it comes to digital forms of communication, it is not always possible for creators to be aware how the information would be used in later contexts and can interfere with an individual's right to forget ([79]). In addition, some laws and/or regulations stipulate that the control of the data needs to take into account cultural needs.[2]

Privacy management is an especially thorny and shifting concept ([12]), and an intersection of research relevant to email research and visualisation which has, thus far, remained largely unexplored. The majority of the studies identified in

[1] https://mako.cc/copyrighteous/google-has-most-of-my-email-because-it-has-all-of-yours.

[2] Australian Laws https://osf.io/68wp4.

[11] made no mention of privacy, and nearly a quarter of the studies were tested on participants' own email collections, or in a slightly smaller sample, on the popular open source email dataset, the Enron dataset.[3] It is highlighted that only two out of the 39 reviewed papers engaged with a personal email archive (cf. [25, 56]) and, of these, one involved the owner of the archive as a co-author. This is a distinct gap within literature pertaining to email visualisation research, one which has arisen, at least in part, due to the difficulties involved in defining privacy.

Debated in scholarship at least since the philosophies of Aristotle ([80]), little has been agreed about the definition of privacy other than that it is a multifaceted concept encompassing legal, ethical, cultural and personal dimensions. There have been 'many attempts to create a synthesis of existing literature' ([81]), but the default approach to protecting privacy for many institutions, archives included, has necessarily been to rely on the more concrete legal definitions of, for example, personal and sensitive data,[4] as well as on the ethical mandate to limit harm.

This situation is not one that will improve with time. In 2012, it was noted that approximately '75% of the email accounts belong to individual users, with only 25% belonging to organisations' ([19]). This statistic is more than a decade old at the point of writing and, therefore, does not necessarily reflect the proportions of email data that are destined to be archived in coming years. Whilst the 'risk adverse attitude' (cf. [15]) of present custodians of data is quite logical given the potential ramifications from mis-managed email data, the great swathes of incoming, culturally significant data necessitates the inclusion of alternative approaches in order to facilitate effective user-driven access and, therefore, research.

There is not, at present, a nuanced and consistent approach for managing privacy with respect to email collections, although [11] presents the first steps towards this. The paper explores existing literature pertaining to the visualisation of emails and the impact of different design choices on the level of privacy consciousness. The five privacy consciousness (PrivCon) levels discussed in the paper ([11]) represent a scale of privacy management strategies that might be applied to the data that forms the basis for different visualisations. These strategies range from full disclosure (PrivCon 0) through to closed to public access (PrivCon 4) with each level representing a category of privacy management as opposed to a specific method. The description of the levels is reproduced in contracted form below:

- $PrivCon0$—the open end of the scale with no accounting for privacy, there are visualisations that contain the full range of the data as would be utilised in, for example, in a forensic examination of the data or an archivist's appraisal when a full collection has been donated.
- $PrivCon1$—the introduction of redaction that 'includes situations whereby the data have been altered or removed in order to obscure the identity of individuals contained within'.
- $PrivCon2$—'the grouping or amalgamation of data to the point that individuals become 'lost in the crowd', minimising the risk that details might be identified'.
- $PrivCon3$—the introduction of noise which 'involves shifting the data through the use of an algorithm, statistical model or encryption, in a way that maintains the statistical characteristics of the data set, but the detail does not consistently reflect the original'.
- $PrivCon4$—it represents a closed collection which has been fully redacted and contain only a descriptive representation of the collection with only a cursory indication of contents. This presentation of the data represents what might be found in an online catalogue for an archival collection that only permits on-site access, or for a fully embargoed collection.

The manner in which the PrivCon levels (0-3) might be applied to a dataset is displayed in Fig. 2. The paper ([11]) reveals a skewed distribution of approaches in the thirty-nine papers reviewed, with a clear tendency leaning towards anonymised/pseudonymised content (PrivCon 1). This is summarised in Table 1 along with pros and cons of each type of strategy.

2.3 Research questions and approach

The present paper sits at the intersection of the knowledge gaps identified above, seeking to explore how email visualisations research, privacy, and useful user-driven access might interact.

To investigate the interplay between usefulness and privacy with regards to visualisations, we generated visualisations utilising data drawn from a filmmaker's personal digital archive. We then filtered these through the privacy-aware strategies that reflect the PrivCon levels discussed in Sect. 2.2. The visualisations were presented to researchers and archivists to explore the extent to which each type of visualisation and each level of privacy was perceived to be useful to their respective workflows.

To advance this research towards a practical solution for privacy management in the archive, this paper engages with

[3] https://www.cs.cmu.edu/~enron/.

[4] Legislation relevant in a UK context includes: European Convention on Human Rights (ECHR); the Charter of Fundamental Rights of the European Union (CFREU); the General Data Protection Regulation(EU) 2016/679 (GDPR); the Freedom of Information Act 2000 (FOIA) and Freedom of Information (Scotland) Act 2002 (FOISA).

Table 1 Distribution of PrivCon levels adopted in the literature review of [11], collection type, and pros and cons of each level

Level	Papers	Example data	Pros	Con
PrivCon 0	13	Forensic evidence	Low labour	High risk
		One's own email data	Full access by user	
PrivCon 1	30	Enron	Some protection	Identity could be inferred
		Email datasets for ML	Automated tools exist	Reliability can vary
			Details accessible	
PrivCon 2	6	Enron curated	Can track flow, activity	Aggregated stats only
		Commercial datasets	Trends accessible	Multiple searches can reveal
PrivCon 3	0	–	Secure in theory	Difficult to implement

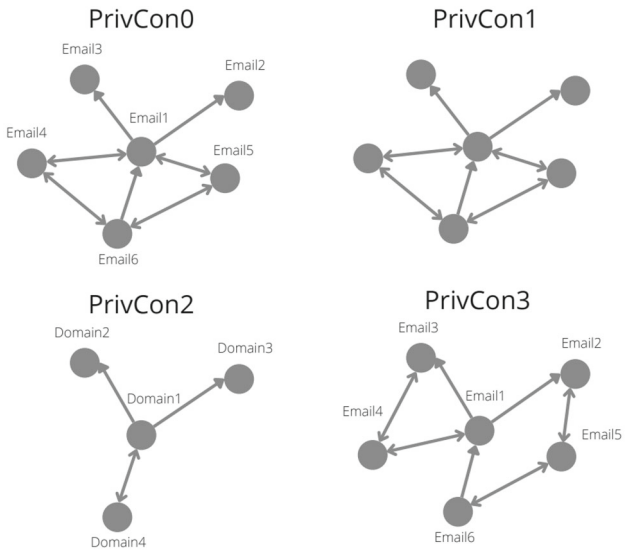


Fig. 2 A representation of different PrivCon levels. PrivCon 0 is a direct representation of the data. PrivCon1 has had the sensitive data removed. PrivCon2 has been amalgamated, in this instance by domain type. PrivCon3 has had noise introduced, at random, to obscure the relationship between certain data points

arts and humanities scholars as well as archive practitioners to explore how perceived usefulness of email visualisations changes for these stakeholders as data are curated to respect different levels of privacy. It approaches this in response to the following research questions:

- RQ_1: What is the relationship between the extent of privacy-awareness applied to visualisations of email collections and the usefulness of these visualisations to researchers/practitioners?
- RQ_2: What design features of the privacy-aware visualisations are the most/least useful for researchers and practitioners as an interface for the email collection?

Given the great variety of visualisation designs in the literature (Sect. 2.1), it was necessary to select a subset to feature within this study. Visualisations are selected to include at least one used in the branches of research identified in [11]. Of the five branches of research (cf. Fig. 1), the area of topic identification was excluded from this study. Topic identification was most often keyed towards the removal of spam, or the automatic categorisation of content (e.g. [64, 82]). Within the bounds of the current study, and given the scope of the available data, it was not deemed viable to train a model for the automatic detection of content. In fact, within archives, the use of AI and machine learning is a relatively new, but growing field ([83]). Whilst it may be possible, at a later date, to incorporate such methods as standard within an archival setting, the sector has not yet arrived at this point.

In Sect. 2.1, we noted the prominence of network graphs for social network analysis. For this reason, we included directed network graphs as one of our visualisation for the study. We further noted that other categories of research were not inclined towards any particular visualisation. As such we selected establish standards for patterns of relationships and behaviour: scatter plots and bar graphs, respectively. We further included two newer forms of visualisations, mountain graphs, and word tree, observed in the literature to be useful for analysing patterns of relationships and content analysis. Examples of all these visualisations are shown in Sect. 4.3.

We describe below how each type of visualisation was implemented to incorporate privacy management strategies reflecting PrivCon levels (Sect. 2.2).

Implementation of network graphs

For the study, the participants were presented with four network graphs[5] at each level of the PrivCon Scale (0-3). At PrivCon0, the visualisation contained full email addresses. These were removed for PrivCon1, leaving only the shape of the connections for interpretation. PrivCon2 involved aggregating the dataset, grouping data points by email domain

[5] The network was created using the Python NetworkX's DiGraph class with visualisation realised using nx.draw, which engages the matplotlib library.

name. The most stringent level of privacy, PrivCon3, is implemented using noise introduced to reflect the general distribution but to avoid issues of identity reconstruction noted with network graphs ([84–87]).

Implementation of mountain graphs

The mountain graphs[6], otherwise known as stacked line graphs, have proved quite popular in literature to present and portray the ebb and flow of relationships over time. For this study, each layer in the graph represents a unique contact and the area within the layer demonstrates the extent of email communication (to, from, CC and BCC) on the given date denoted by the x-axis. Visualisations were created for PrivCon levels 0, 1 and 2. PrivCon0 included a key with full email addresses for each contact. For PrivCon1, the key is removed as a form of anonymisation. For PrivCon2, as a form of aggregation, the emails were grouped utilising the personal, professional, shopping, practical categorisations (discussed in Sect. 3.1). These categories were then used as the different layers of the graph.

Implementation of scatter plots

The second type of visualisation selected, scatter plots,[7] is in line with studies such as [54, 58, 59]. These were created showing the points of contact for each email throughout the date range of the dataset. The email contacts were arranged in order of frequency, from the highest number of contacts to the lowest. Colour was used to denote whether the point of contact was To, From, CC, or BCCing the individual. These visualisations were included only as PrivCon0 and PrivCon1. The former included the email addresses listed on the x-axis and the latter has these redacted.

Implementation of bar graphs

For this study, bar graphs[8] are used to show the number of contact points in the email collection on any given day as well as the type of contact (To, From, CC, BCC). For clarity, the dataset has been limited to show only those contacts with greater more than one connection edge. The PrivCon0 graph depicts the frequency and type of contact (to, from, cc, bcc) for each of the higher frequency email addresses. PrivCon1 is similar to this, but with the email addresses redacted. The PrivCon2 graph aggregates the activity and presents it by date rather than by individual email addresses. This is a similar presentation for the PrivCon3 graph, but, for that graph, noise has been introduced to limit the potential for reconstructing identities.

Implementation of word tree

For this study, the word trees[9] are presented as PrivCon0-2. For PrivCon0, the visualisation includes the word tree with a reading panel on the right-hand side that allows the participant to see the content portrayed in the visualisation in context of the collection as a whole. At PrivCon1, the sensitive information (names, email addresses etc.) has been redacted in both the visualisation and the reading panel. As a form of aggregation, for PrivCon2, the contextualising reading panel is removed.

3 Methodology

3.1 The data

The dataset used for this study is a personal email archive of a filmmaker who used their email account for both personal and professional purposes. Their professional activities revolved around the conception and creation of *avant garde* films. They attended a range of conferences and film festivals and were in contact with film institutes and artist support networks with the intent of archiving their life's work. They worked with students on their own projects and they engaged in a variety of other artistic pursuits. They maintained a strong network of personal and professional relationships, with many individuals sitting within both categories. Furthermore, as a disabled artist, the filmmaker engaged with a variety of individuals and organisations to support them day to day. Each of these activities, alongside things like internet shopping, has left their trail within the email collection, making it a rich source for exploring the artist's life and professional activities.

The email collection was recovered from the legacy internal hard drives of a Mac OSX desktop used by the late filmmaker. The recovered data comprise 5095 emails spanning the time period 2006-2012 (the year of the artist's death), although there are only a very small number dating back to 2006 and nothing after that until 2009. For our study, in addition to the email content, associated metadata was extracted to comprise: email address of sender; email address of recipient; email address of people copied into the email; the date on which it was sent. Although it was not used for the current

[6] Created using Python library Pandas' plot.area.

[7] Created using the Python library Plotly's go.Scatter.

[8] Created using the Python library matplotlib's pyplot.bar with the stacked option.

[9] At the time of conducting the study, there was no Python library for creating word trees. As such, and due to the need for a reading panel, it was decided to use a pre-built word tree creator (jasondavies.com/wordtree/). For future developments of these visualisations, there would be additional funding and resources to create or implement bespoke code more keyed towards the needs of the archive.

study, broader metadata included information on whether the email is junk, has been read, has a high or low priority—an additional layer of complexity for an already complex datatype ([88]) for further exploration in future studies.

To offer more digestible visualisations for the participants of this study, the data source was sampled to include two months of the filmmaker's life. We targeted the date range from December 2010 to January 2011 (351 emails), chosen to offer a good range of personal and professional emails to represent a number of known major events—a major holiday period and the period of production for what would be his last major film. To ensure that the dataset was focused on emails with a high level of interest to the archive and researchers, the data were manually coded into one of five categories (personal, professional, shopping, practical, advertising) based on a review of the content of the email and the presence of, for example personal anecdotes, receipts or unsolicited or periodical content from institutions. This was done by one of the researchers in the research project (and an author of this paper) who had been responsible for exploring the email collection and for advising the archive on its content. Those in the advertising section (e.g. spam, circulars) were excluded from the visualisation as they were judged to be of little interest or value regarding the artist's life and work. This resulted in a dataset of 218 emails out of the two months sample or 5.4% of the collection as a whole.

3.2 The participants

Given the sensitive nature of the material under investigation and the restricted status of the featured archive for public release,[10] the participants for this study included only those individuals who had been granted privileged access to the collection as a part of the associated research project. The benefits of this arrangement were twofold:

- It ensured that participants were familiar with the central subject of the email collection, creating a facsimile of the natural process of discovery experienced by researchers or archival practitioners.
- The arrangement allowed for the testing of visualisations at all levels of the privacy scale without risking the release of sensitive data.

The participants were selected due to their range of expertise and a shared research and/or professional interest in the dataset. A more detailed breakdown of the participants' professional profiles is given in Sect. 4.2 as the findings for Stage One of this research. The participant pool included four individuals, two who were trained archivists at different stages of their career and with different day to day responsibilities as well as two who were Arts & Humanities researchers, each from different disciplines, their work characterised by quite disparate methodologies. All participants had worked with the archive in question for a number of years prior to the commencement of this study, and so they were intimately familiar with an array of contents from within the wider collection. None, however, had engaged with the email collection beyond abstract conversations and reports at team meetings.

The number of participants involved in the study may seem small, yet, it is comparable to similar studies. For instance, MUSE and the professional counterpart ePADD, used for email collections in cultural memory institutions [2, 25, 26], exploit a range of data analysis techniques to promote the exploration of email collections. In [25], the usefulness of the tool was explored through an experiment involving six participants (two archivists, a historian and three working professionals). Working with their own email collections, the participants rated the tool on a five point scale, supplementing this with qualitative comments to contextualise the responses. This model of investigation is a familiar model, repeated in many of the studies to determine the extent to which a selected visualisation design supported the participants' needs. The study is organised as an in-depth three-stage exploration of a complex problem from multiple angles (see details in Sect. 3.3) to compensate for the limited availability of participants.

3.3 The study

The research underlying this paper adopted a delphi study model, a research method designed to 'obtain the most reliable consensus of a group of experts' [89]. It involves 'a series of questionnaires interspersed with controlled opinion feedback' [89]. Specifically, the participants partook of three rounds of questionnaires, with stages two and three incorporating the chance to review aggregated feedback from the previous stage.

3.3.1 Stage One

The focus of Stage One was to establish a baseline for how participants might engage with email data. The questions asked included background information, such as research discipline, interests and common methodologies used. This contextualises the findings from later stages and helps to understand how researchers the researchers might engage with email collections as they are usually presented. The key question of this stage was: "What kind of research can you envision yourself conducting with email data?".

[10] At the time of writing. The archive will, following a full sensitivity review and cataloguing, be released for wider viewing.

3.3.2 Stage Two

In Stage Two of the research, participants were first presented with a summary of the Stage One responses regarding their present research, allowing them to consider and augment their previous response in light of the shared ideas. Then participants were shown a series of visualisations, grouped by type but with visualisation type and level of privacy ordered randomly. By varying the order of presentation, it minimised the potential for bias brought about by increased knowledge of the dataset gained throughout the survey as well as that brought about by varied levels of interest through the experiment (cf. Appendix A, Table 9). Also to minimise potential for bias, the visualisations were given a consistent colour palette, scale, and, as much as possible, presentation (font, title/key placement, background, surroundings). Each visualisation was given a brief description to aid the participant in comprehending its scope and context as well as support them in the process of interpretation.

For each visualisation, participants were asked:

1. What kinds of information can you gather from this visualisation?
2. Does this type of visualisation support your approaches to research?
3. In what ways might visualisations like this help you to address your key questions/themes and/or envisioned outcomes?
4. In what ways could the visualisation be lacking in helping you address your key questions/themes and/or envisioned outcomes?

3.3.3 Stage Three

The final stage of the survey was intended to consolidate the participant's understanding of the varied visualisation and to facilitate their ranking in terms of usefulness compared to levels of privacy protection. This was done by presenting the responses from Stage Two in a collated form for each visualisation presenting these to the participants as a part of the survey for review. The participants were then asked to reconsider the visualisations and give them a score for their usefulness, as well as give reasoning for this score. More precisely, for each visualisation they were asked to consider:

1. Is there anything you would like to add or change in relation to your initial assessment of this visualisation?
2. How useful is this visualisation for your research or practice? 1 (not useful)–7 (very useful)
3. Why have you given this rating?

Table 2 Average usefulness responses, on a scale of 1–7, for each PrivCon level

	No. of responses	Average score
PrivCon 0	20	5.000
PrivCon 1	20	3.800
PrivCon 2	16	4.688
PrivCon 3	8	2.875

Table 3 Average usefulness responses, on a scale of 1–7, for each visualisation type

	No. of responses	Average score
Word tree	12	4.583
Directed network graph	16	3.688
Mountain graphs	12	5.000
Scatter graphs	8	4.125
Bar charts	16	4.188

4 Findings

4.1 Overview

The detailed findings related to each of the three stages of the research are presented in Sects. 4.2, 4.3 and 4.4. Here we present a holistic overview, with numbers drawn from the usefulness ratings from Stage Three of the study. While the results pertain to Stage Three of the process, it is assumed that participants made their assessment informed by their experience throughout all three stages of the study.

The average responses on the usefulness scale (cf. Table 2) revealed that the five PrivCon 0 visualisations were most highly rated (5.0), across all visualisations in this category. This is followed by the four PrivCon 2 visualisation (4.69) and then the five PrivCon 1 visualisations (3.8). Those perceived as the least useful were the two PrivCon 3 (2.88).

It will be shown in the subsequent sections that the preference for the fully disclosed PrivCon 0 reported in Table 2 can actually be variable across different visualisations, suggesting that too much data can cloud the information contained within.

The average rating for all the visualisations was 4.28 and indeed for each visualisation type (cf. Table 3) was above the midpoint. This could indicate that participants are able to envisage uses within their own work for each type of visualisation. Most highly rated were the mountain graphs, a representation of the Patterns of Relationships area of research (5.0). Following after this, and closely rated with 4.58, 4.19 and 4.13, respectively, are the word trees, the bar charts, and the scatter graphs. Lowest rated (3.69) are the directed network graphs.

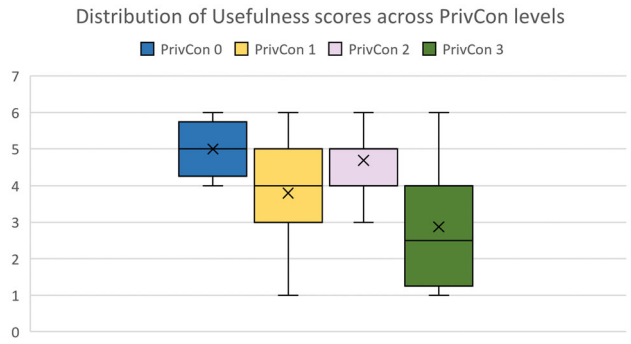


Fig. 3 A box and whiskers diagram displaying the distribution of usefulness scores for each of the PrivCon levels

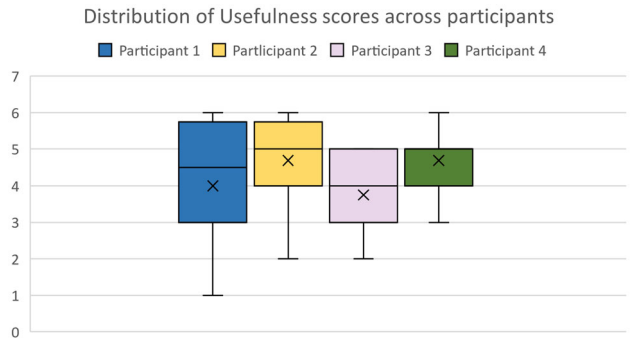


Fig. 4 A box and whiskers diagram displaying the distribution of usefulness scores for each of the participants

For a more detailed perspective on the range of responses given by participants, Fig. 3 presents the range, inter-quartile range, average usefulness scores given for each PrivCon level. This figure also demonstrates that the PrivCon 0 visualisations were consistently the highest rated, receiving only scores between 4 and 6, and with an inter-quartile range across this interval. PrivCon 2 retrieved the next highest score with a range between 6 and 3, but with the majority of responses clustered between 4 and 5. Both PrivCon 1 and 3 received responses across the spectrum of the scale, although PrivCon 1 trended towards the upper end of the scale and PrivCon 3 the lower end. This result is contrary to the expected relationship between privacy awareness and usefulness.

Another important factor to consider is that whilst participants were given a scale upon which to rate the visualisations, there is a degree of subjectivity in the interpretation of this scale.[11] As such, and to provide a point of reference, Fig. 4 displays the range, inter-quartile range, average usefulness scores given by each participant. Participant 1 gave the broadest range of responses and was the only Participant to give a 1 to the visualisations. Participant 2 trended towards the positive, although assigned a few lower scores. Participant three usually responded in the mid to upper range of 3-5, although never gave above a 5. Finally, Participant 4 gave the most compact responses, usually rating visualisations between a 4 and 5, and they never responded below a 3.

[11] This is an area that has been explored most keenly with regards to user's rating behaviour in recommender systems (e.g. [90]).

4.2 Stage One

The first stage of the study helped to build a profile of each participant's current work with and interest in email data, as well as an idea of overarching research interests and methodologies. The summary of the participants' responses to this stage of research is found in Appendix A, Tables 10 to supplement the findings presented here.

The responses represent four individuals with quite diverse interests and focal points for their works. There are two archival practitioners and two Arts & Humanities researchers. The archivists have different roles within their respective institutions with one, Participant 2, focused on the processing of large volumes of digital data within an archive and the other, Participant 3, on an array of legal, theoretical and practical factors relevant to archive management. The two researchers both have diverse research interests with one, Participant 4, centred on interdisciplinary theoretical work and the other, Participant 1, interested in fine art with a focus on film. In terms of envisioned work with email collections, the two archivists were both interested in quite practical aspects of email examination, in particular understanding the dimensions, content and risks associated with the emails as well as the provision of access to content for researchers. As noted in Sect. 2, these are all aspects that have often been the driving force behind email visualisation research. The archivists could therefore, in theory, be both users of the visualisations to aid in their archival workflow and suppliers of access to content through the inclusion of visualisations in, for example, a catalogue. This means that they provide a valuable perspective for the second and third stages of the study. The two researchers included in the study, conversely, are subject matter experts. Their focus is on extracting historical and/or theoretical analysis of an artist's life, work and network reflects the concerns of other researchers and potential users of email visualisations. This distribution of responses establishes a baseline from which to explore the participant's responses to different visualisations in Stage Two of the study, highlighting both the similarities and disparities between their approach to an archival collection. Of particular interest are the threads of similarity between the envisioned work that Participants 1, 2, and 4 might conduct with email collections. Participant 2 describes their work as 'very practical', but the processes involved in appraising, describing, reviewing and providing access to content align, at least in terms of the mechanics, quite closely with the work of the researchers— they each seek to comprehend the collection as a whole, the context within which it sits and the

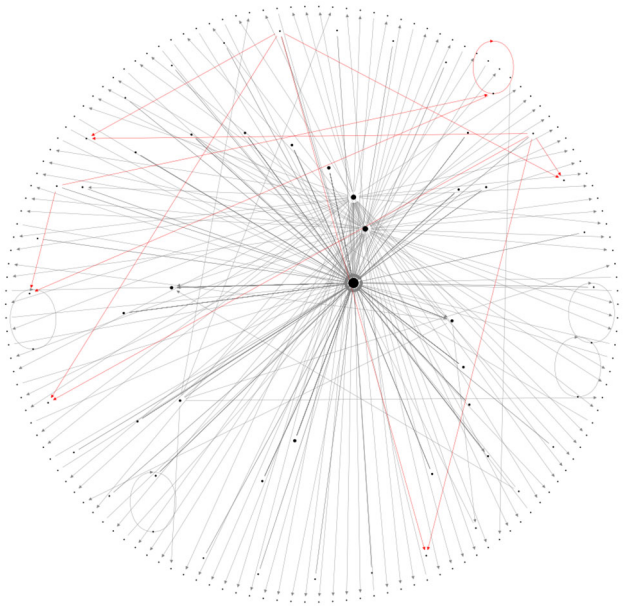


Fig. 5 A sample directed network graph representing PrivCon1, with email addresses redacted

relationships between names individuals/institutions. Whilst not stated as explicitly, this may also align with the 'collections development' described by Participant 3.

4.3 Stage Two

For this stage of the analysis, each type of visualisation is addressed individually with a table of collated responses from participants. Where relevant, direct quotes are taken from the surveys to provide additional context and a deeper analysis. Similarly, specific participants are noted if their circumstances impact on the results. Tables 11, 12, 13, 14, 15 provide the collated responses to this stage of the survey for reference.

4.3.1 Directed network graph

Responses regarding the directed network graphs (see Fig. 5 for an example) were, on the whole, dependent on the level of detail present in the graph (cf. Table 11). PrivCon0 and 2 were viewed most favourably regarding potential for research. Both were described as enabling participants' insights into the wider network, the individuals involved, power dynamics or spheres of influence. Participants also noted an indication of geographical location and professional affiliation from the email domain names. Particularly intriguing for the archivist participants was the potential for using this visualisation to link up with other resources within the archive to aid in or develop wider protocols for content or sensitivity review—an ongoing dilemma for large-scale digital collections (cf. [83, 91–94]). For the researchers, the interest was on the details that might be gleaned from the visualisation, allowing for high-level examination of the email collection and the artists' personal and professional milieu.

PrivCon 1 and 3, those less favourably reviewed by participants, were perceived as giving an idea of overarching patterns in the dataset, but were considered too abstract. For PrivCon1, one of the researchers found the form of the visualisation itself interesting but notes the difficulty of identifying key contacts from those which represented, for instance, food delivery services. This issue was exacerbated for PrivCon3 with all participants noting the lack of information and a difficulty interpreting the data without understanding the impact of the noise introduction.

Some other issues arising with this visualisation included the density of the data points making interpretation quite difficult. In addition, it was noted at all levels that the inclusion of additional data, referencing for example the number of emails each line represented, the date of communication, or the subject of the email would enhance the usability of the visualisation. Given that the visualisation is already information dense, it would be necessary to integrate any additional data using hover over or similarly interactive functions. Making the visualisation interactive may also mitigate the issues of density, supporting a more malleable approach to exploring the data, for instance, by reshaping the network around specific data points.

4.3.2 Mountain graph

Exploring mountain graphs as depicted in Fig. 6, the participants viewed PrivCon 0 and 2 equally well, but were less engaged with PrivCon1. PrivCon0 supported the identification of patterns of communication for all participants, with the archivists postulating additional usages in linking this visualisation to other archive items. Participant 1 found it particularly useful for identifying individual contacts within the dataset. It should be noted however, that Participant 3, an archivist, highlighted that the inclusion of email addresses gave rise to sensitivity concerns. The PrivCon2 visualisation was useful to the participants in a different way, helping to identify workflow and workload over time. The inclusion of categorisation allied some of the participants' concerns over understanding the content of the email collection, particularly for the archivists who saw scope for 'confirming aspects of context and... content'. Despite this, Participant 1 raised the concern that the categorisation applied to emails not to contacts, so could not account for instances where there was 'crossover... between professional and personal'. However, this is a data structure and markup issue, rather than an issue with the visualisation itself. Least well received of the mountain graphs was PrivCon1, for a familiar reason—the lack of detail. Comments such as 'needs names' and 'Content

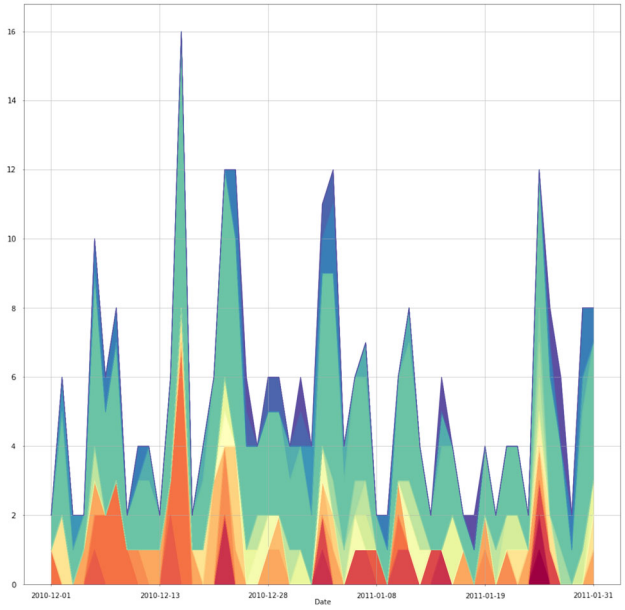


Fig. 6 A sample mountain graph representing PrivCon1, with email addresses redacted

- as always!' highlight the extent to which this is a priority for researchers in particular, although Participant 2 also concurred that the lack of information was a hindrance.

As a general comment on the visualisations, Participants 4 noted an important point about accessibility for this visualisation. As it relies on colour and, more particularly differentiation between layers of colour, there would be issues for any user who was colour blind. Similarly, it was noted that the visualisation was 'quite tiring on the eyes', something echoed by Participant 2. This suggests that prolonged usage may not be tenable, particularly for larger or more complex data sets. This is more of a design issue than one relating specifically to the usage and usefulness of the visualisation, but it is an important consideration nonetheless.

4.3.3 Scatter plots

Of the two scatter graphs (see Fig. 7 for an example), the PrivCon0 was judged to support research and practice for three participants, although Participant 2 did not agree as there was insufficient detail to support the archival work they do. Participant 3 conversely saw some potential of the high level perspective on the email collection 'to link to cataloguing data... and to aid in sensitivity review'. Despite this, they are again cautious of the privacy risks associated with the inclusion of email addresses. The researchers both agreed that the visualisation could support their work, highlighting patterns in the relationships as well as periods of high activity. Despite this, Participant 4 noted that the precise design of the visualisation was against expectation, in that they believed the date should be placed on the x-axis, something which added a barrier to comprehension. The more privacy-aware visualisation, PrivCon1, was again less well received by participants with the primary issue being the level of detail.

4.3.4 Bar charts

Out of all the visualisations, the bar charts as shown in Fig. 8 seemed to be the least supportive of the participants' work. Only the PrivCon0 visualisation was judged useful, although Participant 4 would require for the data to be contextualised to a specific event. Most participants again wanted additional detail, such as the subject or theme of specific emails. Participant 4, for instance notes that 'this visualisation can respond to 'who' questions... [b]ut it can't tell me 'what' the content of those conversations were'. PrivCon1 received the least favourable review out of the whole collection, with no participant seeing potential for it to support their research/practice. Attempts were made by Participants 2 and 4 to interpret the data, but these interpretations were quite broad. Participant 3 also unfavourably compares the efficacy with the mountain and scatter graphs. A possible reason for this is that, within this study, both the mountain and scatter graphs were keyed towards the relationships between people, whereas the bar chart was aimed towards their behaviour. The difference is quite nuanced, but, based on the participants' responses, important. The PrivCon2 bar chart had slightly more potential for the participants. Whilst Participants 1 and 2 could not find a use for the visualisation, Participants 3 and 4 thought that it may support their work. Participant 3, for instance, considered that an archivist may be able to use it to contextualise a collection, particularly if used in conjunction with other resources. Participant 4 was able to gain 'greater insight into the filmmaker's behaviours' as well as contextualising these in terms of the dates. Even Participant 1 considered that the visualisation could be used to compare 'who contacted him more than he them', but this is qualified by some doubt when they note that 'maybe I've misunderstood the information'. This is one of the few times that one of the participants highlighted an issue interpreting the visualisation. For PrivCon3, only Participant 4 considered that the visualisation might support their work, but this is qualified when they note that they felt 'increasingly uncertain... as to what data I am looking at', once more highlighting the need to additional clarification regarding the introduction of noise.

4.3.5 Word tree

The word tree visualisation as shown in Fig. 9 was quite divisive amongst participants and was the only type of visualisation to elicit responses across the spectrum with regards to usefulness. That being said, it also appears to have been

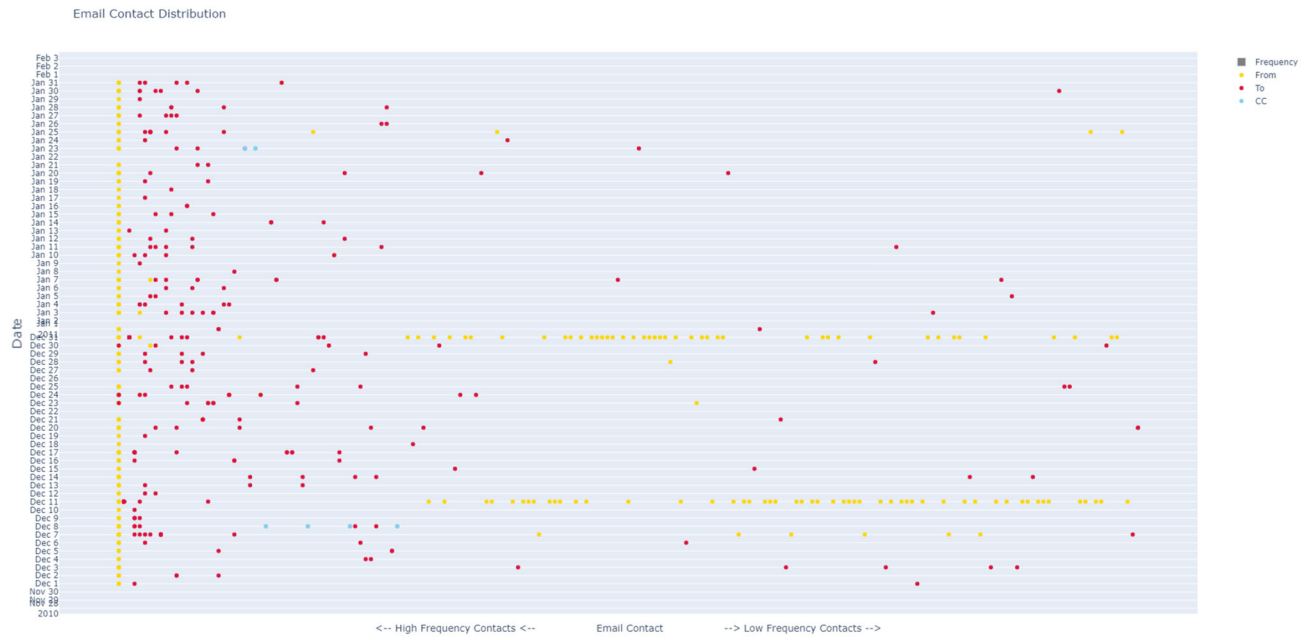
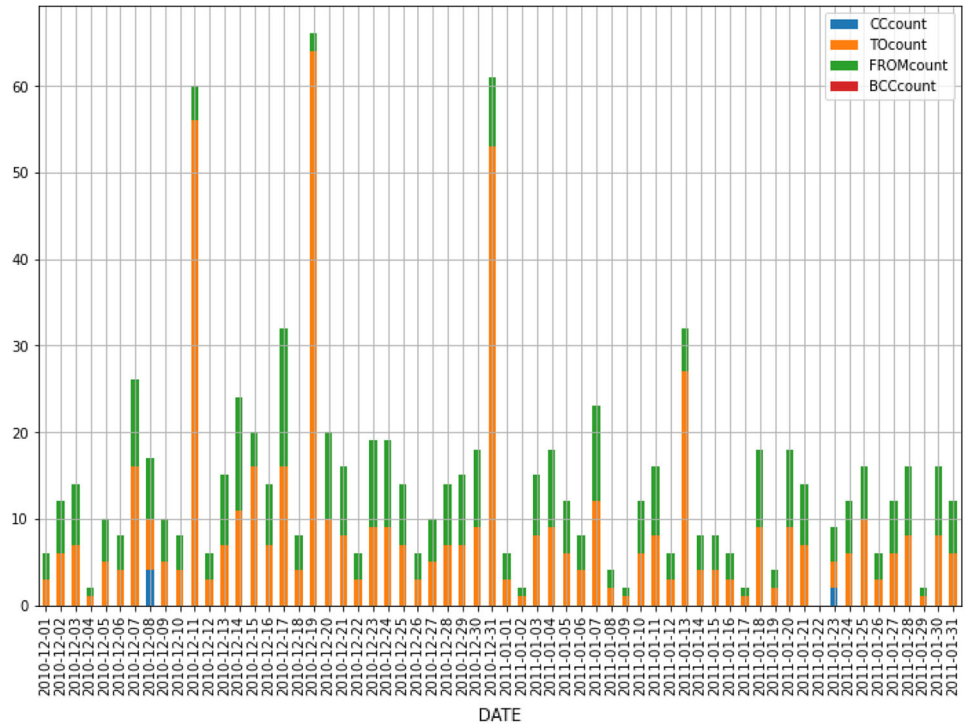


Fig. 7 A sample scatter plot representing PrivCon1, with email addresses redacted

Fig. 8 A sample bar chart representing PrivCon2, with email contacts aggregated



one of the most liked visualisations, allowing participants to delve, in reasonable detail, into the content of the email collection. PrivCon0, for instance, was useful to two participants, one was not certain and one did not feel that it would support their research/practice. Participant 1 consistently did not feel as though the visualisation supported their work, although this is tempered by the note that the visualisation would be improved 'if there were key words', suggesting that the central word was the issue. For Participant 4, PrivCon1 was the most useful for their research, highlighting the frequent structures and themes present within the collection. They also indicate that, whilst not all the information present in the visualisation is relevant to their research, it would likely have wide reaching applicability, encompass-

Fig. 9 A sample word tree representing PrivCon2, with identifiable data redacted

ing items of import for many kinds of researcher. At the higher PrivCon levels, they found that the lack of identifiable information was an impediment to their research. However, in direct contrast, they noted a level of caution about the extent of detail present in PrivCon0 and how that would position them with GDPR, particularly in relation to evidencing research in later work. Whilst not explicitly stated, this is also an issue with PrivCon1, where they indicate that they could 'hazard a guess' as to whether the contents originated from the filmmaker or another individual, suggesting potential for identity reconstruction, even with the redaction of sensitive content.

The archivists agreed at all levels that the visualisation could support their practice. At PrivCon0, Participant 2 notes that the visualisation would be 'great for identifying emails with sensitive content' and, at the other levels they highlight that whilst the pre-redacted content is less useful for the archivist, they could be a 'potential access tool for a user'. Participant 3 largely agrees, noting in particular that it was a 'clever way of looking deeply at content from across the dataset'. Despite this, they do indicate that the usefulness of the visualisation would be dependent on the search strategy, citing the idiom 'garbage in, garbage out'.

The distinction between usefulness at different levels of privacy awareness was the least pronounced for this set of visualisations, with most participants engaging equally at all levels. A potential reason for this is that, even with the sensitive content redacted, or the contextualising window removed, the visualisation still provides targeted insight into the contents of the email, rather than focusing on high-level metadata. It is clear from both the literature review and they participants' responses to the other visualisations that the content of emails is something that is considered key to both research and practice. This is perhaps due, at least in part, to familiar methodologies utilised throughout the Arts & Humanities and archival sciences, whereby researchers/practitioners will manually search through and engage with archival contents.

4.4 Stage Three

The final stage served to solidify and clarify the observations prevalent in Stage Two. The participants' responses to the questions are given in Tables 16, 17, 18, 19, 20 in Appendix A. For clarity, and to avoid repetition, the questions are presented as Q1, Q2 and Q3 rather that written out in full. For reference, these are:

Q1 —Is there anything you would like to add or change in relation to your initial assessment of this visualisation?
Q2 —How useful is this visualisation for your research or practice? 1 (not useful)–7 (very useful)
Q3 —Why have you given this rating?

4.4.1 Directed network graph

The responses relating to directed network graphs (depicted in Table 16) are reasonably coherent between the participants at each of the PrivCon levels. To address first the review of responses from Stage Two, Participants 2 and 3 had nothing to add to their thoughts after considering the collated opinions. Participant 1 supplemented their thoughts on PrivCon 1 and

Table 4 A summary of scores for usefulness given in response to the Stage Three of the study for Directed Network graphs

PrivCon Scale	P1	P2	P3	P4	Average	SD
0	5	5	4	5	4.75	0.433
1	3	2	2	4	2.75	0.829
2	5	5	4	6	5	0.707
3	1	2	2	4	2.25	1.090

3, noting that PrivCon 1 had a potential knowledge gap, prohibiting understanding and that the level of detail in PrivCon 3, particularly regarding names, rendered it unuseful. Participant 4 reflected on each of the visualisations, noting the high level of sensitive data in PrivCon 0 and suggesting that the redacted nature of PrivCon 1 helped them to 'rethink how data analysis might support their research' beyond their usual approaches. For PrivCon 2, they indicate that the visualisation might support 'useful conclusions about the filmmaker's creative activity' and for PrivCon 3 they express an interest in the processes involved in the creation of the visualisation, highlighting a potential knowledge gap that could be acting as a roadblock to understanding.

To turn to the participants' ratings, as summarised in Table 4, the PrivCon 2 visualisation was rated, on average, most highly, closely followed by Privcon 0. The scores from every participant for both of these visualisations were towards the positive end of the scale. Participant 3 gave the lowest score with a 4, noting that PrivCon 0 was 'quite hard to follow' and that, whilst they could envision a usage for PrivCon 2, it was only 'potentially some use'. PrivCon 1 received an average score of 2.75, so below the midpoint on the scale, although Participant 4 does give the visualisation a 4, noting that 'it does look useful' but highlighting that this usefulness is not immediately apparent and requires additional thoughts. Participant 1 describes the visualisation as a 'snapshot' suggesting something without much depth, something echoed by Participants 2 and 3 who indicate that the redacted information and the lack of contextual information reduce the extent to which the visualisation can be useful. Interestingly, the standard deviations indicate that participants were also more in agreement over the usefulness of PrivCon 0 and 2 (0.433 and 0.707 standard deviation, respectively) than they were over PrivCon levels 1 and 3 (0.829 and 1.090).

4.4.2 Mountain graph

The responses relating to mountain graphs (depicted in Table 17) are also coherent across the PrivCon levels and are also more positive than for the directed network graphs. With reference to any additional thoughts the participants had regarding these visualisations, only Participants 1 and 4 noted a shift in their thinking. For PrivCon 0, Participant 4 notes that the potential for diachronic analysis is helpful. Participant 1 agrees with the assessment given in Stage Two that this visualisation is problematic for those with colour blindness, but adds that it is easier to read than the network graph.

The scales of usefulness applied to this type of visualisation (Table 5) indicate that PrivCon0 is regarded as the least useful and PrivCon2 the most useful, although all levels are rated as above the midpoint of the scale. Participant 1's scores for PrivCon 0 and 2 are not explained, but confusingly they rate PrivCon 1 as a 6, noting that it would be 'very useful' with names or dates. This response points to a potential misunderstanding of PrivCon 0, which does contain names, but has been given a lower score. All PrivCon levels within this visualisation also contain dates, again suggesting a lack of clarity or a knowledge gap in the provision of the visualisation. Participant 2, across all privacy levels, highlights the visually appealing nature of the design, but again notes the potential for accessibility issues. Despite this, they give the visualisation a 5 across the board. They indicate that there is potential for these to display frequencies more clearly than, specifically, the network graphs and the word trees—a view that is perhaps indicative that visualisations focusing on patterns of relationships are of more use than those keyed towards networks or email content. Of specific PrivCon levels, Participant 2 notes the potential to mediate sensitivity when presenting the data to a wider audience. Participant 3 sees the possibility for integrating this type of visualisation to events/ trends in activity. PrivCon 1 is rated as less useful, with a score of a 4 compared to 5 for the other two levels, as Participant 3 notes that the visualisation works better with context. They also make an excellent point about PrivCon 2, that might be more widely applicable across all areas of category creation—that the categories could be subjective. This is something that could be mitigated through the inclusion of additional individuals in the tagging process. For PrivCon 1, they return once more to the idea that the visualisations which lack textual elements provoke or inspire them to think differently about digital communication. For PrivCon 2, the score given reflects again on the type of categorisation, much as it seemed to for Participant 3. Participant 4 also suggests that a combination of visualisations, particularly this and the word trees, might provide a more useful perspective, especially if they were cross-referenced with regards to content categorisation.

As well as being, on the whole, more highly rated than the directed network graphs, the mountain graph visualisations also elicited a more cohesive response from participants, with the standard deviation sitting between 0.433 and 0.707.

Table 5 A summary of scores for usefulness given in response to the Stage Three of the study for mountain graphs

PrivCon Level	P1	P2	P3	P4	Average	SD
0	4	5	5	5	4.75	0.433
1	6	5	4	5	5	0.707
2	6	5	5	5	5.25	0.433

Table 6 A summary of scores for usefulness given in response to the Stage Three of the study for scatter plots

PrivCon Level	P1	P2	P3	P4	Average	SD
0	6	4	4	6	5	1.000
1	3	4	3	3	3.25	0.433

4.4.3 Scatter plot

The responses relating to scatter plots (Table 18) are harder to compare with the other types of visualisations, as there are only two PrivCon levels represented. Only Participant 1 augmented their original assessment upon reviewing the collated responses, noting of PrivCon 0 that the timelines supported the mapping of activity and individuals. They also suggested that this feature could be cohesive with the paper side of the hybrid archive.

Looking at the usefulness assigned to each visualisation (cf. Table 6), the scatter plots more closely reflect the responses given in response to the directed network graphs, whereby PrivCon 0 is rated as much more useful that PrivCon 1, receiving an average score of 5 and 3.25, respectively. That being said, PrivCon 0 was quite divisive between the participants, with the researchers both rating the visualisation with a 6 and the archivists a 4, leading to an overall standard deviation of 1.000. The reasons the researchers give for mapping this visualisation so highly, the highest score given to any visualisation in fact given that none were rated with a 7, include its usefulness for understanding patterns of behaviour and honing in on 'key frequency/volumes of contacts'. Sensitivity is again an issue for Participants 4, but they suggest that it might work best as an 'internal team tool'. This perspective is highlighted by their response to PrivCon 2, where they indicate that the pattern alone is unhelpful, except in the case where it was 'tied to a specific output e.g. a film or a specific person or contact'. If this more focused approach could be taken, then a redacted version of the visualisation could be published alongside the results, therefore reducing the risk posed by the personal data included in PrivCon 0. Participant 2 explains their score of a 4 by noting that scatter plots can be difficult to interpret due to the quantity of information it contains. They do suggest, however, that this may support a more detailed analysis rather than, for instance, the overview provided by the mountain graph. Intriguingly this comparison is interesting given that both forms of visualisations are intended to display patterns to be found in the relationships evidenced by the archive. This is perhaps something Participant 2 was able to engage with, suggesting a certain cohesion between the visualisation types. Participant 3 admits that they are uncertain how they might use the visualisation, although the seem to see potential within it, perhaps for users or alternative job roles, given the score that they assign.

PrivCon 1 elicited a lower, if less diverse response from participants, with Participants 1, 3, and 4 assigning a 3 and Participant 2 a 4. As with previous visualisations, Participant 1 indicates that the redaction of names is problematic for their interpretation and use of the visualisation. This is something agreed by Participant 3, who notes that potentially this drawback could be mitigated by comparison with additional datasets. Participant two once again ascribes the same score to each visualisation within this set, noting the redaction as a useful possibility rather than a hindrance. When contrasted with their perspective on the directed network graphs, it is possible that the focus of the visualisation on patterns in relationships is indeed the difference in the levels of usefulness that they perceive.

4.4.4 Bar chart

The responses relating to bar charts (depicted in Table 19) are perhaps the most disparate of any within the dataset, eliciting both the highest and lowest scores for usefulness, even in relation to the same PrivCon level. All participants other than Participant 2 augmented their original thoughts upon reviewing. Participant 1, for instance, notes that PrivCon 2 allowed for conception of contact initiated and reciprocated. Participant 3 agreed and liked that the PrivCon 0 visualisation could supplement the catalogue, and that PrivCon 2 was primarily useful in relation to specific research questions. Participant 1 once more notes that the introduction of noise acts as a barrier to using PrivCon 3, something echoed by Participant 4 in an inability to identify the types of data presented in PrivCon 3 compared to PrivCon 2.

As can be seen in Table 7, this series of visualisations follows the pattern seen in relation to directed network graphs and, to a lesser extent, the scatter plots. The PrivCon level rated as the most useful was PrivCon 0, followed by PrivCon 2 and then, on an even level, PrivCon 1 and 3. PrivCon 0 and 2 are also the most cohesive in the participants' opinions, although the deviations are still quite high with scores of 0.029 and 1.118, respectively. The PrivCon 0 visualisation receives the highest average score, an accolade shared with the PrivCon 2 Mountain graph and the PrivCon 0 Word Tree. This contrasts with initial considerations identified in Stage Two of the study, in which PrivCon 2 received two 'maybe's

Table 7 A summary of scores for usefulness given in response to the Stage Three of the study for bar charts

PrivCon level	P1	P2	P3	P4	Average	SD
0	6	6	5	4	5.25	0.829
1	1	6	3	4	3.5	1.803
2	4	6	3	5	4.5	1.118
3	1	6	3	4	3.5	1.803

and two 'no's compared to PrivCon 0 which received 3 'yes's and one 'maybe'. The reasoning offered by participants for their scoring of PrivCon 0 includes that the visualisation allows for the user to understand the frequency with which filmmaker contacted people and that there are other practical applications, especially if cross-referenced with other visualisations. Participant 2 highlighted that the form of the visualisation, a bar chart, is fairly familiar and so would have a reduced learning curve making it useful for the majority of users, something that this participant echoes across the spectrum of this type of visualisation, no matter the PrivCon level applied. They add, again, that the option for redaction is useful and that the aggregated visualisation presented for PrivCon 2 'can provide a lot of detail'. Other participants were less optimistic about the PrivCon 2 visualisation, noting that it 'gives some information' but that 'other visualisations probably do this job better'.

Participants supplied identical responses for PrivCon 1 and 3, with the lowest score attributed by Participant 1 and the highest by Participant 2. The distinction between these two scores is quite extreme and is in part responsible for the standard deviations of 1.803. Both Participants 2 and 4 judged these visualisation to be equally as useful as that supplied for PrivCon 0, with Participant 2 again citing the familiarity of the design and Participant 4 pinning this level of usefulness on the potential to, at some point, identify the highest frequency contacts in PrivCon 1, and for it to be possible to identify 'any time critical period in the filmmaker's life' for PrivCon 3. Participant 3, conversely, judged PrivCon 1 and 3 to be of equal usefulness to PrivCon 2, indicating that PrivCon 1 lacked contextual information and acknowledging that they could not determine how PrivCon 3 would be helpful. Participant 1 was most critical of these two PrivCon levels, indicating that PrivCon 1 contained no useful information, whereas PrivCon 3 was obscured by the noise, making it hard to read—both of these sentiments echo this participant's feelings about other visualisations of similar PrivCon levels.

4.4.5 Word tree

The responses relating to word trees (depicted in Table 20) follow a pattern not seen in any of the other visualisation types, where the usefulness score decreases as the PrivCon level increases. The participants also offered more and lengthier considerations on the visualisations, based on review of the collated responses from Stage Two. For PrivCon0, they reiterate the importance of the search term and note that connecting it to a 'more advanced search interface' and other datasets would be beneficial. Participant 1 goes even further indicating that they would like to see more of the email contents for each search term. Participant 4 focuses in on the content of the visualisation, noting the potential for research into the emotion of the email collection, one of many related and valuable branches of research within the humanities. For PrivCon 1, Participant 1 again notes that redaction reduces the usefulness of the visualisation, something that is advanced by Participant 4, who indicates that, in particular, it could cause issues in tracing threads of conversations. Opinions on PrivCon 2 are slightly divergent with regards to the collated material. Participant 3 notes that the 'absence of the reading panel makes this much less usable,' but Participant 4 notes that it is 'slightly easier to read' making it more accessible as an interface.

As can be seen in Table 8, the responses to this visualisation are quite cohesive, with all the highest standard deviation to be found for PrivCon 1. As noted in relation to the bar charts (Sect. 4.4.4), the PrivCon 0 word tree has one of the highest average ratings of all the visualisations. Participants highlight that the reasons for this include increasing the discoverability of the information, particularly with the reading panel for quick access. They also highlight that it limits the need for manual search of emails, although there is a caveat to both of these depending on the search term utilised. Participant 2 believed that this visualisation would be 'easily understood by the majority of out remote users'. Participant 4, once again, returns to the idea of personal and sensitive data, considering that this visualisation might be 'too revealing'. The feedback from participants relating to PrivCon 1 is similar in nature, again highlighting the possibility for more efficient navigation and discoverability of content. Participant 2 once more notes the usefulness of being able to redact content and Participant 4 considers the role this type of visualisation could play in relation to 'questions of attribution', but once more highlights that this would need to be in 'in concert with other tools'. Interestingly Participant 1 rates the PrivCon 2 visualisation as being equally useful at PrivCon 0, noting that the usefulness depends on the search term. Each other participant rates this visualisation as the lowest of this set, with Participant 2 noting that the lack of a reading pane might be 'frustrating', something that is reflected by Participant 3, who notes that the removal of the panel doesn't bring any advantage and is, in fact, illogical when there is the 'opportunity to link a finding aid to the content'. Participant 4 reflects that the visualisation would be useful, but may require them to adapt 'the nature of that research in engaging with the visualisation'.

Table 8 A summary of scores for usefulness given in response to the Stage Three of the study for Word Trees

PrivCon Level	P1	P2	P3	P4	Average	Standard Deviation
0	5	5	5	6	5.25	0.433
1	3	5	5	5	4.5	0.866
2	5	4	3	4	4	0.707

5 Discussions and implications

5.1 Discussions

On the whole, participants were able to engage creatively and productively with the majority of the visualisations. The responses given by both researchers and archivists indicate that they participants were able to envision how visualisations might support their work should they engage with email collections. At times, this usage was concomitant with existing practice, for instance, supporting existing activities or research questions. At others, the visualisations prompted participants to consider new perspectives on how they might engage with the data. Participant 4 regularly reflected on new areas of thought prompted by surveying the email collections through the use of visualisations, or, in some cases, by the creation of the visualisation itself. Similarly, Participants 2 and 3 noted several possibilities for integrating visualisations into the archival workflow, supplementing the catalogue, or providing a point of access for users. These findings support established thought[12] that visualisations support holistic, exploratory behaviour of data, encouraging a user to engage with existing modes of thought but also facilitating them to gain new insights and therefore, potentially, prompt new approaches and questions in individual subject areas, cross-disciplinary research or professional practice.

In terms of the impact of different levels of privacy awareness on usefulness, the findings demonstrated that although each of the PrivCon levels achieved at least one score of 6, the distribution of the other scores varied quite dramatically and yielded unexpected results. It was postulated in [11] that 'when considering email data from the perspective of humanities researchers, whose standard methodologies involve the close and usually manual examination of data, the scale of privacy may well be considered inversely related to the degree of useful access'. However, in this empirical approach to investigating the issue, it was demonstrated that the situation is more nuanced than that with the usefulness dependent on the underlying focus of the data and associated analysis as much as the restrictions introduced by the privacy management strategy.

As revealed in [11], PrivCon 1—particularly anonymisation, pseudonymisation and redaction—represents the most popular privacy management strategy employed by those conducting research into email collections, through the use of visualisations. Most participants, however, viewed this redaction as removing key information (e.g. names) that was essential to their work. The sense, for most, was that simply viewing the overarching pattern made by individual data points was insufficient for detailed analysis within an arts & humanities and archival workflow context. To a degree, this might be minimised by the use of a different techniques, such as pseudonymisation, whereby participants would still be able to follow the threads of specific individuals even if that individual was not explicitly named. This option, however, is more risky in terms of the potential for re-identification (cf. [84–87]). Conversely, Participant 2 acknowledged that the opportunity to redact content was beneficial to allow the wider release of email data. In line with this, Participant 4 revealed a level of anxiety regarding the amount of information available at the lower PrivCon levels, especially as it pertained to disseminating their research. This, therefore, indicated that the higher PrivCon levels might have specific purposes for the public facing side of research or practice, after the data have been surveyed and analysed without the use of a filter. In fact, this follows the pattern found in many of the studies identified as associated with PrivCon 0 datasets in [11]. These papers would facilitate open access to the data for researchers involved (often utilising the participants' own email collections) and then anonymise, pseudonymise and/or redact content to allow for publication of examples. In terms of active research or practice, however, not only do these approaches provide a lower level of privacy for the data subjects, but they also provide little usability for follow on work.

The results for PrivCon 2 were most strikingly contrary to the expectation of the relationship between privacy awareness and usefulness. Whilst, on the whole, not viewed as being quite as useful as PrivCon 0, visualisations in this category are well regarded by the participants. In one notable instance, the directed network graphs, the PrivCon 2, received a slightly higher score than PrivCon 0. Based on the participants' responses and proposed usages for this privacy awareness level, it suggests that this higher level of protection concurrently offers a greater range of opportunities for researchers and practitioners to engage with email collections. By grouping data points so that the individual is hidden in a crowd, this type of visualisation offers a summary or intermediary form of analysis that can inform and inspire

[12] As discussed briefly in introduction and more full in, for example [20–24, 26, 27] and [25].

the user in their work. Such holistic perspectives are increasingly proving valuable within the humanities with the advent of data-driven studies such as those associated with, to name a few areas, distant reading (cf. [95–97]), digital humanities (cf. [98–101]) or machine learning and AI (cf. [102–104]). In addition, an email collection results in a large, potentially untenable number of data points. The dataset utilised for this study, for example, was a small sample of the complete email collection (approximately 5.4%) and this, in turn, was a relatively small email collection compared to those that exist in more recent archival datasets ([3, 105]). Even at the scale presented in this study, participants raised concerns about the level of detail present in some of the visualisations, the network graphs in particular, suggesting that they might become unsustainable if expanded to larger experiments. The introduction of interactive elements (e.g. the ability to zoom, re-centre, include hover over information) is one solution to mitigate these issues, but these demand a higher level of technical skill on the part of the creator of the visualisation, as well as greater hardware and software requirements. The amalgamated nature of PrivCon 2 style visualisations is another possibility, and one with both a high level of usability and privacy awareness.

The final PrivCon level explored in this paper, PrivCon 3, was regularly judged to be the least useful to the participants' work. The reasoning behind this appears to be, in the first instance, one of a knowledge gap. There were a number of instances throughout the study where participants were uncertain about engaging with the visualisations. In fact, the majority of issues arose from the level of detail and context (or lack there of) for the visualisations. The only issue where the participants consistently exhibited anxiety about their ability to comprehend the visualisation, both at Stage Two and Stage Three, was for PrivCon 3. Here participants expressed the need to more completely understand the processes underlying the generation of noise and how this might impact upon their analysis of the data.

Within these overarching patterns, there were some possible influencing factors or points requiring further investigation. There was one instance where PrivCon 1 was rated more highly than PrivCon 2 and that was in relation to the Word-Trees. This disparity from the overarching pattern is perhaps best accounted for by the removal of the reading panel for PrivCon 2. Similarly, there was evidence of anomalous results for the Mountain Graphs. Each PrivCon level for this set received very similar usefulness results from the participants. The distinction came from Participant 1 who gave a rating of 4 for PrivCon 0 and a 6 for levels 1 and 2. In principle, this is a truly intriguing result; however, when exploring their reasoning behind the score, there appears to have been some confusion given that under PrivCon 1 it is noted that 'if it had the names/dates it would be very useful'. Each of these graphs does have the date included and the PrivCon 0 graph also has the names, but was given a lower usefulness score. Unfortunately, there is no reason given for the PrivCon 0 score. Additional investigation would be required to facilitate a more concrete analysis.

5.2 Implications

This work has implications primarily for the archival sector, but also for any researcher who might engage with email research data. Archives are increasingly faced with a daunting challenge of managing the ingest, review, and management of data from large scale digital([16, 83]). Of especial concern is how archives might process increasing archival digital collections and conduct sensitivity reviews in a timescale to allow researchers timely access to the data contained within. The visualisation presented in this paper offers a viable, but adaptive level of privacy protection to the individuals named within an email collection in a rapid, relatively resource-light manner. More than this, these visualisations enable useful access on the part of researchers, supporting them to engage with a collection that, in its present condition, would otherwise remain closed to their work. More specifically, this work has revealed that commonly utilised methods for protecting privacy whilst facilitating access—notably those that fall under PrivCon 1—may not actually be the most advantageous methods to utilise, and instead research suggests that researchers and archivists both favour meaningfully amalgamated perspectives on the data, such as those contained within PrivCon 2.

6 Conclusion

The research presented in this paper represents a strong step towards the integration of privacy-aware visualisations with an archival email collection to facilitate access to content that might otherwise need to be closed due to privacy considerations. The experiment was conducted with a small specialised group of researchers and archivists, each of whom had privileged access to a filmmaker's email collection. This allowed for the examination of email visualisations from across the spectrum of privacy levels identified in [11], without the need to modify or review the collection in a way which might interfere with the results. The participants offered valuable insights into how the visualisations might support their work, helping to identify the relationship between privacy-aware strategies and their perceived usefulness in relation to their research and/or practice. Their comments and justifications helped to identify email data analysis, visualisation and communication features of note that can facilitate their work.

There are several valuable avenues available for advancing the work presented in this paper. The most immediate would be to replicate the study with a larger participant pool

that encompasses a wider range of research and professional expertise. There are certain impediments to this with regards to disclosing potentially personal and sensitive data to a wider user group; however, with proper ethical considerations and a carefully selected dataset, such an experiment could provide essential validation for the results of this study. Incorporating additional datasets would also allow elucidation on the impact of scale that is raised in Sect. 5.1. Beyond this, there is an established consideration that must be investigated with regards to ratings. It can be difficult to accurately compare ratings between individuals (cf. [90, 106]), particularly in situations where the data may be sparse (cf. [107]. Increasing the dataset and applying more rigorous methods of averaging out the responses, for instance similarity metrics such as Pearson's Correlation or Jaccard similarity (cf. [108–110]), would be step forward in understanding this effects. In addition, it would allow for a greater diversity of responses, incorporating not only additional research and practitioner interests, but also participant backgrounds to allow for a more democratic review of the visualisations.

Similarly, the breadth and depth of each PrivCon level and each branch of research have, by necessity, been addressed on a surface level in this study to allow for an overarching perspective on the situation as a whole. The PrivCon levels identified in [11] encompass many more privacy management strategies that could be reflected in an extended study. Equally, the branches of research interests could be deconstructed to explore further facets and a variety of other visualisation designs. Additional investigations, perhaps focusing on the nuances of each area of research or each PrivCon level, would allow for a more nuanced analysis of usefulness. Equally, this more focused investigation would permit opportunities for exploring more detailed privacy management strategies, such as ascertaining the impact of choosing between direct redaction, anonymisation, and pseudonymisation or examining the more advanced techniques involved in PrivCon 3 (e.g. differential privacy).

Another area of interest would be to touch more deeply on the potential of the visualisations to facilitate and even provoke new areas of investigation, or new research questions within established fields. A recent AHRC funding opportunity entitled 'Embed digital skills in arts and humanities research'[13] is one indicator amongst many that humanities research is increasingly leveraging the flexibility of the digital format to encourage new areas of thinking. Visualisations sit well within this remit as an approachable and flexible interface for data driven research. The archivist participants in our study were, in fact, able to see the potential in these to act as a mediator for their collections, enhancing discoverability, and accessibility. This is something that has also been touched upon in, for example, [111] and [112] as well as the increasing presence of heritage datasets such as Digital Bodleian[14] and the National Library of Scotland's Data Foundry.[15] The research contributed to the wider research community's understanding of privacy and privacy management engaging more closely with trained archivists to assess how the levels of privacy managed by visualisations compare to that of a thorough sensitivity review. This is likely to augment the viability of incorporating the visualisations into an interface for an email collection, something that could be of tangible benefit to both archive practitioners and users alike.

Acknowledgements This work was supported by the Arts and Humanities Research Council (AH/R007012/1). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

Author Contributions Equal contribution from all authors.

Funding Arts and Humanities Research Council funding [AH/R00 7012/1].

Data availability Original email data are held at Special Collections, University of Reading (currently pending review for sensitive information). The survey data are included in the article as Appendix.

Code Availability Not integral to the study.

Declarations

Conflict of interest The authors declare no conflict of interests.

Ethics approval Approved by the University of Glasgow Ethics Committee—application 100200091.

Consent to participate Signed by each participant.

Consent for publication Obtained from each participant.

[13] https://www.ukri.org/opportunity/embed-digital-skills-in-arts-and-humanities-research/ (accessed 28/06/2022).

[14] https://digital.bodleian.ox.ac.uk/.

[15] https://data.nls.uk/.

Appendix

[The following pages present the participant responses for Stages 1-3 of the study collated into twelve tables for reference and re-use.]

Table 9 A table noting the order in which visualisations were presented to participants: e.g. 5.4 in the second row of the column for Participant 3 means that directed network graphs were shown in the fifth group of the survey questions and PrivCon 1 graph of that group was shown as the fourth visualisation

Visualisation	Participant 1	Participant 2	Participant 3	Participant 4
Directed Network PrivCon0	1.1	1.3	5.1	2.1
Directed Network PrivCon1	1.2	1.1	5.4	2.3
Directed Network PrivCon2	1.4	1.2	5.2	2.2
Directed Network PrivCon3	1.3	1.4	5.3	2.4
Mountain PrivCon0	5.2	3.3	1.3	3.3
Mountain PrivCon1	5.3	3.2	1.1	3.1
Mountain PrivCon2	5.1	3.1	1.2	3.2
Scatter Graph PrivCon0	4.1	5.2	2.2	4.1
Scatter Graph PrivCon1	4.2	5.1	2.1	4.2
Bar Chart PrivCon0	3.4	2.4	3.4	1.1
Bar Chart PrivCon1	3.1	2.3	3.1	1.2
Bar Chart PrivCon2	3.2	2.1	3.2	1.3
Bar Chart PrivCon3	3.3	2.2	3.3	1.4
Word Tree PrivCon0	2.1	4.2	4.3	5.3
Word Tree PrivCon1	2.3	4.3	4.1	5.1
Word Tree PrivCon2	2.2	4.1	4.2	5.2

Table 10 A table of responses for Stage One of the study: collecting participant background information, such as research discipline, interests, and common methodologies used.

	Primary field of research/practice	Research/practice interests	Research/practice methodologies	Envisioned research/practice with email collections
Participant 1	Fine Art	Artists video	art methodologies (thinking through making), theoretical and empirical	historical analysis of working methods of artist or contextual understanding of their milieu
Participant 2	Archivist looking at cataloguing and processing large collections of personal digital archives	Interested in how archivists can apply our professional practice to large volumes of digital data which may be deposited in our collections. How do we review this material, provide catalogue descriptions of the material. how do we appraise, organise and conduct sensitivity reviews on the material	N/A	Very practical work - appraise, describe, review, provide access
Participant 3	Archives	Broad professional and research interests in copyright and compliance, cross-domain and interdisciplinary working, photograph collections, performing arts and events data standards (and intangible cultural heritage more broadly), and preservation management	Practical archive (and cross-domain heritage) work	Leadership of collections development, management and access
Participant 4	An interdisciplinary scholar with broadly defined fields of research as being centred on studies of visual culture, creative practice, cultural studies, and critical theory/philosophy	Critical Disability Studies, Gender and Sexuality, Embodiment, theories of affect and the emotions, Cultural Institutions (methods, practices, policies, ideologies), Art practice and creativity, Theory and Philosophy, more recently accessibility, equalities and digital practices	Work conducted between close attentive studies of artworks and their contexts, and broader theoretical models that might support or explain the emerging trends in these studies	An examination of confluences of power/ agency/ influence and how these might (or might not) reflect models of creative practice and the dissemination of art works. Particularly interested in the relationships between the filmmaker and key cultural agents or institutions

Table 11 A table of collated responses for the perceived usefulness of directed network graphs

Directed network graph				
PrivCon level	What kinds of information can you gather from this visualisation?	Does this type of visualisation support your approaches to research?	In what ways might visualisations like this help you to address your key questions/themes and/or envisioned outcomes?	In what ways could the visualisation be lacking in helping you address your key questions/themes and/or envisioned outcomes?
PrivCon0	Pattern of communication in terms of network, names, geographical locations, and whether the email is a business or a person. Supports integration of personal/historical knowledge to understand email usage. Spheres of influence	3x yes 1x maybe	Description/Catalogue work Confirm context Sensitivity review Identify key contacts and milieu Interrelation of contacts	Difficult to follow Incorporation of content Links to project/timelines Nature of contacts Construction of vis and underlying infrastructure
PrivCon1	Frequency of emails from different sources. Most emails were between two contacts rather than a range of people. Constellation of influences/connection. Identification of key nodes of contact and mutual connections at the periphery	2x maybe 2x no	As a comparator. Social milieu Reflections on cultural influence and agency	Hard to understand. Numbers of emails. Names Subject matter.
PrivCon2	Pattern of communication and network. Deeper analysis without sensitivity concerns. Geographical location of contacts and whether it is business/professional/commercial correspondence or personal. Focus on institutions and organisations	1x yes 3x maybe	Identify gaps in the record. Confirming context and content. Geographical locations Type of contact Confirm or challenge other info from the archive.	Useful in a company archive, less in personal. Numbers of emails. Timeline information.
PrivCon3	Shows that someone sent and received a number of emails from different sources. Otherwise not much without additional information, especially as this is a combination of redaction and noise	1 x maybe 3 x no	As a comparator Key questions/themes around cultural influence and agency.	Hard to understand Lacking in all ways Names Clarity on underlying processes used to protect privacy.

Table 12 A table of collated responses for the perceived usefulness of mountain graphs

Mountain graph				
PrivCon level	What kinds of information can you gather from this visualisation?	Does this type of visualisation support your approaches to research?	In what ways might visualisations like this help you to address your key questions/themes and/or envisioned outcomes?	In what ways could the visualisation be lacking in helping you address your key questions/themes and/or envisioned outcomes?
PrivCon0	Identifies wide range of conversants and amount of correspondence - pattern of communication over time. Analysis of network and workflow/workload. E.g. a person who was helping him professionally at that time through filming would shows the amount of input. Detail enables easier links to other sources	2x yes 2x maybe	Practical archives administration (identify gaps, confirm context/ content for catalogue description, link to catalogue data, aid sensitivity review.) Who was supporting him professionally and who he was conversing with. Collaborate with known events in his life.	Some lines v.close together - hard to distinguish/difficult on the eye esp. for colourblind individuals Risky due to sensitive content Subject of correspondence
PrivCon1	Understanding of patterns of communication over time & analysis of network, workflow/load ad intersections of these. Frequency of contacts over time	2x maybe 2x no	Practical use in archives administration (identify gaps in record & confirm context/content for catalogue description. Link to catalogue data and aid in sensitivity review.) Understand how groupings of 'conversation' appear and disappear like hearing the sounds of a crowd/group at a distance, without being able to make out individual voices or words. Might support a detailed biographical piece, or seeking a deep-level understanding of working methods / networks	More detail (e.g. geographical location, subject matter, names)
PrivCon2	Understanding of patterns of communication over time & analysis of network, workflow/load and intersections of these. Ratio of type of email. Early-stage qualitative coding/cohorts for email analysis. High volume of personal email	2x yes 2x maybe	Practical use in archives administration (identify gaps in record & confirm context/content for catalogue description. support summary description about type of email.) Pattern of working and social behaviour. Question the role of qualitative researchers in analysing large textual datasets.	More detail (e.g. geographical location, subject matter, names) Coding according to types of content more specific to the content generated by the filmmaker. Crossover between professional and personal (some are both).

Table 13 A table of collated responses for the perceived usefulness of scatter plots

Scatter Plots				
PrivCon level	What kinds of information can you gather from this visualisation?	Does this type of visualisation support your approaches to research?	In what ways might visualisations like this help you to address your key questions/themes and/or envisioned outcomes?	In what ways could the visualisation be lacking in helping you address your key questions/themes and/or envisioned outcomes?
PrivCon0	Understanding of patterns of communication over time & analysis of network, workflow/load and intersections of these. Detail allows links to other sources. Who he was in contact with and quantity of contact. Details of unusual patterns of communication	3x yes 1x no	Practical use in archives administration (identify gaps & Confirm context/content for catalogue description. support summary descriptionabout type of email) Prominent data patterns/trajectories Useful, especially if tailored to questions like disability, sexuality, creative process - determined by identification of correspondents and their roll in his life	Risky due to sensitive data More detail Unusual axes
PrivCon1	Understanding of patterns of communication over time & analysis of network, workflow/load and intersections of these. Frequency of individual contact over time	2x maybe 2x no	Practical use in archives administration (identify gaps in record & confirm context/content for catalogue description support summary description about type of email.) Would need to be used in conjunction with wider information in the non-digital archive. Might support detailed biographical piece of deep-level understanding of working methods/networks	More detail (e.g. time of day, emails addresses)

Table 14 A table of collated responses for the perceived usefulness of bar charts

Bar Chart				
PrivCon level	What kinds of information can you gather from this visualisation?	Does this type of visualisation support your approaches to research?	In what ways might visualisations like this help you to address your key questions/themes and/or envisioned outcomes?	In what ways could the visualisation be lacking in helping you address your key questions/themes and/or envisioned outcomes?
PrivCon0	Patterns of communication over time & analysis of network, workflow/load and intersections of these. Frequency of individual contact over time, geographical location, type of contact (personal, institutional, organisational) estimate of volume. Detail enables deeper analysis and links to other sources	3x yes 1x maybe	Practical archives administration (find gaps in record, confirm context/content/extent for catalogue description, development of conversations.) his milieu Triangulate between this new form of information (visualisations), and existing processes, to find useful common ground. Adapt research questions to new methods.	Risky due to sensitive information Subject of email exchange/ detail of email content Detail on his work and working methodology.
PrivCon1	One contact received a large number of emails, three with quite high frequency	4x no	Helpful to examined processes of how the visualisation was built for secondary research (interests in research collaboration, leadership in archives and collections)	Lack of detail (e.g. qualitative data points)
PrivCon2	Understanding of patterns of communication over time. Daily to/from ratio. No CC's/BCCs suggesting he primarily messaged on a one-to-one basis rather than a group. Some periods where no messages were sent	2x maybe 2x no	Inform archivists looking in detail at context. General idea of email use. Focus on specific known period of the filmmaker's life, to understand how and when he used email as a digital tool. Tendency in arts and humanities to assume separation between creative and life process, but this contradicts that	Less useful than mountain and scatter examples, unless to/ from/cc is of particular interest. Lacking in detail (content, sender, nature of contact.
PrivCon3	Information on amount of emails the filmmaker was sending/receiving in a day	1x maybe 3x no	It doesn't/unsure	Clarity of what 'noise' means. Increased privacy makes understanding of data increasingly uncertain.

Table 15 A table of collated responses for the perceived usefulness of word trees

Word Tree				
PrivCon level	What kinds of information can you gather from this visualisation?	Does this type of visualisation support your approaches to research?	In what ways might visualisations like this help you to address your key questions/themes and/or envisioned outcomes?	In what ways could the visualisation be lacking in helping you address your key questions/themes and/or envisioned outcomes?
PrivCon0	Contains phrases used and specific content/contexts (e.g. weather) that are directly linked to contacts. It also contains a large quantity of sensitive and personal data and is a clever way to look deeply at content from across the dataset.	2x yes 1x maybe 1x no	Identifying emails with sensitive content, creating the catalogue and viewing content in context. Analysed in aggregate to determine patterns across different individuals. Pinpointing emails linked to specific words (film titles, festivals or people) to aid access. Types of language relating to themes (e.g. health). Some sense of who is writing to whom.	Addition of email header + footer info to support identification of individuals and dates/times Refined search strategy Sensitivity concerns Focus on keywords Differentiation between incoming and outgoing
PrivCon1	This provides context across the dataset for phrases and links directly to a contact. Frequency of terms allows a sense of communication style.	3x yes 1x no	Identifying sensitive content, creating catalogue Viewing content in context. Gain sense of written communication styles. Substitutes reading a lot of text.	Addition of email header/footer info. Refined search strategy Pre-redacted information less useful for cataloguer. Additional context. Detail on the nature of the email thread Guidance on how to cite
PrivCon2	Detail about the type of correspondence, phrases used (e.g. weather) and amount of correspondence. Details about the personal situation of the filmmaker. A clever way to look deeply at content from across the dataset.	2x yes 1x maybe 1x no	Identifying sensitive content, creating catalogue Viewing content in context. Types of language relating to themes/ keywords. Some sense of who is writing to whom.	Addition of email header/footer info. Refined search strategy Sensitivity concerns Harder to understand context without reading panel

Table 16 A table of responses to Stage Three of the study for Directed Network Graphs: each participant's understanding of the visualisation (columns 3-6) and usefulness ranking (last column)

Directed Network Graph						
PrivCon	Question	Participant 1	Participant 2	Participant 3	Participant 4	Average
0	Q1		No	no change	I'm aware how much sensitive data is displayed in this visualisation, so I'd likely encounter data protection/sensitivity issues.	
	Q2	5	5	4	5	4.75
	Q3	It is useful to see milieu through an email snapshot.	It is very 'busy' and could be very complicated and off putting for users, but thinking about applying this type of visualisation to a simple set of data then this could work well.	I do find this quite hard to follow	Knowing how much personal data is included in this visualisation, I think I would run into problems showing my sources/evidence. Also the large number of data points makes me wonder how I could usefully summarise it!	
1	Q1	hard to understand but may be useful if explained more.	No	No change	This kind of visualisation challenges my reliance on text-based communication. Engaging with it helps me to rethink how data analysis might support my research and what it can do with/for traditional scholarship in the arts and humanities	
	Q2	3	2	2	4	2.75
	Q3	It is useful to see his milieu through an email snapshot.	I am not sure that this type of visualisation with information redacted is particularly useful to my work or users	Without contextual information, it is hard to see how this might be useful in my work	It does look useful, but I think I would need to work out *how* I might incorporate this kind of vis into my research.	
2	Q1		No	No change	It helps me to hone in on key interlocutors/domains, which might lead to useful conclusions about the filmmaker's creative activity during that time.	
	Q2	5	5	4	6	5
	Q3		The ability to pinpoint one item of information is very useful for giving an overview and could be also used in a simplified way to show catalogue levels.	There is potentially some use in enabling the categorisation of email contacts, which might be hard to do without a visualisation	If the visualisation reveals information about institutions and organisations, this could be helpful.	
3	Q1	without names this is not useful	No	no change	I'm interested in the arrangement of nodes, and what parameters arranged them in this particular way. Learning how to 'read' this visualisation would be a useful step I think.	
	Q2	1	2	2	4	2.25
	Q3	meaningless without names	I find this visualisation too 'busy' and very hard to understand with the added redaction and noise. I don't think this would be useful.	Without names I can't see much use for this in my work	Hard to say exactly how I would use this visualisation! I think it would need to be supplemented by more analysis	

Table 17 A table of responses to Stage Three of the study for mountain graphs: each participant's understanding of the visualisation (columns 3–6) and usefulness ranking (last column)

Mountain Graph						
PrivCon	Question	Participant 1	Participant 2	Participant 3	Participant 4	Average
0	Q1	I find this much easier to read than the network graph. Blocks seem to convey the proportion of contact better for me. For colour blind people maybe a tonal one rather than a colour one?	No	no change	Just that the addition of a timeline is helpful to observe patterns over time.	
	Q2	4	5	5	5	4.47
	Q3		Visually appealing with the use of colour, I think simplified it could be used to show 'amounts' more clearly than maybe the word tree and network graph. But again accessibility could be an issue due to use of colour.	This could be useful in understanding how records relate to events / trends over time, e.g. spikes in activity	I think this potentially shows more detail of the kind that would be helpful to my research than the previous visualisations.	
1	Q1	useful if you know who it is and the dates - can be mapped on to events and film production - who might be doing what.	No	No change	I don't think so!	
	Q2	6	5	4	5	5
	Q3	if it had the names/dates it would be very useful	Visually appealing with the use of colour, I think simplified it could be used to show 'amounts' more clearly than maybe the word tree and network graph. But again accessibility could be an issue due to use of colour. The ability to redact could be useful	This works better with some contextual information	This visualisation is useful for inspiring me to think differently about digital communications!	
2	Q1			No change	I don't think so!	
	Q2	6	5	5	5	5.25
	Q3	useful for understanding patterns of behaviour	Visually appealing with the use of colour, This shows 'amounts' more clearly than maybe the word tree and network graph. But again accessibility could be an issue due to use of colour.	the potential weakness of this is in the subjective choice of categories	It looks like a useful starting point - it would be even more helpful to categorise more correspondence 'types' (perhaps in cross-reference to the word cloud visualisations?)	

Table 18 A table of responses to Stage Three of the study for scatter graphs: each participant's understanding of the visualisation (columns 3–6) and usefulness ranking (last column)

Scatter Plot						
PrivCon	Question	Participant 1	Participant 2	Participant 3	Participant 4	Average
0	Q1	useful for mapping activity and who is involved at different time lines. Useful in tandem with paper archive in this respect.	No	no change	I don't think so!	
	Q2	6	4	4	6	5
	Q3	useful for understanding patterns of behaviour	I think these scatter plots can be difficult to interpret/ read - this example is quite busy. However, being able to add more information on the axis could make the interpretation easier for users than maybe the Mountain graph. Could therefore be used to provide more detailed analysis rather than an overview.	I am not entirely sure how I might use this	Helpful for honing in on key frequency/volumes of contacts - though obviously would need redaction before I could disseminate! So perhaps an 'internal' team tool, rather than an external research dissemination tool.	
1	Q1	need names	No	no change	No	
	Q2	3	4	3	3	3.25
	Q3	without names not v useful	I think these scatter plots can be difficult to interpret/ read - this example is quite busy. However, being able to add more information on the axis could make the interpretation easier for users than maybe the Mountain graph. Could therefore be used to provide more detailed analysis rather than an overview. The ability to redact is useful	other than if used in comparison with other datasets, this is less useful than the version with names	I'm struggling to think of situations where I would need analyse on a frequency/timeline basis the filmmaker's correspondence UNLESS it was tied to a specific output e.g. a film or a specific person or contact.	

Table 19 A table of responses to stage three of the study for bar charts: each participant's understanding of the visualisation (columns 3-6) and usefulness ranking (last column)

Bar Chart						
PrivCon	Question	Participant 1	Participant 2	Participant 3	Participant 4	Average
0	Q1		No	I like the idea that this could provide "extent" information for cataloguing	No further responses	
	Q2	6	6	5	4	5.25
	Q3	learn about the frequency with which the filmmaker emailed people	I suspect the familiarity of a bar chart would make this visualisation a useful tool for the majority of our users.	I can see practical applications that I hadn't picked up on before	Potentially useful if cross-referenced with other visualisations.	
1	Q1		No	no change	Not at the moment	
	Q2	1	6	3	4	3.5
	Q3	no useful information	I suspect the familiarity of a bar chart would make this visualisation a useful tool for the majority of our users. The ability to redact is useful	I think the lack of contextual information makes this hard to use	This data visualisation could be useful if it is subsequently possible/justifiable to identify those highest-frequency contacts.	
2	Q1	allows me to see how much he initiated and how much was reciprocated	No	I agree with this comment "Less useful than mountain and scatter examples, unless to/from/cc is of particular interest"	No	
	Q2	4	6	3	5	4.5
	Q3	gives some information	I suspect the familiarity of a bar chart would make this visualisation a useful tool for the majority of our users. Can provide a lot of detail.	Other visualisations probably do this job better	Again - potential for usefulness is in conjunction with other visualisations.	
3	Q1	i'm not sure how the 'noise' affects the information	No	no change	I think I'm struggling to identify the key differences between the type of data being presented here and in the previous visualisation?	
	Q2	1	6	3	4	3.5
	Q3	not sure how to read this in relation to the actual information	I suspect the familiarity of a bar chart would make this visualisation a useful tool for the majority of our users.	not sure how this is helpful	I think it's difficult to say how useful this would be. I'd need to reflect on any time-critical periods in the filmmaker's life where his email correspondence became particularly important for understanding his creative practices.	

Table 20 A table of responses to Stage Three of the study for Word Trees: each participant's understanding of the visualisation (columns 3-6) and usefulness ranking (last column)

Word Tree						
PrivCon	Question	Participant 1	Participant 2	Participant 3	Participant 4	Average
0	Q1	The search term is key. Seeing a bit more of each the emails from the search term.	No	The comments indicate that this could be useful, but could be even more useful with some refinements. I like the idea that it could be linked to a more advanced search interface, and also the idea of connecting it to other datasets.	It's interesting that the summarised responses above don't hone in on the emotions and affective responses. It refers to an emotion-based word "hope" and it might be useful for research on emotions and relationships.	
	Q2	5	5	5	6	5.25
	Q3	It could help find emails that would be useful to look at - with more depth and information for research. Help with not having to wade through every one.	I think that this type of visualisation would be easily understood by the majority of our remote users. Keywords could be identified and used to highlight areas of catalogue descriptions and/or digital archives. Having the extra panel could allow a 'quick view' into e.g. an area of the catalogue.	Its usefulness will be highly dependent on the choice of search term	It feels like a useful way of sifting through large amounts of email data, though I wonder whether it might be too revealing in terms of sensitive and personal data. It is a helpful insight into using data analysis tools to work through substantial amounts of digital correspondence.	
1	Q1	redacted information reduces the usefulness for my research.	No	no change	The redaction could make it difficult to identify ongoing conversations between relevant individuals for the research. Perhaps such data could be sifted and applied for release in a more focused way once the overall sense of the emotive content of the correspondence has been established.	
	Q2	3	5	5	5	4.5
	Q3	It could help find emails that would be useful to look at - with more depth and information for research. Help with not having to wade through every one.	Again this visualisation is appealing and easy to read. Having an option in this display to redact information could be useful if dealing with sensitive dataset of archive material	The usefulness of this depends on the choice of search term	As an arts & humanities scholar, questions of attribution are often high on the agenda. The data analysis tool could prove useful in this context, but would need to be in concert with other tools.	

Table 20 continued

Word Tree						
PrivCon	Question	Participant 1	Participant 2	Participant 3	Participant 4	Average
2	Q1		No	I think the absence of the reading panel makes this much less usable	Only that this interface is slightly easier to read. So from a presentations perspective, it is a slightly more accessible interface.	
	Q2	5	4	3	4	4
	Q3	depending on the term could be very useful	It could give a good easily accessible overview of instances of keywords within a catalogue/dataset. Possibly the lack of the pane might prove frustrating for a user, however, as an overview to decide if a collection contains useful content for particular research this works well.	When there is an opportunity to link a finding aid to the content - as with a reading panel - there needs to be a good reason not to do that, and I don't think removing the reading panel brings any advantage	I think I would probably adapt my methodologies to fit around this visualisation: in other words, the visualisation would be useful for my research/practice, but/and would also change the nature of that research in engaging with the visualisation.	

References

1. Scerri, S., Handschuh, S., Decker, S.: Semantic email as a communication medium for the social semantic desktop. In: European Semantic Web Conference, pp. 124–138 (2008). Springer
2. Hangal, S., Chan, P., Lam, M.S., Heer, J.: Processing email archives in special collections. In: DH, pp. 208–211 (2012)
3. Schneider, J., Adams, C., DeBauche, S., Echols, R., McKean, C., Moran, J., Waugh, D.: Appraising, processing, and providing access to email in contemporary literary archives. Arch. Manuscr. **47**(3), 305–326 (2019)
4. Jaillant, L.: After the digital revolution: working with emails and born-digital records in literary and publishers' archives. Taylor & Francis (2019)
5. Noonan, D.W.: Email: an appraisal approach. J. Arch. Organ. **13**(3–4), 146–151 (2016)
6. Prom, C.J.: Preserving Email. Digital Preservation Coalition Heslington (2011)
7. Baker, F.: E-mails to an editor: safeguarding the literary correspondence of the twenty-first century at the university of manchester library. New Rev. Acad. Librariansh. **21**(2), 216–224 (2015)
8. Decker, S., Kirsch, D.A., Kuppili Venkata, S., Nix, A.: Finding light in dark archives: using ai to connect context and content in email. AI & SOCIETY, 1–14 (2021)
9. Koven, J., Bertini, E., Dubois, L., Memon, N.: Invest: intelligent visual email search and triage. Digit. Investig. **18**, 138–148 (2016)
10. Bendersky, M., Wang, X., Najork, M., Metzler, D.: Search and discovery in personal email collections. Found. Trends® Inf. Retriv. **15**(1), 1–133 (2021)
11. Bartliff, Z., Kim, Y., Hopfgartner, F.: A survey on email visualisation research to address the conflict between privacy and access. Arch. Sci., 1–22 (2022)
12. Doss, E., Loui, M.C.: Ethics and the privacy of electronic mail. Inf. Soc. **11**(3), 223–235 (1995). https://doi.org/10.1080/01972243.1995.9960194
13. Jakobi, T., von Grafenstein, M., Smieskol, P., Stevens, G.: A taxonomy of user-perceived privacy risks to foster accountability of data-based services. J. Responsib. Technol. **10**, 100029 (2022)
14. Gharib, M., Giorgini, P., Mylopoulos, J.: Copri v. 2-acore ontology for privacy requirements. Data Knowl. Eng. **133**, 101888 (2021)
15. Moss, M., Gollins, T.: Our digital legacy: an archival perspective. J. Contemp. Arch. Stud.**4** (2017)
16. Jaillant, L.: Introduction. In: Jaillant, L. (ed.) Archives, Access and Artificial Intelligence: Working with Born-digital and Digitized Archival Collections, pp. 7–28. Bielefeld University Press (2022)
17. Biber, K., Luker, T.: Evidence and the archive: Ethics, aesthetics, and emotion. Taylor & Francis (2014)
18. Li, J., Hu, X., Xiong, P., Zhou, W., et al.: The dynamic privacy-preserving mechanisms for online dynamic social networks. IEEE Trans. Knowl. Data Eng. (2020)
19. Carpenter, L., Jackson, T.W., Matthews, G., Thomas, D., Spencer, A.: The role of it in email preservation and archiving. In: 18th International Conference on Automation and Computing (ICAC), pp. 1–6 (2012). IEEE
20. Hendery, R., Burrell, A.: Playful interfaces to the archive and the embodied experience of data. Journal of Documentation (2019)
21. Louis, A., Engelbrecht, A.P.: Unsupervised discovery of relations for analysis of textual data. Digit. Investig. 7(3–4), 154–171 (2011)
22. Kaczmarek, J., West, B.: Email preservation at scale: Preliminary findings supporting the use of predictive coding. Open Sci. Framew. (2019). https://doi.org/10.17605/OSF.IO/6YP9J
23. Moss, M., Thomas, D., Gollins, T.: Artificial fibers-the implications of the digital for archival access. Front. Digit. Hum. **5**, 20 (2018)
24. Stadlinger, J., Dewald, A.: A forensic email analysis tool using dynamic visualization. J. Digit. Forensics Secur. Law **12**(1), 6 (2017)
25. Hangal, S., Lam, M.S., Heer, J.: Muse: Reviving memories using email archives. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology, pp. 75–84 (2011)
26. Hangal, S., Piratla, V., Manovit, C., Chan, P., Edwards, G., Lam, M.S.: Historical research using email archives. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, pp. 735–742 (2015)
27. Borden, B.B., Baron, J.R.: Opening up dark digital archives through the use of analytics to identify sensitive content. In: 2016 IEEE International Conference on Big Data (big Data), pp. 3224–3229 (2016). IEEE
28. Nix, A., Decker, S.: Using digital sources: the future of business history? Bus. Hist. 1–24 (2021)
29. Langdon, J.: Describing the digital: the archival cataloguing of born-digital personal papers. Arch. Rec. **37**(1), 37–52 (2016)
30. Srivastava, S.B., Goldberg, A.: Language as a window into culture. Calif. Manag. Rev. **60**(1), 56–69 (2017)
31. Dabbish, L.A., Kraut, R.E.: 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, pp. 431–440 (2006)
32. Whittaker, S., Sidner, C.: Email overload: exploring personal information management of email. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 276–283 (1996)
33. Stolfo, S.J., Hershkop, S., Hu, C.-W., Li, W.-J., Nimeskern, O., Wang, K.: Behavior-based modeling and its application to email analysis. ACM Trans. Internet Technol. (TOIT) **6**(2), 187–221 (2006)
34. Bird, C., Gourley, A., Devanbu, P., Gertz, M., Swaminathan, A.: Mining email social networks. In: Proceedings of the 2006 International Workshop on Mining Software Repositories, pp. 137–143 (2006)
35. Golbeck, J., Hendler, J.A.: Reputation network analysis for email filtering. In: CEAS, pp. 1–8 (2004)
36. Chapanond, A., Krishnamoorthy, M.S., Yener, B.: Graph theoretic and spectral analysis of enron email data. Comput. Math. Org. Theory **11**(3), 265–281 (2005)
37. Rowe, R., Creamer, G., Hershkop, S., Stolfo, S.J.: Automated social hierarchy detection through email network analysis. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis, pp. 109–117 (2007)
38. Pszota, D.: Email communication transformation into knowledge base. Published on: Dec (2012)
39. Bellotti, V., Ducheneaut, N., Howard, M., Smith, I.: Taking email to task: the design and evaluation of a task management centered email tool. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 345–352 (2003)
40. Garriss, S., Kaminsky, M., Freedman, M.J., Karp, B., Mazieres, D., Yu, H.: Re: Reliable email. In: NSDI, vol. 6, pp. 22–22 (2006)
41. Zou, C.C., Towsley, D., Gong, W.: Email virus propagation modeling and analysis. Department of Electrical and Computer Engineering, Univ. Massachusetts, Amherst, Technical Report: TR-CSE-03-04 (2003)
42. Hershkop, S.: Behavior-based Email Analysis with Application to Spam Detection. Citeseer (2006)
43. Kennedy, R.: Affecting evidence: Edith thompson's epistolary archive. Aust. Fem. Law J. **40**(1), 15–34 (2014)

44. Golbeck, J., Gerhard, J., O'Colman, F., O'Colman, R.: Scaling up integrated structural and content-based network analysis. Inf. Syst. Front. **20**(6), 1191–1202 (2018)
45. Heibi, I.: A visual framework for graph and text analytics in email investigation. Master's thesis, University of Bologna (2017)
46. Magalingam, P., Rao, A., Davis, S.: Identifying a criminal's network of trust. In: 2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems, pp. 309–316 (2014). https://doi.org/10.1109/SITIS.2014.64
47. Park, P., Weber, I., Macy, M.: The mesh of civilizations in the global network of digital communication. PLoS ONE **10**(5), 0122543 (2015)
48. Smilkov, D.: Understanding email communication patterns. PhD thesis, Massachusetts Institute of Technology (2014)
49. Straub, K.M.: Data mining academic emails to model employee behaviors and analyze organizational structure. Master's thesis, Virginia Tech (2016)
50. Tsetini, M.: Computer forensics on financial crimes. PhD thesis, Thessaloniki - Greece (2015)
51. Štorga, M., Mostashari, A., Stanković, T.: Visualisation of the organisation knowledge structure evolution. J. Knowl. Manag. (2013)
52. Wen, Q., Gloor, P.A., Fronzetti Colladon, A., Tickoo, P., Joshi, T.: Finding top performers through email patterns analysis. J. Inf. Sci. **46**(4), 508–527 (2020)
53. Zhang, J.: Miteams: quick organizational mapping by combining email and survey data. Master's thesis, Massachusetts Institute of Technology (2018)
54. Bulkley, N.: Email and output: Communication effects on productivity. PhD thesis, University of Michigan (2006)
55. Mondal, S., Shukla, M., Lodha, S.: Privacy aware temporal profiling of emails in distributed setup. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1229–1238 (2017)
56. Perer, A., Shneiderman, B., Oard, D.W.: Using rhythms of relationships to understand e-mail archives. J. Am. Soc. Inform. Sci. Technol. **57**(14), 1936–1948 (2006)
57. Lu, Q., Zhang, Q., Luo, X., Fang, F.: An email visualization system based on event analysis. In: CCF Conference on Computer Supported Cooperative Work and Social Computing, pp. 658–669 (2019). Springer
58. Perer, A., Smith, M.A.: Contrasting portraits of email practices: visual approaches to reflection and analysis. In: Proceedings of the Working Conference on Advanced Visual Interfaces, pp. 389–395 (2006)
59. Viégas, F.B., Golder, S., Donath, J.: Visualizing email content: portraying relationships from conversational histories. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 979–988 (2006)
60. Cadman, R., MacDonald, B.H., Soomai, S.S.: Sharing victories: characteristics of collaborative strategies of environmental non-governmental organizations in canadian marine conservation. Mar. Policy **115**, 103862 (2020)
61. Luo, S.J., Huang, L.T., Chen, B.Y., Shen, H.W.: Emailmap: Visualizing event evolution and contact interaction within email archives. In: 2014 IEEE Pacific Visualization Symposium, pp. 320–324 (2014). IEEE
62. King, V.: Self-portrait with mortar board: a study of academic identity using the map, the novel and the grid. Higher Educ. Res. Dev. **32**(1), 96–108 (2013)
63. Thanh Tung, V., et al.: Email search visualization: An efficient way for searching email. Master's thesis, Helsingfors universitet (2014)
64. Schreck, T.: Visual-interactive analysis with self-organizing maps-advances and research challenges. In: Self-Organizing Maps, pp. 83–96. IntechOpen (2010)
65. Mandic, M., Kerne, A.: Visualizing rhythms of intimacy in email communication. Interface Ecology Lab, Center for Digital Libraries (2004)
66. Viégas, F.B.: Revealing individual and collective pasts: Visualizations of online social archives. PhD thesis, Massachusetts Institute of Technology (2005)
67. Save, M.V., et al.: People oriented email: A social approach to email interfaces. Master's thesis, North Carolina State University (2020)
68. Whittaker, S., Jones, Q., Nardi, B.A., Terveen, L.G., Creech, M., Isaacs, E., Hainsworth, J.: Contactmap: using personal social networks to organize communication in a social desktop. In: CSCW Videos, p. 7 (2002)
69. Weisgerber, C., Butler, S.: Visualizing the future of interaction studies: Data visualization applications as a research, pedagogical, and presentational tool for interaction scholars. Electron. J. Commun. **19**(1–2) (2009)
70. Butavicius, M.A., Lee, M.D., Pincombe, B.M., Mullen, L.G., Navarro, D.J., Parsons, K.M., McCormac, A.: An assessment of email and spontaneous dialog visualizations. Int. J. Hum. Comput. Stud. **70**(6), 432–449 (2012)
71. Sperr, E.: Word Trees for Visualizing PubMed Search Results (2019)
72. Vane, O.: Text visualisation tool for exploring digitised historical documents. In: Proceedings of the 2018 ACM Conference Companion Publication on Designing Interactive Systems, pp. 153–158 (2018). http://hdl.handle.net/10724/38661
73. Scells, H., Zuccon, G.: Searchrefiner: A query visualisation and understanding tool for systematic reviews. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 1939–1942 (2018)
74. Cuper, M.: Researching pandemics through time: A covid-19 inspired data-driven approach to explore historical newspapers. In: International Conference on Theory and Practice of Digital Libraries, pp. 227–231 (2021). Springer
75. Shen, H., Bednarz, T., Nguyen, H., Feng, F., Wyeld, T., Hoek, P.J., Lo, E.H.: Information visualisation methods and techniques: state-of-the-art and future directions. J. Ind. Inf. Integr. **16**, 100102 (2019)
76. Schneier, B.: Data and Goliath: The Hidden Battles to Capture Your Data and Control Your World, 1st edn. W. W. Norton & Company (2015)
77. Ellen, L.: Breaking rules for good? How archivists manage privacy in large-scale digitisation projects. Arch. Manuscr. **46**(3), 289–308 (2019). https://doi.org/10.1080/01576895.2018.1547653
78. Jaillant, L.: More data, less process: a user-centered approach to email and born-digital archives. Am. Archiv. 85(2), 533–555 (2022). https://doi.org/10.17723/2327-9702-85.2.533
79. Crossen-White, H.L.: Using digital archives in historical research: What are the ethical concerns for a 'forgotten' individual? Res. Ethics **11**(2), 108–119 (2015). https://doi.org/10.1177/1747016115581724
80. DeCew, J.W.: In Pursuit of Privacy: Law, Ethics, and the Rise of Technology. Cornell University Press (1997)
81. Buchanan, T., Paine, C., Joinson, A.N., Reips, U.-D.: Development of measures of online privacy concern and protection for use on the internet. J. Am. Soc. Inform. Sci. Technol. **58**(2), 157–165 (2007)
82. Mujtaba, G., Shuib, L., Raj, R.G., Majeed, N., Al-Garadi, M.A.: Email classification research trends: review and open issues. IEEE Access **5**, 9044–9064 (2017)
83. Jaillant, L., Caputo, A.: Unlocking digital archives: cross-disciplinary perspectives on ai and born-digital data. AI Soc. 1–13 (2022)
84. Liu, P., Wang, L.-e., Li, X.: Randomized perturbation for privacy-preserving social network data publishing. In: 2017 IEEE Inter-

national Conference on Big Knowledge (ICBK), pp. 208–213 (2017). IEEE
85. Majeed, A., Lee, S.: Anonymization techniques for privacy preserving data publishing: a comprehensive survey. IEEE Access **9**, 8512–8545 (2020)
86. Bourahla, S., Laurent, M., Challal, Y.: Privacy preservation for social networks sequential publishing. Comput. Netw. **170**, 107106 (2020)
87. Chong, K.M., Malip, A.: Trace me if you can: an unlinkability approach for privacy-preserving in social networks. IEEE Access **9**, 143950–143968 (2021)
88. Fang, Y., Zhao, C., Huang, C., Liu, L.: Sankeyvis: visualizing active relationship from emails based on multiple dimensions and topic classification methods. Forensic Sci. Int. Digit. Investig. **35**, 300981 (2020)
89. Okoli, C., Pawlowski, S.D.: The Delphi method as a research tool: an example, design considerations and applications. Inf. Manag. **42**(1), 15–29 (2004)
90. Ashokan, A., Haas, C.: Fairness metrics and bias mitigation strategies for rating predictions. Inf. Process. Manag. **58**(5), 102646 (2021)
91. Johnson, V., Ranade, S., Thomas, D.: Size matters: The implications of volume for the digital archive of tomorrow—a case study from the UK national archives. Rec. Manag. J. (2014)
92. Sloyan, V.: Born-digital archives at the wellcome library: appraisal and sensitivity review of two hard drives. Arch. Rec. **37**(1), 20–36 (2016)
93. Gooding, P., Smith, J., Mann, J.: The forensic imagination: interdisciplinary approaches to tracing creativity in writers' born-digital archives. Arch. Manuscr. **47**(3), 374–390 (2019)
94. Özdemir, L.: The inevitability of digital transfer: How prepared are UK public bodies for the transfer of born-digital records to the archives? Rec. Manag. J. **29**(1–2), 224–239 (2019)
95. Moretti, F.: Distant Reading. Verso Books (2013)
96. Buurma, R.S., Heffernan, L.: Search and replace: Josephine miles and the origins of distant reading. Modern. Modernity **3**(1) (2018)
97. Martos Núñez, E., Martos García, A.: Categorizations of reading and cultural praxis in the digital age: distant reading vs. close reading. Investigación bibliotecológica 32(74), 19–33 (2018)
98. Murrieta-Flores, P., Howell, N.: Contested spaces: Creating computational approaches for the holistic analysis of space and place in digital humanities. In: DH (2017)
99. Joo, S., Hootman, J., Katsurai, M.: Exploring the digital humanities research agenda: a text mining approach. J. Doc. (2021)
100. Dobson, J.E.: Critical Digital Humanities: the Search for a Methodology. University of Illinois Press (2019)
101. Bartliff, Z., Kim, Y., Hopfgartner, F., Baxter, G.: Leveraging digital forensics and data exploration to understand the creative work of a filmmaker: a case study of Stephen Dwoskin's digital archive. Inf. Process. Manag. **57**(6) (2020)
102. Lastilla, L., Ammirati, S., Firmani, D., Komodakis, N., Merialdo, P., Scardapane, S.: Self-supervised learning for medieval handwriting identification: a case study from the vatican apostolic library. Inf. Process. Manag. **59**(3), 102875 (2022)
103. Kirschenbaum, M.G.: The remaking of reading: data mining and the digital humanities (2007)
104. Bassett, C., Berry, D.M., Fazi, B., Pay, J., Roberts, B.: Critical digital humanities and machine-learning. In: ADHO 2017-Montréal (2017)
105. Task Force on Technical Approaches for Email Archives: The Future of Email Archives: A Report from the Task Force on Technical Approaches for Email Archives. Council on Library and Information Resources (2018). www.clir.org/pubs/reports/pub175/
106. Misztal-Radecka, J., Indurkhya, B.: Bias-aware hierarchical clustering for detecting the discriminated groups of users in recommendation systems. Inf. Process. Manag. **58**(3), 102519 (2021)
107. Seyedhoseinzadeh, K., Rahmani, H.A., Afsharchi, M., Aliannejadi, M.: Leveraging social influence based on users activity centers for point-of-interest recommendation. Inf. Process. Manag. **59**(2), 102858 (2022)
108. Bag, S., Kumar, S.K., Tiwari, M.K.: An efficient recommendation generation using relevant Jaccard similarity. Inf. Sci. **483**, 53–64 (2019)
109. Silveira, T., Zhang, M., Lin, X., Liu, Y., Ma, S.: How good your recommender system is? A survey on evaluations in recommendation. Int. J. Mach. Learn. Cybern. **10**(5), 813–831 (2019)
110. Bobadilla, J., Serradilla, F., Bernal, J.: A new collaborative filtering metric that improves the behavior of recommender systems. Knowl.-Based Syst. **23**(6), 520–528 (2010)
111. Kenderdine, S.: Experimental museology: immersive visualisation and cultural (big) data. Exp. Museol. 15 (2021)
112. Rehm, G., Lee, M., Moreno-Schneider, J., Bourgonje, P.: Curation technologies for cultural heritage archives: Analysing and transforming a heterogeneous data set into an interactive curation workbench. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage, pp. 117–122 (2019)