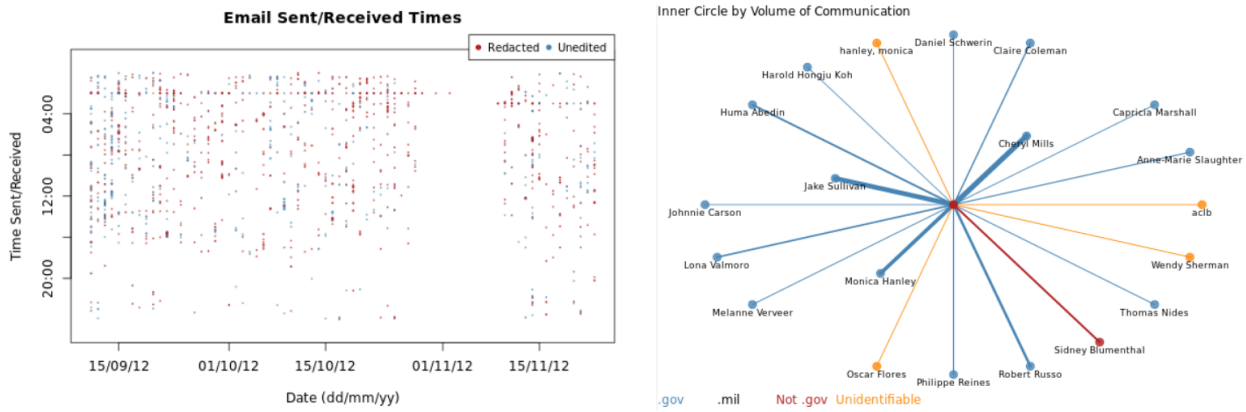
Interactive Filter and Display of Hillary Clinton's Emails: A Cautionary Tale of Metadata

Christopher D. Salahub and R. Wayne Oldford



(a) An empty email block: October 30 - November 9

(b) Inner Circle: Correspondent info by colour; volume by thickness.

Fig. 1. Metadata: Time, date, redaction, correspondents, volume, source type. Filtered: Sept. 11 to Nov. 23, 2012.

Abstract—We present a web-based visualization that allows the user to interactively filter and display characteristics of 32,795 Hillary Clinton's emails as provided by Wikileaks.

The visualization focuses on the meta-data of each email, including its senders, receivers, and the timestamp the email appeared on the Clinton server. An interactive time range slider filters all email and all displays automatically update to changes in the slider. The main display shows Clinton's most frequent correspondents arranged as nodes of a spoked graph with Clinton at the centre. Volume determines the thickness of each spoke and high volume determines an inner circle whose spokes are shortened. Correspondents and their edges are coloured according to whether that email account could be identified as being an approved Federal government account or not. A second display shows two daily time series: the total number of emails for that day, and the number meeting selection criteria. A third display shows a scatterplot of the time of day versus the day on which that email appeared. Scatterplot points are coloured by whether the email was redacted or not.

Other displays add some information beyond metadata. FOIA exemption codes appear as a selectable list and a barplot shows email counts by FOIA code. The (stemmed) terms having highest frequency in the displayed email, and those having highest tf-idf are listed in separate displays. All displays are interactively filtered by time range and selected FOIA codes.

We illustrate how the filtered displays can be used to generate hypotheses and uncover interesting information. These touch on contentious issues including the handling of classified information, the 2012 attack on the Benghazi U.S. diplomatic compound, and emails apparently missing from those released publicly.

The data are extracted from Wikileaks HTML files, cleaned, and stored in a form useful for interactive exploration. A local `R shiny` server provides the interactive displays as a public service online tool to explore and uncover patterns in the meta-data and summary contents of Clinton's email. Coupled with publicly available sources of information, these interactive tools uncover surprising amounts of information about an individual, especially one holding public office. The ease with which this can be accomplished and shared should serve as a clear warning as to what can be learned about anyone from metadata.

Index Terms—Exploratory data analysis, metadata, text mining, web-scraping, interactive web visualization, `R`, `shiny`

1 Introduction

The 2016 U.S. Presidential election was one of the most contentious in history. The existence and possible content of Hillary Clinton's private email server dogged former Secretary Clinton's bid for the U.S. presidency and was doubtless a contributing factor to her surprising defeat by Donald Trump.

On March 16, 2016 Wikileaks published a searchable archive [49] containing the contents of more than 30,000 emails (and attachments) that were sent to and from Secretary Clinton on her private server. The documents were provided as pdfs by the U.S. Department of State in response to Freedom of Information Act (FOIA) [14] requests. The State Dept. also provided a searchable web archive of the documents, released in several instalments from May 2015 to March 2017 [11]. Both sites provide a useful tool for anyone searching for particular terms in the documents. The Wikileaks site was put to much use by investigative reporters, and others, to search for topical news items.

What is critically missing from either site is a facility to learn summary, or statistical, features across all, or groups of, emails. A separate

- *Christopher D. Salahub, University of Waterloo, Canada E-mail: csalahub@uwaterloo.ca.*
- *R. Wayne Oldford, University of Waterloo, Canada E-mail: rwoldford@uwaterloo.ca.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

site where visual displays of summary features of the emails, which can be filtered by user selection, will go a long way towards filling this need. The site would be *complementary* to the more typical "content search and record display" site as provided by either online archive. Add general internet search in a third browser, and the three combined would provide the means to actively explore the contents of any email archive. One imagines, for example, having three browsers open simultaneously, one for the data base search (e.g. Wikileaks), one for statistical visualization and exploration (proposed here), and one for internet search (to provide important context for the other two services). Each fills a need not met by the other two; the synergy of all three provide the user a powerful set of investigative tools.

In this paper, we describe an implementation of such a visual analytic service and illustrate it on the Secretary Clinton's e-mail archive. The implementation is located at `rshiny.math.uwaterloo.ca/clinton` where any visitor may interactively explore summary features of the entire corpus of emails. Different visual displays present different salient features from all selected emails; emails can be selected via a variety of interactive filters. All displays are reactive and update simultaneously in response to user interaction with the data filters. Together, filters and displays provide a visual analysis service that can be used to quickly learn more about, and to discover some (possibly unanticipated) patterns in the emails.

The implementation is described over the next three sections, beginning with discussion of the data in Section 3 which discusses the source of the data, the extraction and cleaning process in Subsection 3.1, the details of the metadata in Subsection 3.2, the use of redaction codes in Subsection 3.3 and content data used in Subsection 3.4. The displays of the data and their patterns over the entire unfiltered data are then described in Section 4. Within this section, Subsection 4.1 describes the spoked network plot, Subsection 4.2 addresses the email volume time series plot, Subsection 4.3 discusses the scatterplot of daily sending times, Subsection 4.4 outlines the exemption code barplot, and Subsection 4.5 explains the content tf-idf and frequency information. Interactive filters which are implemented in the application are discussed in Section 5. Finally, brief analysis of some interesting findings is included in Section 6, alongside a summary of the timeline of the Clinton email imbroglio in Section 2.

2 A BRIEF TIMELINE ON THE PRIVATE EMAIL SERVER

While the following timeline is somewhat abbreviated, it should serve to raise some of the major issues and concerns related to the private email server and its contents. It also introduces some of the principal characters involved. More complete and in depth timelines are readily available elsewhere (e.g. [2, 22, 28, 45, 51]).

On November 21, 2008, the New York Times reported that Hillary Clinton had decided to accept the position of U.S. Secretary of State. On January 13, 2009 the internet domain name `clintonemail.com` was registered [3]; eight days later Senator Clinton was confirmed as Secretary of State.

Public knowledge that a private email server being used by Secretary Clinton and others for State Department and personal communications did not surface until March 2015 [42] during the course of a U.S. Congressional investigation [8] of the September 11, 2012 attack by militants on U.S. compounds in Benghazi Libya.

The State Department had difficulty fulfilling public FOIA and House Benghazi Committee requests [32] for Secretary Clinton's government emails because she had exclusively used the private server for all her email. On March 10, 2015, Clinton told reporters that she turned over 30,490 emails to the State Department and deleted 31,830 emails deemed to be personal [22]. Clinton had tasked three lawyers Cheryl Mills (Clinton's former chief of staff), David Kendall (Clinton's personal lawyer), and Heather Samuelson (a State Department staffer during Clinton's tenure) to make the determinations as to which emails were work related and which were not [9, 45].

On March 10, 2015 the House Benghazi Committee requested that the private email server be turned over to a neutral third party to determine which emails are personal and which are government records [39], but was informed March 27 by David Kendall that no emails remained on the private server for any kind of review [40]. Between March 25 and 31, 2015, Paul Combetta (then the server's system administrator), erased all backup copies using `BleachBit` (see `www.bleachbit.org`).

Combetta, Mills, and Samuelson will later be granted partial immunity by the Justice Department during the FBI investigations into the private email server, as were two others: Bryan Pagliano (original server manager) and John Bentel (former director of Information Resources Management for the State Department's Executive Secretariat) [18, 19, 36].

On April 12, 2015 Clinton announces that she is running for the U.S. Presidency. On July 24, 2015, inspectors general for the State Department and the national intelligence agencies announce finding classified information in the emails and that the information they found was classified at the time sent [41], though her campaign declared that they must have been classified after the fact. On August 19, 2015, Clinton calls the allegation of mishandling classified information a "disagreement between agencies" [54].

Nearly one year later, July 5, 2016, FBI Director James Comey recommended that no charges be laid against Clinton on use of private email server [24]. On October 28, 2016, Comey revealed that in a separate investigation into former Congressman Anthony Weiner, that emails belonging to his wife Huma Abedin have been found on his laptop. Since Abedin was a close aid to former Secretary Clinton, FBI investigations were reopened into the private server usage but closed again by Comey on November 6, 2016 without charges being laid [55]. In both cases, Comey and the FBI are criticized by pundits from different political parties.

3 THE DATA

As of March 3, 2017, a total of 32,795 available emails, either to or from Hillary Clinton, have been made publicly available in pdf form as a searchable database [11]. Many of these have been redacted according to the FOIA exemption codes [14].

Wikileaks [49] has provided the same redacted pdfs and, more usefully for analysis purposes, an HTML version of each pdf. Consequently, have used the Wikileaks database as our data source. All data extraction, cleaning, analysis, and presentation is done using the open source statistical programming language R [35].

3.1 Data extraction and cleaning

The raw HTML of each message was programmatically downloaded from the Wikileaks archive using R packages `RCurl` [25] and `XML` [26]. This took several hours and required the use of manually inserted system pauses to prevent time-outs in the connection, most likely due to Wikileaks DDoS (distributed denial of service) protection software. Besides avoiding such DDoS protection, these pauses are considered web-scraping etiquette and best practice. Once downloaded, the HTML data were processed using the R packages `tm` [15, 16], `stringr` [48], and `SnowballC` [5].

After downloading the raw HTML and extracting the data of interest, the resulting data was stored in `csv` files to provide the flexibility to load the data into any architecture or analysis tool desired. These will be transferred to a relational data base should our server traffic warrant it.

3.2 Metadata

For each email, from the HTML we extract as best we can the identity of the persons sending and receiving the email as well as the date and time at which the email was processed by the server. The fields used were the address to, address from, contact name to, contact name from, subject line, and time. As well, forwarding chains of email addresses were captured through the identification of any to or from fields followed by emails within the text. In cases where no contact name was present the address was substituted. When the address was missing no imputation was completed. Carbon copy information was not extracted.

The HTML was constructed from email printed out, redacted, and then provided as pdfs by the State Department. Consequently, detailed email header information as would normally be available electronically is mostly missing. All time stamps appear to be the local date and time at which the server sent or received that email (e.g. no time zone or other source time or IP chain information is available). Moreover, because of redaction (typically FOIA exemption B6 [14]), sender and receiver emails may have truncated domains, contain only the person's name, or be missing altogether. In cases where both `From` and `To` are entirely missing, but there is an email chain within the message, we do not impute values (e.g. see `https://wikileaks.org/clinton-emails/emailid/31599`).

Using regular expressions to extract the metadata was occasionally challenging and it is always possible that some fringe cases have been mishandled. On the whole the metadata is fairly consistent and only rarely do some impure and messy addresses and contact names arise due to irregular spacing or placement of text within the HTML code. One such fringe case is Huma Abedin's `clintonemail.com` email account which will show in the displays as an overly long string. We have chosen not to special case this but leave it as is.

Where possible, the email addresses have also been parsed so that they may be classified into one of four categories: those which are `.gov`, those which are `.mil`, those which are identifiable as coming from a domain that is neither `.gov` nor `.mil`, and those whose domain was not identifiable from the data.

3.3 Redaction information

Emails that are redacted are marked as such by the string "RELEASE IN PART" and by the presence of one or more of nine FOIA exemption codes B1, B2, ..., B9 marking the place where text is missing (redacted). This provides two further pieces of quasi-metadata (since it is not actual email content) that can be used in analysis.

3.4 Content information

The entire (redacted) content of each email is available on the State Department and Wikileaks sites and the user is encouraged to view it there (the pdf forms are more informative, especially in appreciating the extent of the redactions). To provide some coarse statistical summaries of the content, all word tokens are extracted and partially processed to reduce their number. For example, "stopwords", as identified by the `stopwords` function (with both `kind = "en"` and `kind = "SMART"`), or as from a set of custom stopwords, were removed. Remaining words were stemmed by the `stemDocument` function from `tm` and `SnowballC`. While neither the stopword lists nor the stemming tools are particularly well tuned to this corpus of emails, they nevertheless provide some hints about the topics covered.

4 Displays

TODO Each of these four displays is described and commented on in context of whole time line Four displays of the metadata and quasi-metadata

4.1 Inner circle

This display takes the most frequent correspondents (max. 20) in the selected emails and arranges their email addresses at the ends of equiangular spokes having Secretary Clinton as hub. The hub is coloured red to show that this account is known not to be a government sponsored account (either `.gov` of `.mil`). Every account that is identifiably not a government sponsored account is coloured red. Those that are identifiably government are coloured either blue (for `.gov`) or black (for `.mil`, if any appear). Those which can not be determined as either government or not, are coloured orange.

Figure 2 shows the top 20 correspondents (or "Inner Circle") over all emails in the collection. Note that Huma Abedin appears both as a blue `.gov` account and as a red non-government (`clintonemail.com`) account. The other red account is that of Sidney Blumenthal. This has been a source of some controversy [13, 38], since he had been rejected by the State Department yet appears to provide Clinton advice throughout her tenure. Wendy Sherman appears as orange. Sherman

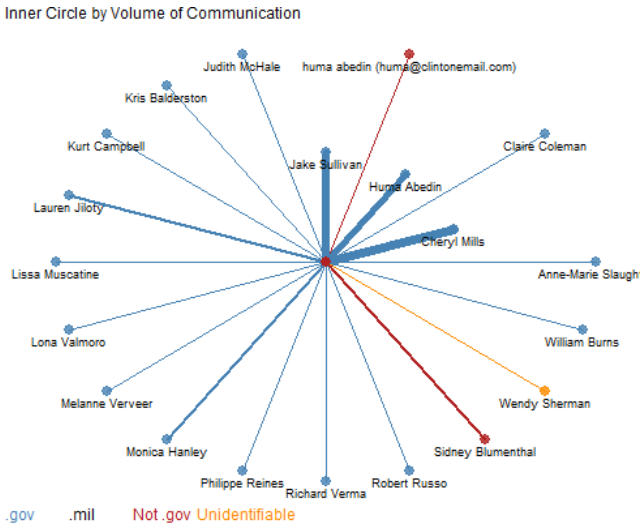


Fig. 2. Inner Circle over all correspondence

was appointed by Clinton to be Under Secretary of State for Political Affairs in 2011 and the high frequency of Sherman emails dates from this time. Nevertheless, from this archive the email address could not be identified definitively as being either government or non-government. Either the address was redacted for privacy reasons (a B6 FOIA exemption), or it was simply unavailable in the header provided.

The greater the number of emails between correspondents, the wider is their connecting spoke. Whenever the difference in volume is great enough, correspondents are separated into two groups: those with the greatest correspondence have shorter spokes visually placing them "closer" to Clinton. As Figure 2 shows shows, Clinton's closest inner circle are her closest aides Huma Abedin, Cheryl Mills, and Jake Sullivan. TODO Chris, can you be more precise about the algorithm?

4.2 Email volume

Figure 3 presents two time series showing the volume of emails for

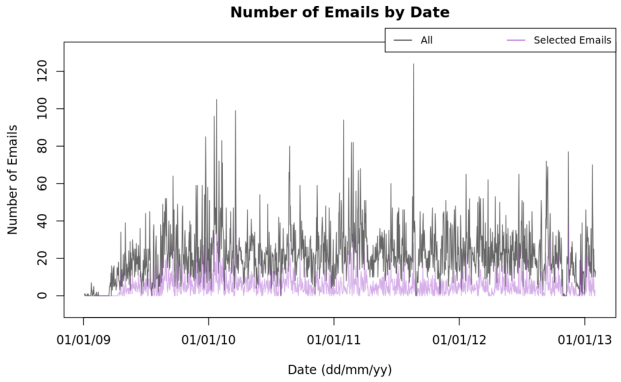


Fig. 3. Email Volume: All emails (top) ; Those sent by Clinton (bottom)

each day. In grey at top is the daily total; in magenta at bottom is the total for the subset chosen by the filters. In Figure 3, the bottom series is the total daily emails sent from Clinton to others.

Over the four year scale shown in Figure 3, the display is quite busy. Even so, a few features stand out. For example, there is a notable lack of email from the first few months of Clinton's tenure; there is essentially no email from Clinton at the beginning and very little to her.

One also notices the spikes of email activity. These can be checked against world events to generate hypotheses about what might be occurring to cause such spikes. For example, the largest spike occurs on August 21, 2011, the beginning day of the battle for Tripoli and on which it was reported that two of Qaddafi's sons had been captured [50]. It is also the day on which the famous "tick tock Libya" email is composed by Jake Sullivan providing a timeline crediting Secretary Clinton with leading the U.S. policy on Libya "from start to finish" [43]. Much of the email that day is redacted (B5 primarily). Other peaks (and valleys) of activity could be similarly investigated.

4.3 Email times

Figure 4 shows all emails sent by Clinton over the four year period.

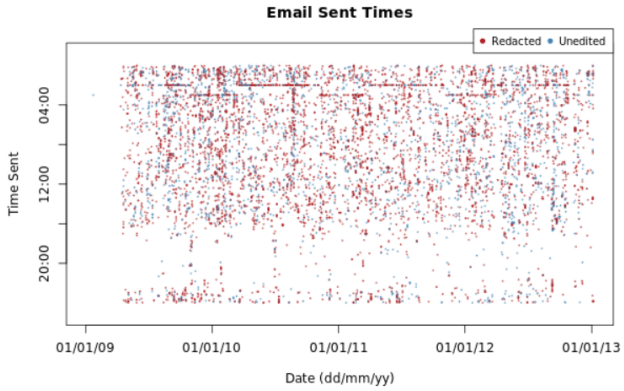


Fig. 4. Email time patterns: all emails sent from Clinton

Each email appears as a point determined by the time shown that it was sent. The calendar date determines the position on the horizontal axis, the (24hr) time of day determines the vertical. Points are coloured red if any part of that email was redacted, and blue if it was released in full. Alpha blending is used to minimize the effect of over-plotting.

A few patterns are easily discerned. For example, the least amount of email appears for a few hours around 20:00 hours, or 8 PM. Evenings appear to be when email traffic is lightest. Otherwise, for the most part, it seems fairly uniformly distributed throughout. Midnight is at the top (and bottom) of the plot, so it would seem that email will occur mainly from midnight until about 4 or 5 PM.

If these are the actual times Secretary Clinton composed and sent email, it suggests that much of that activity occurred in the middle of the night. To reflect the low email activity period as actual "downtime" between "days" (now interpreted as separable 24 hour email cycles), we offer the possibility of choosing 6 PM as the "end" of a day rather than midnight. This places the least active email period as a "quiet" boundary between days. Unfortunately, though it better represents the rhythm of correspondence, it can be confusing since it means that the "next day" begins just after 6 PM.

Note also the surprisingly regular horizontal lines that appear across the top of the plot. This regularity is suggestive of some automated schedule for the server which causes the email to stack up before being recorded at either 2 or 3 AM. The location of these lines switches exactly whenever daylight savings time switches in North America.

When the received emails are added to the plot, the pattern is essentially the same (though much denser) and the horizontal lines where the server has scheduled something are even clearer.

4.4 FOIA exemptions

U.S. Freedom of Information Act (FOIA) exemption codes B1 through B9 are used in each place in an email where information has been redacted. For the authoritative definition the act itself should be consulted [14]. Briefly, they mark each redaction as B1 for national security and foreign policy matters, B2 for personnel practices, B3 for statutory exemptions, B4 for trade secrets or financial information obtained in confidence, B5 for inter- or intra-agency memorandums, B6 for personal privacy, B7 for records compiled for law enforcement, B8 for records prepared in relation to financial monitoring institutions, and B9 for geophysical and geological information concerning oil and gas wells. Each email can contain redactions for any number of exemption codes.

Figure 5 is a barplot of all FOIA exemption codes found in all of the

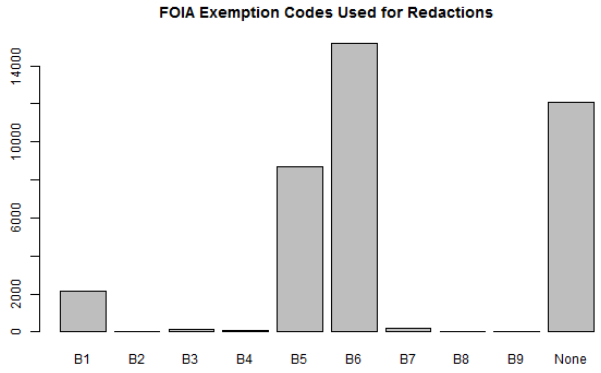


Fig. 5. FOIA barplot: Exemption codes from all email

emails. As can be seen, B1 redactions made for national security and foreign policy reasons account for a small portion of all redactions made though affect more than 2,000 emails in the collection. Nearly half of the emails contain B6 exemptions, those made to protect personal privacy. The inter-agency, intra-agency exemption B5 is used in more than 8,000 emails. This code has come under some criticism from advocates for transparent government as being overly applied [21]. The remaining codes appear in very few emails and about 12,000 emails, or about a third of them, are not redacted at all.

As with all other displays, the FOIA barplot will update in reaction to all filters making it a useful tool to identify how codes co-occur over any selection of emails. Conversely, filtering on FOIA codes can be interesting to see who sees correspondence of varying sensitivities, especially B1 and B5.

4.5 Term frequency and TFIDF

The 20 terms (stemmed words excluding stopwords) which appear in the greatest number of the emails selected are displayed as those having "Highest Frequency". Those terms which appear frequently

20 Highest Frequency Terms	20 Highest TFIDF Terms
will, state, pm, said, secretari, can, call, depart, time, presid, govern, offic, work, usa, one, also, meet, new, us,	msg, pr, pager, folder, fyi, nternet, tx, folderid, rim, ticker, pls, cheryl, send, dev, sorri, fw, soon, jake, delet, true

Fig. 6. Top 20 terms appearing in all emails

within some emails but less frequently across emails will have a high tf-idf (term frequency - inverse document frequency) score. The top 20 scoring of these in the selected emails are shown in the "TFIDF" display. TODO Chris, check that I have described these two measures correctly The top 20 shown in Figure 6 are for all emails in the corpus; these will change depending on the filtering.

The terms give some limited insight into the contents of the email. As seen in Figure 6, for example, the first names of two of Clinton's closest aides, Jake Sullivan and Cheryl Mills, appear under TFIDF but not as high frequency terms. One problem with tf-idf for this e-mail corpus is that many emails contain email chains which grow as each person replies which could magnify the within email frequency for some terms.

As mentioned in Section 3.4, the stemming and stopword removal provided by the package `tm` in R is also challenged by this messy and non-standard data. The list of stopwords had to be supplemented so as to avoid uninformative action verbs such as 'will' and 'can'. The stemming algorithm was also challenged by typographical errors in emails and the proliferation of acronyms (e.g. for individuals and government abbreviations).

5 Filters

There are only three filters but each one affects all data displays. As each filter is changed, every display redraws itself on the filtered data, immediately in reaction to the change. In this way, the filters can be used together to focus on particular subsets of the emails or simply to observe how the display patterns change with the filter being applied.

5.1 Time filtering

A sliding time window filters the emails displayed to those whose date

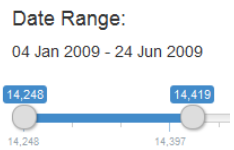


Fig. 7. Time slider: move either end, or move the interval

lies between the two end points. The size of the time window is changed by moving either end, either by mouse click and drag or by selecting an end to move with the arrow keys. The whole time window is moved by selecting the middle bar and either dragging it or moving it with the arrow keys.

This filter is simple but powerful. All emails are displayed when the range covers all dates. Moving the end points towards one another allows the user to focus the displays on any particular range, down to as fine as all emails on a single day. With a fixed range of days, for example a two week period, dragging the middle bar from left to right will have each display smoothly update over time.

5.2 Sent or received

A drop down menu is used to select those emails which Clinton sent, or

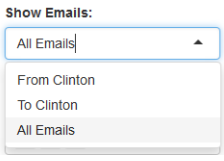


Fig. 8. Filtering on whether Clinton sent or received the email, or either

received, or either sent or received. The default is "All Emails", which is the same as no filter being applied for sent or received. Choosing "From Clinton" selects for display only those mails sent by Secretary Clinton, choosing "To Clinton" selects only those emails which she received. This might be used, for example, to explore whether the inner circle of correspondents changes depending on whether Secretary Clinton is emailing them, or they are emailing her.

5.3 FOIA exemptions

Multiple filtering by FOIA exemption codes is supported using a box selection area. Figure 9 displays the codes that are selected for email

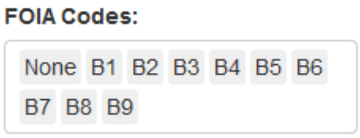


Fig. 9. FOIA codes: Emails containing any of these codes appear

display. Any email having one or more of these codes will appear in the displays. To not have emails display with a B1 exemption code, for example, then the user selects and deletes the B1 from the list. Or to have only the emails that contain B1 redactions displayed, all other codes can deleted from the list. Codes are returned to active display at any time by clicking in the display box and selecting them again from a drop down menu of the excluded codes. Any number of codes (one or more) can be selected to be included in the displays.

5.4 Auxiliary information

To demonstrate how other sources of information can easily be used to supplement the metadata found in the email archive, we consulted the State Department website and retrieved Secretary Clinton's official foreign travel schedule [46]. This information was added to the display as a simple check box "Display Foreign Travel Schedule". When checked, light blue vertical bars appear on the email volume time series display. Appearing underneath the series, these mark time periods when Secretary Clinton was travelling outside the U.S. on State Department business.

6 Some Interactive Analyses

Each of the displays and filters described in the previous sections gives a slightly different and informative view of summary features of the email archive. However, their analytic power is considerably amplified when used together. Rather complex queries can be formed involving date, classification code, and whether Clinton is sending or receiving email. Moreover, given a particular pattern in one display, its relation to that in another can be of considerable interest. Together with access to the archive contents and an internet search engine, an enormous amount can be learned from these visual analytic tools.

In this section, we explore a few avenues of enquiry that might naturally arise. At best this only hints at the analytic power that interactive filtering and display can bring to bear on an email archive. To get a better appreciation, the user is encouraged to try it themselves at `rshiny.math.uwaterloo.ca/clinton`. The tool is useful both for generating and for testing hypotheses.

6.1 The last two years

One might begin an interactive analysis by looking at the email patterns in the last two years of Clinton's tenure as Secretary of State. This is done by simply having the right end of the time slider of Figure 5.1 as far right as possible and then moving the left most end towards the middle of the time period until January 1, 2011 is reached. Figure 10 shows the resulting email volume time series over this period. Note, however, that in addition to filtering on time we have chosen to also filter on FOIA exemption code, selecting only those emails containing one or more B1 (national security and foreign policy matters) redactions. The magenta time series at the bottom then shows all emails, sent or received by Clinton, in the last two years of her tenure which were redacted in part by the State Department (after the fact, not at the time) as being B1 FOIA exempt. As can be seen, there are relatively few of these.

Recall from Section 4.1 that the tallest spike of Figure 10 corresponds to the beginning of the battle for Tripoli (August 21, 2011). Another series of the next largest spikes appear at the far left of this

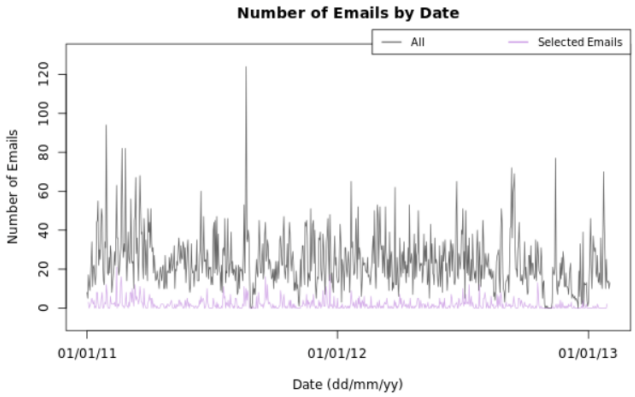


Fig. 10. Last two years: Selected Emails are those that are B1 redacted

time window, at the beginning of 2011. We might choose to compress the time window more, to focus on this section of the emails.

Bringing the right end of the time slider in to April 15, 2011 produces the series of Figure 11. As can be seen, there continue to be highs and

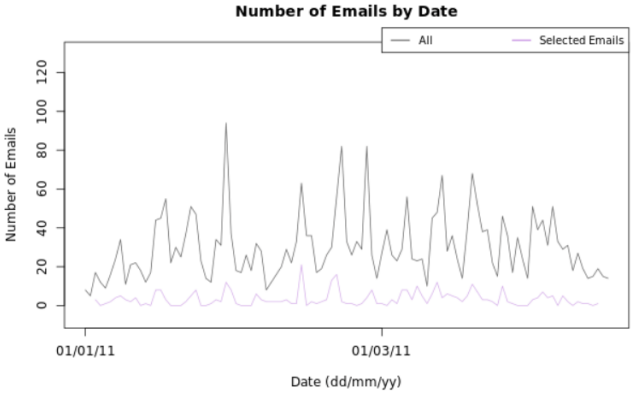


Fig. 11. Jan. 1 – Apr 15, 2011: Selected Emails are those that are B1 redacted

lows of email activity. The highest peak now is at the far left. This is January 29, 2011, the day after President Mubarak of Egypt has ordered the army into the streets of Cairo to quell protests [4, 44]. Dipping into the contents of the Wikileaks email archive, much of the email on this day is seen to be about managing the Egyptian Crisis.

Slightly to the right of this peak, there are several more. At this time, international forces are gathering to consider Libya and its leader Muammar Gaddafi. Consulting Jake Sullivan's "tick tock on Libya" timeline describing Secretary Clinton's leadership on this, we move the sliders in to focus on the time range from February 25 when Secretary Clinton announced the suspension of the Libyan embassy in Washington, to March 18, the day after Secretary Clinton had secured "... Russian abstention and Portuguese and African support for UNSC 1973, ensuring that it passes. 1973 authorizes a no-fly zone over Libya and 'all necessary measures' - code for military action - to protect civilians against Gaddafi's army ..." [43]. On March 19, 2011 NATO military operations in Libya began [52].

The readjusted time series are shown in Figure 12. The left most spike is the rightmost spike of approximately 80 emails from Figure EmailVolumeB1JanApril2011. Of course all other displays have been updating every time the time slider is adjusted.

Some sense of the content of the emails can be had from the updated term frequency displays shown in Figure 13. Not surprisingly, Libya and its leader figure prominently in this restricted set of emails.

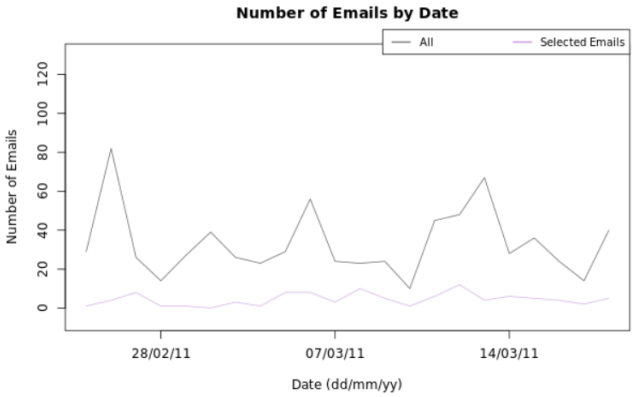


Fig. 12. Feb. 25 – March 15, 2011: Selected Emails are those that are B1 redacted

20 Highest Frequency Terms

d, govern, will, b, email, offic, assist, new, import, qadhafi, libyan, windrush, also, stategov, regist, libya, ventur, limit, interim, state

20 Highest TFIDF Terms

broadway, lp, swh, windrush, obl, ventur, wale, england, virus, pari, limit, outlin, legitimaci, softwar, london, qadhafi, articul, minim, tribal, mahmoud

Fig. 13. Frequent terms: B1 only emails, Feb. 25 – March 18, 2011

Perhaps more surprising is a term like "windrush". A little investigation shows that this is a reference to "Windrush Ventures" (e.g. see [27, 30]), a company owned by former U.K. Prime Minister Tony Blair . "Windrush" appears at the end of emails as part of Mr. Blair's electronic signature. Mr. Blair is part of this B1 selected correspondence.

What other restrictions, as measured by the FOIA exemption codes, can be seen in Figure 14. There appear to be about 100 emails in the

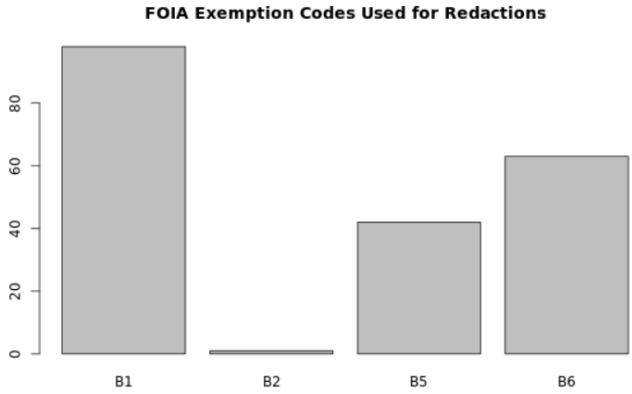


Fig. 14. FOIA code distribution: All B1 redacted, Feb. 25 – March 18, 2011

selected set. In addition to having B1 redactions, many of them also have B5 (inter- or intra-agency) and B6 (personal) redactions.

We might now consider the composition of the inner circle for this set of emails. Figure 15 shows the set of correspondents. Not surprisingly, Jake Sullivan is the principal correspondent. Huma Abedin is next, though she is using two emails, one a government account the other a personal `clintonemail.com` account. Melanie Verveer has the least amount of traffic (judging by the lightness of her spoke colour). Sidney Blumenthal, a non-government employee, is in this inner circle of B1

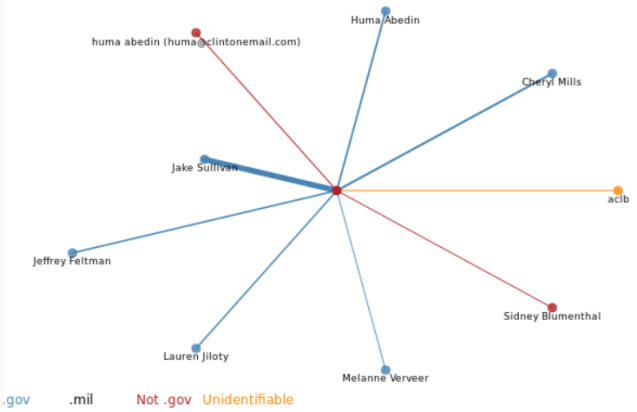


Fig. 15. Inner Circle: All B1 redacted, Jan. 1 – Apr 15, 2011

exempted emails. So too is former U.K. Prime Minister Tony Blair; this is the unidentifiable domain account `aclb`. It should be noted that there could be other accounts which do not appear in the inner circle because their emails were not present in the email record. See Section 3.2.

We could continue to drill down into this information, checking the (unredacted) contents of the emails themselves from either the Wikileaks or State Department sources. We might also more closely connect email dates and correspondents with world events. Instead, we now turn our attention to another feature which appears in Figure 10.

6.2 An email gap?

Towards the very right of Figure 10 there appears to be an anomaly in both time series. There is a noticeable flat area at zero in both series. If we move the leftmost time sliders to the June 1, 2012, and the rightmost to the end, February 1, 2013, we have the email volumes for the last 8 months of the archive. The resulting series are shown in Figure 16. There appears to be no emails for a fairly large chunk of time. In fact,

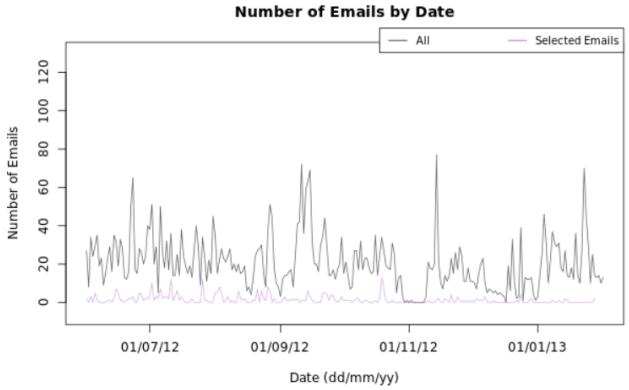


Fig. 16. June 1, 2012 – February 1, 2013: Selected Emails are those that are B1 redacted

there are few or no emails in the archive from October 30 to November 9, 2012. This seems curious, given that the archive is supposed to be complete in Secretary Clinton's work-related email during this time period. The block of no email appears as an even more unusual pattern in Figure 1(a). There the block is conspicuous by the absence amongst the emails from September 11 to November 23, 2012. The email on which the displays of Figure 1 are based, is no longer filtered by FOIA code. It contains all email in the archive in this time period. Direct examination of the archives at Wikileaks and the State Department show no mail in this empty block beyond that seen in Figure 1(a).

The inner circle for all emails from September 11 to November 23, 2012 is shown in Figure 1(b). Those closest to Clinton in terms of volume of correspondence are Clinton aides Jake Sullivan, Cheryl Mills, and Monica Hanley. Monica Hanley is also an account whose identity as government or not could not be determined. Tony Blair and Sidney Blumenthal both appear in the inner circle, as does Wendy Sherman (though the account could not be identified as government or not). Oscar Flores is an assistant to the Clintons, often working at their home. The rest are identified as corresponding through government accounts.

One might ask what is happening around this time period? In a word, Benghazi. Five days after the assault, The U.S. Ambassador to the United Nations, Susan Rice, made a series of television appearances explaining that the attack on the U.S. Benghazi mission was a spontaneous response a "hateful video" which had hours earlier caused violent protests in Cairo against the U.S. Embassy [6]. Four years later, it will be revealed as part of the House Select Committee on Benghazi report [8] that Secretary Clinton had as early as September 12, 2012, communicated to Egyptian Prime Minister Kandil that the attack was known to be planned by Al Qaeda affiliates, it was not a protest, and that it had nothing to do with the film [10]. By the end of September the U.S. administration was calling it a terrorist attack [23]. A month after the attack the administration was defending itself against charges that it deliberately downplayed a terrorist attack in Libya for political reasons in a presidential election year [17]. In October, focus shifted to the security at Benghazi and whether it had been reduced prior to the attack [7]. The House Committee on Oversight & Government Reform began looking into the security aspect [1] in earnest and found evidence that security issues were known and that requests from the U.S. Embassy in Libya for additional security personnel had been turned down by the State Department [29]. On October 26, CIA operatives who had defended the Benghazi mission against the terrorist attacks came forward with their story of what went on, on the ground. They maintain that it was an organized attack and tell reporters that they were told to "stand down" [20]. From October 26 to November 9 2012, no email appears in the archive as being sent from Secretary Clinton.

Moving the time sliders to cover either side of the email gap, October 22 to November 17, 2012, results in the term frequency displays of Figure 17. It would seem that emails were very much related to the

20 Highest TFIDF Terms

nternet, msg, birthday, fbi, folder, pr, cantor, mangoush, sharia, ansar, pager, casualti, petraeus, misrata, militia, heavi, happi, captur, haftar, tx

20 Highest Frequency Terms

state, will, date, forc, said, m, benghazi, call, time, usa, secur, militia, one, sourc, r, hous, secretari, nation, presid, pm

Fig. 17. Around the gap: All emails, Oct. 22 to Nov.17, 2012

unfolding political crisis of Benghazi. Even the name of the terrorist group, Ansar al-sharia, is picked up in the tf-idf terms.

Whatever the reason for the gap seen in the archive, one thing is certain. The story of Benghazi and its aftermath, with its piecemeal and contentious disclosures, and unfolding as it did immediately before a U.S. Presidential election, would consume much of the attention and energy of the U.S. administration, Congress, and the U.S. electorate. It was an event that would haunt Secretary Clinton for months and years to come [53].

6.3 A hunt for no email

The discovery of the obvious email gap in November 2012 raises the possibility that there may be other, less obvious, gaps in the email archive.

To explore this possibility, the time slider was adjusted to the relatively small span of thirty days. This would allow us to detect gaps in email of smaller sizes. To focus attention on Secretary Clinton's behaviour, only those emails sent from Secretary Clinton were displayed.

Moving the slider, from left to right, from early 2009 to early 2013, would allow us to move a one month wide magnifying glass across the whole of the archive. A few contiguous days with no email from Clinton should appear visually as a small horizontal line at zero in the email volume (magenta coloured and lower) time series. Any such horizontal patch will be identified by its date.

We found ten more patches, in addition to the original obvious email gap of Section 6.2. So as to get some idea of the context of the patch, and to illustrate the advantage of coupling a visual analytic tool with an internet search engine, for every patch we found we performed a coarse internet search to see what might turn up, if anything to explain it. If, for example, we found a patch of no emails from Clinton in March 2009, then a Google search as shown in Figure 18 would be tried to see

Fig. 18. A coarse Google search

what might be found in the news that could relate to that date.

6.3.1 You have no mail . . .

The first, and largest, patch occurs at the very beginning of Clinton's tenure as Secretary of State. It is obvious even from the full time series of Figure 3. Essentially no email from Secretary Clinton, and little from anyone else appears until mid to late April 2009.

In time order, the patches and what we learned from each Google search was as follows.

We found ten more places in the archive which we identified as gaps to be explored.

where there were zero emails from Clinton

Dates:

- None until mid April 2009
- June 8 or 9, June 15 to about the 20th, 2009
- July 17-20, around 27, 2009
- Day or 2 around Oct 13, 14, 2009
- April 10-15 or so? 2011
- one or two days late August, 2011. (no mail)
- end of march first half of April 2012
- August 27-30 2012
- low first week of September, 2012
- 2 weeks end of october, up to Nov 10 approx 2012
- Dec 7- 17? 2012, Again about Dec 25 2012 to Jan 1, 2013

Contentious issues:

- private email server
- classified documents (outside of state)
- Sidney Blumenthal
- Benghazi spin handling of media (Susan Rice)
- scrubbing of her server (bleachbit)
- missing email from online email State department

Learn:

- inner circle over time (state or not)
- spike in email around Libyan revolution
- gap in the email
- daily email patterns, server behaviour (daylight savings time)

Filter:

- time
- redacted or not
- FOIA exemption codes

Content:

- Term frequency, TFIDF

Tool:

- Web-based, interactive filter and display tool

Discovery from visualization & connecting discovery sources

- Nothing on this ... Preparation for Benghazi (security considerations)

- House Oversight and Government Reform Committee (standing committee) Darrell Issa, Chair Jason Chaffetz, Chair (Gowdy a member) (discovered email server) ... - House Select Committee on Benghazi (Summer 2014) Trey Gowdy, Chair

7 Concluding Remarks

In providing the service, we purposely focused the displays (for the most part) on simple characteristics of the emails, the so-called metadata. There are several related reasons for this focus.

First, metadata is often more reliable, regular, and, being easily collected, more generally available. For Secretary Clinton's emails, some metadata is lost, either because their source is a printed form or because it has been redacted. Examining metadata also arguably intrudes least upon the privacy of the individual correspondents, compared at least to the email's content.

Second, since at least the Snowden revelations (e.g. [37]), public discourse has grown on the potential value of metadata to those who have it. By providing a web service where anyone can see for themselves what might be learned from metadata, we hope to contribute positively to this important discussion. Moreover, the work-related only metadata of a public figure as senior as Secretary Clinton will hopefully resonate more strongly with other public figures engaged in the debate (e.g. [12, 31, 33, 34, 47]) than might that of an ordinary citizen.

Finally, it is important that users realize that the value of metadata is not just in itself, where one might expect to easily discern general day-to-day habits such as one's circle of correspondents. Rather it is that when coupled with other sources, which are abundantly and publicly available for Senator Clinton, much more can be learned than from any one source alone. Those who have knowledge of, or access to, other sources may filter metadata to test previously held hypothesis; conversely, exploration of metadata could uncover patterns that generated hypotheses to be tested elsewhere.

Acknowledgments

This work was supported in part by a Discovery grant from the Natural Sciences and Engineering Research Council of Canada..

REFERENCES

[1] S. Attkisson. Congress to probe security flaws for Libya diplomats. *CBS Evening News*, October 18, 2012.

[2] S. Attkisson. Hillary Clinton's email: the definitive timeline. *Sharyl Attkisson: Untouchable Subjects, Fearless Reporting (online)*, November 2016.

[3] Automatic. clintonemail.com. *WHOis.net*, March 2017.

[4] A. Bloomfield. Egypt protests: troops and tanks ordered out in bid to quell protests. *The Telegraph*, January 28, 2011.

[5] M. Bouchet-Valat. *SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library*. CRAN R-project, 2014.

[6] CBS News. "Face the Nation" transcripts, September 16, 2012: Libyan Pres. Magariaf, Amb. Rice and Sen. McCain. *Face the Nation*, September 16, 2017, 2012.

[7] CBS News. Libya consulate: Was security added or taken away? *CBS This Morning*, October 5, 2012.

[8] T. G. (Chairman), L. Westmoreland, J. Jordan, P. Roskam, M. Pompeo, M. Roby, S. Brooks, E. E. Cimmings, A. Smith, A. Schiff, L. Sanchez, and T. Duckworth. *U.S. House of Representatives Select Committee On the Events Surrounding the 2012 Terrorist Attack in Benghazi*, volume 114-848. U.S. Government Publishing Office, Washington, D.C., December 2016.

[9] J. Chia. The brains behind Clinton's email 'cover-up': How top aide decided which messages were deleted, sat in on her FBI interview and is set to follow her to the White House. *Mail Online*, September 4, September 2016.

[10] H. Clinton. The secretary's call with egyptian pm kandil. In *U.S. House of Representatives Select Committee On the Events Surrounding the 2012 Terrorist Attack in Benghazi*, volume September 12. U.S. Government Publishing Office, 2012.

[11] H. R. Clinton. Secretary Clinton emails. In *Virtual Reading Room*. U.S. Department of State, 2017.

[12] D. Cole. 'We Kill People Based on Metadata'. *The New York Review of Books*, May 10, 2014.

[13] N. Confessore and M. S. Schmidt. Clinton friend's memos on libya draw scrutiny to politics and business. *The New York Times*, May 18, 2015.

[14] U. S. Congress. *The Freedom of Information Act, 5 U.S.C. Sect 522, As amended by Public Law No. 104-231, 110 STAT. 3048*, volume PUBLIC LAW NO. 104-231, 110 STAT. 3048. U.S. Department of Justice, 1996.

[15] I. Feinerer and K. Hornik. *tm: Text Mining Package*. CRAN R-project, 2017.

[16] I. Feinerer, K. Hornik, and D. Meyer. Text mining infrastructure in r. *Journal of Statistical Software*, 25(5):1–54, 2008.

[17] A. Gearan and C. Lynch. U.S. Ambassador Susan Rice. *The Washington Post*, October 15, 2012.

[18] J. Gerstein and N. Gass. Top Clinton aide Cheryl Mills granted partial immunity in email investigation. *Politico*, September 23, 2016.

[19] A. Goldman and M. S. Schmidt. Justice dept. granted immunity to specialist who deleted Hillary Clinton's emails. *The New York Times*, September 8, September 2016.

[20] J. Griffin. Cia operators were denied request for help during benghazi attack, sources say. *Fox News Politics*, October 26, 2012.

[21] N. Jones. The next FOIA fight: The B(5) "withold it because you want to" exemption. *UNREDACTED: The National Security Archive*, 2014.

[22] G. Kessler. Hillary Clinton's e-mails: a timeline of actions and regulations. *The Washington Post*, March 10, March 2015.

[23] E. Kiely. Benghazi Timeline: The long road from "spontaneous protest" to premeditated terrorist attack. *FactCheck*, June 29, 2016.

[24] M. Landler and E. Lichtblau. F.B.I. Director James Comey recommends no charges for Hillary Clinton on email. *The New York Times*, July 5, 2016.

[25] D. T. Lang and the CRAN team. *RCurl: General Network (HTTP/FTP/...) Client Interface for R*. CRAN R-project, 2016.

[26] D. T. Lang and the CRAN Team. *XML: Tools for Parsing and Generating XML Within R and S-Plus*. CRAN R-project, 2016.

[27] D. Leigh and I. Griffiths. The mystery of Tony Blair's finances. *The Guardian*, December 1, 2009.

[28] J. Linshi. What to know about the Hillary Clinton email controversy. *Time*, August 2015.

[29] C. McGreal. Benghazi attack testimony claims state department ignored warnings. *The Guardian*, October 10, 2012.

[30] R. Mendick. Blair Inc: How Tony Blair makes his fortune. *The Telegraph*,

January 7, 2012.
[31] B. Obama. Statement by the President on the Section 215 Bulk Metadata Program. *The White House: Office of the Press Secretary*, March 27, 2014.
[32] R. O'Harrow Jr. How Clinton's email scandal took root. *The Washington Post*, March 27, March 2016.
[33] E. O'Keefe. Bush credits Obama for continuing NSA's metadata program. *The Washington Post*, April 21, 2015.
[34] M. Pompeo and D. B. Rivkin Jr. Time for a rigorous national debate about surveillance: Post 9/11 measures have been weaked or discarded, a coherent new approach is needed,. *The Wall Street Journal*, January 3, 2016.
[35] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.
[36] C. Ross. The immunized five: Meet the people covering for Hillary. *The Daily Caller*, September 23, 2016.
[37] A. Rusbridger. The Snowden Leaks and the Public. *The New York Review of Books*, November 21, 2013.
[38] M. S. Schmidt. A closer look at Hillary Clinton's emails on Benghazi. *The New York Times*, May 21, 2015.
[39] M. S. Schmidt. House Bengazi Committee requests Hillary Clinton email server. *The New York Times*, March 20, 2015.
[40] M. S. Schmidt. No copies of Clinton emails on server, lawyer says. *The New York Times*, March 27, March 2015.
[41] M. S. Schmidt and M. Apuzzo. Hillary Clinton emails said to contain classified data. *The New York Times*, July 24, July 2015.
[42] M. T. Schmidt. Hillary Clinton used personal email account at State Dept., possibly breaking rules. *The New York Times*, March 2, March 2015.
[43] J. Sullivan. tick tock on Libya https://wikileaks.org/clinton-emails/emailid/23898. In *Hillary Clinton Email Archive*. WikiLeaks, 2017.
[44] Telegraph Staff. Egypt protests: World leaders call on Egypt to address its citizens' grievances. *The Telegraph*, January 28, 2011.
[45] P. Thompson. Clinton email investigation timeline. *Thompson Timeline*, November 2016.
[46] U.S. Department of State. Hillary rodham clinton's travels as secretary of state. *United States Office of the Historian*, 2016.
[47] K. Waddell. Trump's CIA Director Wants to Return to a Pre-Snowden World. *The Atlantic*, November 18, 2016.
[48] H. Wickham. *stringr: Simple, Consistent Wrappers for Common String Operations*. CRAN R-project.
[49] Wikileaks. Hillary Clinton email archive. *https://wikileaks.org/clinton-emails/*, March 2017.
[50] Wikipedia Contributors. Battle of Tripoli (2011). *Wikipedia, The Free Encyclopedia*, 2017 Web.
[51] Wikipedia Contributors. Hillary Clinton email controversy. *Wikipedia, The Free Encyclopedia*, March 2017 Web.
[52] Wikipedia Contributors. Timeline of the 2011 Libyan Civil War. *Wikipedia, The Free Encyclopedia*, 2017 Web.
[53] Wikipedia Contributors. Timeline of the investigation into the 2012 Benghazi attack. *Wikipedia, The Free Encyclopedia*, March, 2017 Web.
[54] A. Yuhas. Hillary Clinton: alleged classified emails simply 'disagreement between agencies'. *The Guardian*, August 19, August 2015.
[55] A. Yuhas, S. Siddiqui, B. Jacobs, and S. Ackerman. FBI has found no criminal wrongdoing in new Clinton emails, says Comey. *The Guardian*, November 7, November 2016.