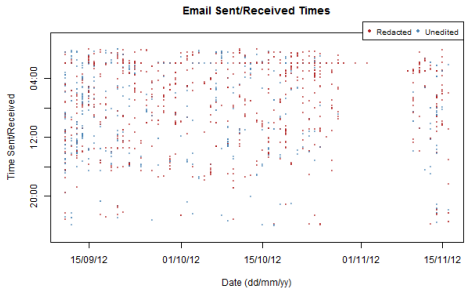
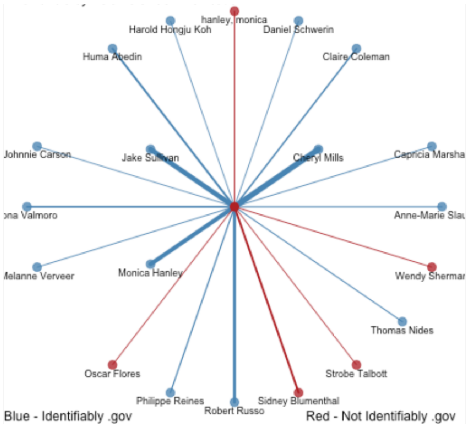
Interactive Filter and Display of Hillary Clinton's Emails: A Cautionary Tale of Metadata

Christopher D. Salahub and R. Wayne Oldford



(a) Missing data in November.



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

Fig. 1. Metadata: Time, date, correspondents, volume. Filtered: Sept. 11 to Nov. 15, 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 [20] containing the contents of more than 30,000 emails (and attachments) that were sent to and from Secretary Clinton's private server. The documents were provided as pdfs by the U.S. Department of State in response to Freedom of Information Act (FOIA) [6] requests. The State Dept. also provided a searchable web archive of the pdf documents, released in several installments from May 2015 to March 2017 [5]. 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, however, is the ability to easily conduct statistical analyses of their contents. Only then can

- *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

patterns be easily discerned.

In this paper, we describe a complementary website service (`rshiny.math.uwaterloo.ca/clinton`). This service does not provide the search facilities found on Wikileaks, but rather provides the user with interactive visualizations that allow them to uncover patterns in the emails and, through a variety of filters on the emails, to explore these patterns at some depth. The service is intended to be used in conjunction with either the Wikileaks service or the State Dept. FOIA service. The latter two provide the search mechanisms as well as the actual contents of individual emails. One imagines having three browsers open, one for the data base search (e.g. Wikileaks), one for statistical visualization and exploration (our service), and one for internet search (to provide important context for the other two services). Each one fills a need not met by the other two; the synergy of all three provide the user a powerful set of investigative tools.

In providing the service, we have purposely restrict consideration (for the most part) to displaying patterns of simple characteristics of the emails, displaying simple characteristics of the emails, the so-called metadata.

TODO MORE TO COME HERE

Meta data ... Cautionary

1.1 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 in the whole affair. More complete and in depth timelines are readily available on the internet (e.g. [1, 7, 10, 12, 19]).

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 [2]; 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 [18] during the course of a U.S. Congressional investigation [3] 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 [13] 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 [10]. 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 [4, 19].

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 [15], but was informed March 27 by David Kendall that no emails remained on the private server for any kind of review [16]. 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) [8,9,14].

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 [17], 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" [21].

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 [11]. 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 [22]. In both cases, Comey and the FBI are criticized by pundits from different political parties.

2 OLD STUFF

TODO This material is to be reused wherever All work related emails were eventually released to the State Department who subsequently published more than 30,000 on the U.S. To add to the confusion, only those work government emails, as vetted Secretary Clinton's lawyers

The emails and the server became a source for criticism of Secretary Clinton, and others, and came to dog her bid for the presidency in 2016, in what may be one of the most contentious U.S. elections in history. Part of this criticism was levelled More complete timelines are publicly available on the internet (e.g. [1, 7, 12]). For our purposes, it is enoug

33,000 supposedly missing emails Encouraged and three other American citizens. Public knowledge of the private email server surfaces State business being conducted on a private server surfaced only after the a

The 2016 US Presidential election was one of the most contentious in history. The existence and possible content of Hillary Clinton's private email server played a significant role in the campaign until the end.

The timeline of events related to use of this private server may be found elsewhere (e.g. Despite their significant nature, very few of the individuals commenting on the significance of this server had spent any significant time viewing this content. This is no fault of theirs, the officially released data has previously only been stored in databases on the United States' State Department Website (see https://foia.state.gov/Learn/New.aspx) and Wikileaks (see https://wikileaks.org/clinton-emails/) in individually searchable form. Furthermore, the email data is stored in the inconvenient format of individual PDF documents, and in the case of Wikileaks a slightly more convenient but still cumbersome HTML representation based on the official PDF content. This granularity and separation of individual emails on different webpages prevents a great deal interesting analysis of the data, including any aggregate analyses of patterns in content or metadata. As such, the emails are incredibly misunderstood documents, and current searches for patterns and points of interest have relied upon searching for terms of interest and sifting through the results of the search. Such a method is not only tedious but problematic from a statistical point of view, as approaching any data set with a specific hypothesis before exploring its structure and patterns in abstract leads to biased conclusions and poor analysis. In the interest of simplifying the general exploration of this important data set, the Wikileaks HTML email versions were extracted and a series of visualizations were generated around the metadata included in these emails. An interactive web application in `R shiny` was constructed utilizing these visualizations and the extracted data. Incorporated in this app was a short article explaining the use of the app, providing links to background information, and outlining some of the findings of the more interesting trends observed. This framework provides any individual with interest the ability to generate hypotheses by exploring the data without necessarily searching for any particular conclusion.

In generating and exploring this data, a considerable amount of respect is gained for the power of the often overlooked metadata of emails, especially with public figures. Simply having access to the network of communications, simple indicators of identifiable features of the individuals communicating, and times sent is enough to generate interesting questions which simple Google searches of important events and days can shed a great deal of light on. Leveraging data beyond this metadata, including the Freedom of Information Act (FOIA) exemption codes (see https://vault.fbi.gov/explanation-of-exemptions) used

to redact the emails, only bolsters this simple metadata driven investigation. Metadata is often viewed as less significant or somehow less intrusive data, but the exploration performed in this paper demonstrated, at the very least to the authors, what a powerful tool metadata is for generating hypotheses about data. Given the lack of privacy present in the modern internet age and the amount of information most individuals make publicly available, there is a clear warning here about the utility of metadata for bad or for good. For a public official like Hillary Clinton, the data become even more interesting due to a number of contentious issues which arose during her tenure as the Secretary of State, independent of and as a result of her use of a private server.

3 Data

TODO This needs to be short and to the point. Just say what is done. Before any of the analysis could begin, the emails had to be converted from the individual PDF and HTML form into a more easily used data structure. The choice was made to use the Wikileaks HTML emails as the source for data extraction. This choice was made primarily due to the relative ease of loading the raw HTML pages on the Wikileaks database due to their regular method of storage (the email with ID ¡id¿ is stored at https://wikileaks.org/clinton-emails/emailid/¡id¿) and ease of loading. These qualities made the HTML emails amenable to programmatic extraction, which was completed using a host of web-scraping and string parsing tools in `R`. In particular the packages `RCurl` and `XML` proved invaluable to extract the raw HTML pages for each email, and the packages `tm`, `stringr`, and `SnowballC` were indispensible to process and clean the raw HTML into a useable form. 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.

Unfortunately, this data is not perfect. The PDF files released by the state department do not include the typical email server header with information about the time stamp inforamtion, the addresses of all involved, and other useful and regular information. Instead, these files appear more similar to screen captures of the emails, providing only the visual display a user would see without the useful information used to generate that display. As a result, the email addresses are occasionally not included, and there were initially concerns over the time zones used to generate the time stamp metadata. Furthermore, the HTML data on Wikileaks is frequently irregular in form, and although work has been done to make the cleaning and extraction functions as general as possible, there are still certainly fringe cases not considered which occur frequently enough to have an impact. It is also important to keep in mind the fact that the data have already been filtered twice. First by the emails chosen to be released by Clinton, and next by the emails and email sections which the State Department chose not to redact. Indeed, even the metadata can be affected by this redaction, as a number of individuals have their email addresses redacted in all communications, preventing certain analysis from being performed.

4 Displays

TODO Each of these four displays is described and commented on in context of whole time line

4.1 Inner circle

The spokes are separated into two groups by the radial length of the spokes, these groups are determined automatically by the largest difference in volume of correspondence present in the data between the ordered counts of number of emails sent between individuals. As well, the width of the edges of the graph is determined by the volume of correspondence which occurred between Clinton and the correspondent in either direction. Finally, all edges and points are coloured according to a simple scheme. If the name associated with a node in the graph is at any point associated with an email address hosted at some '.gov' domain in the correspondences, they are coloured blue. Individuals with email addresses which are identifiably not hosted at a '.gov' domain are coloured red. Finally, individuals which cannot have their email host domain identified are coloured grey. It is important to note that for those individuals with grey spokes, the reason the email is not identifiable is the removal of the email by the state department. Such a redaction is not a guarantee that the email said individual used is not an official state email, but it is an indication that the email address is viewed as personal or private information, which suggests some email hosted outside of the '.gov' domain. Still, such speculation cannot be conclusively made, and so these emails are simply reported as they are, unidentifiable.

4.2 Email volume

This plot has two lines, the first is a line showing the total volume of emails for each day, and the second shows the number of emails satisfying the classification filter criteria for a given day. This allows users to view which days have an exceptionally high or low number of classified conversations. Returning to the motivation of hypothesis generation, referencing world events on any days of interest provides a number of interesting insights on how public figures like Clinton react and respond to world events.

4.3 Email times

This plot takes advantage of the time-stamps on every email and then plots the emails as points with times on the vertical axis and the day sent on the horizontal axis. The points in this scatterplot are then coloured according to whether the corresponding email was redacted, in which case the point is coloured red, or released in full, in which case the point is coloured blue. Alpha blending and transparency of these points was utilized to make the patterns present more obvious and prevent overplotting from obscuring any interesting results.

4.4 FOIA exemptions

. It is important to notice that this barplot is not trivialized by the possibility of selecting individual FOIA exemption codes, as many of the emails in the data set are exempted on a number of grounds, and so have numerous FOIA exemption codes. Thus, this barplot is not only interesting when all codes are selected, but can be used to see which codes commonly co-occur in documents.

4.5 Term frequency and TFIDF

5 Filters

TODO Each filter is described and displayed

5.1 Time filtering

5.2 Sent or received

5.3 FOIA exemptions

5.4 Auxiliary information

TODO e.g. Clinton's foreign schedule

6 Some further exploration

TODO Some of our findings. Unless done in the previous section

7 Shiny App

TODO Again short and to the point The app (HOST ADDRESS) provides the ability for users to place custom filters on the time frame of server time stamps and classification tags of the emails and view how a series of three displays change.

Finally, a barplot of the FOIA exemption codes used in the selected data is displayed

Historically important, possibly had an impact on the 2016 US presidential election.

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

8 Exposition

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

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

[3] 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.

[4] 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.

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

[6] 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.

[7] Crowdsource. Hillary Clinton email controversy. *Wikipedia*, March 2017.

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

[9] 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.

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

[11] 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.

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

[13] R. O'Harrow Jr. How Clinton's email scandal took root. *The Washington Post*, March 27, March 2016.

[14] C. Ross. The immunized five: Meet the people covering for Hillary. *The Daily Caller*, September 23, 2016.

[15] M. S. Schmidt. House Bengazi Committee requests Hillary Clinton email server. *The New York Times*, March 20, 2015.

[16] M. S. Schmidt. No copies of Clinton emails on server, lawyer says. *The New York Times*, March 27, March 2015.

[17] M. S. Schmidt and M. Apuzzo. Hillary Clinton emails said to contain classified data. *The New York Times*, July 24, July 2015.

[18] M. T. Schmidt. Hillary Clinton used personal email account at State Dept., possibly breaking rules. *The New York Times*, March 2, March 2015.

[19] P. Thompson. Clinton email investigation timeline. *Thompson Timeline*, November 2016.

[20] Wikileaks. Hillary Clinton email archive. *https://wikileaks.org/clinton-emails/*, March 2017.

[21] A. Yuhas. Hillary Clinton: alleged classified emails simply 'disagreement between agencies'. *The Guardian*, August 19, August 2015.

[22] 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.