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

Christopher D. Salahub & R. Wayne Oldford
University of Waterloo
Waterloo, Ontario, Canada

Abstract

We present a web-based visualization that allows the user to interactively filter and display characteristics of 30,322 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.

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 an significant role in the campaign until the end.

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

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