Visualization of Email Data in R

We developed a web-based application using shiny in R which allows a user to interactively explore an email data set. The particular data set chosen was the official United States State Department release of emails from Hillary Clinton's server sent during her tenure as the United States Secretary of State, due to this data set's relevance and importance in the recent United States presidential election. Emphasis was placed on the metadata of these emails, both due to its relative cleanliness and as a demonstration of the power of such simple data when leveraged with other information.

The data was extracted from HTML representations of the official PDF releases of emails on Wikileaks using the "getURL" function from the package "RCurl." This raw HTML was then processed using regular expressions as supported in the R packages "tm," "stringr," and "snowballC."

Four visualizations are used to present the data. The first of these is a purpose-built spoked network plot with Clinton as a central node and other nodes corresponding to individuals she communicated with. The edge joining the two nodes has a length inversely proportional to the square root of the volume and a width directly proportional to the volume of correspondence between Clinton and the individual. Edges and nodes are coloured by the email domain. Next, a plot of the daily volume of emails sent is displayed, with both the total and selected volume indicated. Third is a scatterplot of the time of day of each individual email against the date coloured by whether the email has been redacted or not, which allows for the user to view emailing habits and their patterns. The final plot displays redaction codes present in the emails, applied in the official release by the State Department to exempt certain email data from release.

All of these displays update in response to a series of filters which the user can control. These include filters on time, redaction codes, and whether the email was sent by Clinton or to Clinton. Together, all of these tools allow the user to explore a vast number of possible hypotheses and explore this controversial data set.

To provide context to such exploration, a brief guide is also provided in a panel in the application. Provided here are examples of some analyses we performed, which focused on conspicuous gaps in the emails, the peak email day, and the general pattern of email sending times. Such guided analyses are meant to inspire and entice users to look for their own hypotheses to test and examine.

This application is currently public, and it is hoped that it will not only allow interested individuals to explore this significant collection of documents, but will expand the users' perspectives on the utility of metadata. Interesting results can be found using only our application and quick Google searches, and the ease with which public information can be leveraged by a simple tool on nondescript data will hopefully compel the user to think about how simple data about them and others is used.