Capstone Project 1 – modelling FCC Net Neutrality comments for fake or real

Milestone Report

The Client: The Public

The FCC has voted to repeal Net Neutrality laws. When open to the public there were 22 million submissions, many apparently in favor of repealing these laws. Existing analysis documented by the press has already revealed that many of these are fake. The public should now just how many of these are fake – estimates exceed 2 million.

Data Set: 22 million records

I contacted Jeff Kao, author of “More than a Million Pro-Repeal Net Neutrality Comments were Likely Faked”, and he provided me with the SQL database of the 22 million public comments. From these I took a sample of 3 million records.

To identify the zip codes I split out the data from the address dictionary fields (domestic addresses were stored differently than international addresses). I matched the zipcodes with their respective US State or Washington DC. I associated each state with the census population estimates for 2016

I cleaned up the time stamps to remove the micro seconds.

To identify fake comments there were certain patterns.

1. If contact email was in ALL CAPS, the domestic street address for the city and zip code did not exist.
2. If multiple comments had been filed in the exact same second (‘000s) – their name typically did not match the name on the email, and the comment was identical to the 5,000+ others filed in that exact same second.
3. Legitimate filings had empty indicators showing that the fields were blank. If a field had a Null value that indicated it was machine generated.
4. Some domestic addresses had an IP number instead of a zip code
5. Fake comments with international addresses were filled in with US Zip codes but consistently missing the city and have “United States of America” written in.
6. There were contact email accounts with non existing companies : hurra.de, or pornhub.com 200,000 employees all with Russian names.

I took a sample of 400 emails and classified them as fake or not.

I aggregated and created a column to show duplicate time stamp and duplicate emails, and identified likely fake US addresses in the international addresses field.

Other Data Sets: Broadband access, Political leanings

My initial study was to see if the proportions of comments allegedly from each of the 51 areas (DC + 50 states) would be consistent with the population in each area. It was not.

Also further analysis could inspect the text comments and see if comments are coming from places that do not even had broadband access, or if the proportion of Repeal comments are consistent with the proportion of Republicans in that area.

Initial Findings: Many are clearly fake

The distribution of comments across the country are not consistent with the proportion of the population. It appears to be possible to analyze several aspects of the comments to determine their authenticity simply by exploring how the comments were submitted and what is omitted or Null.