The Federal Communications Commission’s public comment process for Net Neutrality was deliberately manipulated using false identities

*Startup Policy Lab preliminary report estimates that 3.5 to 7.5 million pro-repeal net neutrality comments were submitted using stolen identities.*

By Charles Belle, Gina Cooper, Jeffery Kao, Sarah Rigdon

**LEARN MORE ABOUT US AT**
startuppolicylab.org

This is a preliminary report. Final data and calculations will be made publicly available in the final report.
About Startup Policy Lab

Startup Policy Lab works at the intersection of technology and government. Our mission is to promote innovative, open, and data-driven policymaking. Our work seeks to inform, educate, and empower those who craft public policy to support those goals. Startup Policy Lab, a fiscally sponsored project of Community Initiatives – a 501c3 nonprofit organization.
Summary

Startup Policy Lab (SPL) does not engage in political advocacy and has no position on Net Neutrality. SPL generates independent research that is publicly accessible to all. Our goal is to inform better data-driven policymaking, not to achieve a particular political outcome.

For democracy to function properly, all citizens must trust that their voice is being heard and counted. Advances in technology have made it easier for citizens to participate, but have also changed the way malicious actors can affect that participation.

The Truth in Public Comments (TiPC) Project examines the concerns related to the authenticity of public comments submitted to the Federal Communications Commission (FCC) as part of the Restoring Internet Freedom (Docket FCC-17-108) proceeding.

The research and lessons from TiPC, however, are not limited to the FCC. These insights apply to any governmental agency (Federal, State, or Local) that receives public comments. Our goal is to provide data to inform development of tools and practices that nurture trust and confidence in our democratic processes for all citizens.

![Emails Sent](chart.png)

- **Bounces**: 25k
- **Completes**: 26k
- **(People who said that they did not submit comments)**

**Emails Sent**

- **77k** Batch 1
- **77k** Batch 2
- **77k** Batch 3
- **74k** Batch 4
- **74k** Batch 5
- **74k** Batch 6
The data presented in this report was generated from a survey of over 450,000 randomly selected emails associated with the FCC comments, sent in 6 batches of approximately 75,000 each. Based on the 14,000 responses, the following two findings emerged:

- The discrepancy in the proportion of submissions using stolen identities between pro-repeal and non-repeal respondents raises concerns about unknown actors manipulating FCC public comment systems to influence public policy for unknown purposes.
  - 88% of survey respondents whose emails were used to submit pro-repeal comments likely had their identity appropriated to submit a comment, compared to just 4% of pro-net neutrality respondents.

- Pro-repeal spam campaigns appeared far more sophisticated, which indicates a greater level of intent, resources, and preparation.
  - Bounce rates were low across pro-repeal campaigns, despite strong evidence of a high proportion of fake comments.

This report indicates that there was a sophisticated and deliberate campaign to manufacture popular opinion, which disrupted democratic discourse.

There was a deliberate attempt to interfere with the FCC’s public comment process. And this type of organized cyber-aggression employed an understanding of how information flows through government. It should be treated as evidence of threats to the systems of democracy that impacts all citizens.

Experts in technology and the law have been encouraging the FCC to examine this additional data attached to each public comment, while the New York Attorney General has also submitted multiple requests to access the data to pursue evidence for prosecution.

More information is needed to understand the nature of the threat in order to develop safeguards. The issue would benefit from further study by a range of stakeholders (SPL, government, and others) to implement legal and technical processes that prevent malicious efforts to undermine citizen confidence in democratic processes.

An expanded study is possible because the breach and evidence of this is on government servers. This, actually, is cause for hope. The data logs and other information needed to further investigate this fraud against the FCC is in the hands of the federal government. For example, IP address information can be used to identify ISPs hosting spammers and hone in on the locations of spam serving computers.
Nonetheless, the manipulation of the FCC’s public comment process indicates that the questions plaguing the FCC are more canary in the coal mine, than final chapter. **This is because the intended target of deception is not the public, but government itself.** The purpose of public comments is to create opportunities for citizens and stakeholders to participate in the rulemaking process. Whether government officials take the veracity of these comments at face value or they choose to ignore them all because they’ve lost faith in the process of surfacing authentic sentiment, citizens lose because government does not have the information it needs from the public to make an informed decision.

The challenges faced by the FCC are systemic and apply to government agencies at all levels (Federal, State, and Local). Accordingly, SPL applauds government initiatives to adopt new technologies that make information accessible to citizens and society,¹ but it is critical to be thoughtful about how those new systems can be abused.

**The substantial number of reported fake submissions raises concerns about the ability of unknown actors to manipulate current FCC public comment systems.**

*SPL’s TiPC Project* conducted the survey by sending emails to commenters to the FCC’s Restoring Internet Freedom (Docket FCC-17-108). The survey asked recipients whether they actually submitted the comment attributed to them in the record.

Approximately 16% of all respondents reported that they did not submit the comment attributed to them. This implies that someone, or some entity, submitted a public comment to the FCC using the individual's name without their permission. A stark discrepancy emerges. Of the respondents that indicated they did not make the comment, 88% of survey respondents whose emails were used to submit pro-repeal comments likely had their identity appropriated to submit a comment. Conversely, just 4% of respondents whose emails were associated with pro-net neutrality comments said that they did not submit the comment attributed to them.

---

The figure below provides a detailed breakdown of the aggregate findings for each Public Campaign from which the SPL survey received at least 20 responses. Altogether, these comments comprise approximately 8.6 million pro-repeal public comments in the FCC record. Applying the 88% of likely stolen identities indicates that 7.5 million pro-repeal comments were filed using an appropriated identity.
SPL adopted a conservative approach to reflect assumption noted in the Methodology section for the number of pro-repeal emails associated with pro-net neutrality responses that used a stolen identity.

**Pro-repeal spam campaigns appeared far more sophisticated, which may represent a greater level of intent, resources, and preparation**

Instances of spam comment campaigns from both sides of the net neutrality debate have been well reported. But evidence in these survey results can be interpreted to understand the qualitative differences between pro-repeal and pro-net neutrality spam campaigns.

Across commenters surveyed, SPL found that 39% have an invalid email address, implying that about 8,500,000 comments were submitted with an invalid email address.
Breaking down the bounce rates of each individual campaign, it is apparent that a high number of pro-net neutrality Public Campaigns submitted comments using invalid email addresses.

SPL sent 450,000 emails over five days and 14,000 responses were collected. The response rate, however, varied greatly between pro-net neutrality and pro-repeal, with pro-net neutrality having a much higher response rate.
Data suggests sophisticated and deliberate actions were taken to mask stolen identities for public comments in support of pro-repeal

Spam campaigns were conducted on both sides of the net neutrality issue, but it is surprising to see that all of the pro-repeal campaigns have a low bounce rate, even though our findings suggest that many of the identities used to submit comments under these campaigns are illegitimate.

The large discrepancy in the bounce rates between the two groups suggests a sophisticated effort to manipulate the public comment process and avoid detection. By appropriating the information of real people with functioning emails, these actors could disguise their intention of manipulating public comment data by avoiding the first line of detection – bouncing emails. SPL is open to alternative interpretations of these remarkable results, but the data strongly indicates that a large-scale effort to steal and use identities is the most likely explanation.

If the sophistication of the spam campaigns were equal on both sides, one would expect to see at least one pro-repeal campaign that had similarly high bounce rates. But the sophistication of the spam comments on the pro-repeal campaigns indicates a higher level of coordination and intent to conceal their true origins.

**Batman vs. Stolen Identities**
The FCC received spam comments that supported both the pro-net neutrality and pro-repeal. The difference, however, is that the majority of spam comments associated with email addresses supporting pro-net neutrality were ignored by the FCC because they were obviously
fake. Conversely, we must conclude that the spam comments associated with email addresses that supported pro-repeal email addresses were a deliberate campaign to evade the eyes of regulators and influence the rulemaking process.

The discrepancy rests in the nature of the bounceback of emails. The survey resulted in a high bounce rate for emails associated with pro-net neutrality using unsophisticated approaches. Examples of an unsophisticated spam comment are those the FCC acknowledged are, “[o]bviously, fake comments [...] by the Flash, Batman, Wonder Woman, Aquaman, and Superman are not going to dramatically impact our deliberations on this issue.”

By contrast, it appears that the spam comments for emails associated with pro-repeal comments reflect deliberate action to use stolen identities. In these instances, millions of Americans may have had their identity harvested for the political objectives of supporting the repeal of net neutrality laws, regardless of whether that individual agreed with the position or even had a position on the proposal. Accordingly, unlike the submission from Batman, which the FCC was correct to ignore, millions of Americans had their voice taken and repurposed without their consent.

Methodology

This population surveyed includes all of the public comments submitted online to the Federal Communications Commission (FCC) in response to the proposal titled Restoring Internet Freedom (Docket FCC-17-108). The data used in this study was submitted using the Electronic Comment Filing System (ECFS) the FCC uses to receive public comments submitted online. All of the information on the FCC’s website is made accessible to the public.

Public comments included in this report were submitted between the period of April 27, 2017 and October 27, 2017. The original docket for public comment was opened on April 27, 2017 and scheduled to close on August 16, 2017. The FCC, however, extended the comment period to August 30, 2017.

Public comments timestamped between August 31 and October 27 are not part of the portion the FCC will officially consider, but they are a part of the official public record for this rulemaking.

In addition to the ECFS, the FCC accepts public comments by phone and letter. Comments submitted by phone and letter, however, are not publicly accessible and were excluded from the study.

The data was downloaded via the FCC API over two-days: October 27-28. The comment scraping script, however, suffered from a couple of disconnections an estimated 50,000 comments were lost as a result. Although the public comment period ended on August 30, 2017, the FCC’s ECFS system continued to take comments afterward and those comments were included in the analysis.

The following data was collected from the FCC website via the API.

- ID (FCC’s internal measure)
- Proceeding
- Name of Filer
- Type of Filing (comment, other)
- Filing Status (i.e., disseminated)
- Viewing Status (i.e., Sunshine, unrestricted, etc.)
- Date Received
- Date Posted
- Email address
- Address
- City
- State
- ZIP
- Brief Comment

The data collected from the FCC API was kept in its original state.

Comment Clustering Process

Once collected, the public comments were grouped into clusters. Clusters are comprised of public comments with exact or substantially similar text. For the purposes of this report, clusters are referred to as Public Campaigns. Clustering the public comments into Public Campaigns relied on the early work by Jeffery Kao.\footnote{Jeffery Kao, More than a Million Pro-Repeal Net Neutrality Comments were Likely Faked, (11/23/17). https://hackermoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6} Kao identified groups of comments with exact or substantially similar text.

To identify duplicate comments, Kao applied a hash function to find and aggregate duplicates. Submissions without any express comment text were removed. No other text preprocessing was applied before encoding and clustering in order to preserve artifacts in the text that may give clues as to the method of submission.
Removing duplicate comments identified 2,955,182 unique comments. These comments were mapped into semantic space vectors and several clustering algorithms were applied to the meaning of the comments.

Two different approaches were made at clustering the document vectors. HAC and DBSCAN with a euclidean distance metric at a very low epsilon were used to identify obvious clusters, which were then culled out manually using a string signature. This left approximately 2 million unique comments.

From the 2.9M unique comments, HDBSCAN was applied on a 100,000 comment sample with euclidean distance metric between l2-normalized doc vectors to identify ‘looser’ clusters; then approximate_predict() was used to classify remaining comments as either within those identified clusters or as outliers.5

This method generated nearly 150 clusters of comment submission texts of various sizes.

The top 100 clusters (of the 150 total clusters) constitute the Public Campaigns as defined above. Of these 100 Public Campaigns, the largest had 7,542,184 comments and the smallest 2,920 comments. And the top 20 comment Public Campaigns accounted for 17,000,000 of the more than 22,000,000 submissions.

How was the survey conducted?

Following guidance from technical experts at QuestionPro, the survey was conducted and a z-score applied, similar to the way pollsters determine population proportion with random sampling.

A random sample of public comments was taken from each of the 100 Public Campaigns (described above). Only Public Campaigns having at least 12,000 comments were included. This reduced the 100 Public Campaigns to 42 that were available for use in the survey. One additional unique Public Campaign was included, for a total of 43 Public Campaigns.

4 Id.

A large proportion of [the approximately] 3 million “unique” comments were essentially duplicates — only differing by a few characters or words or having a different signature. In order to conclusively and exhaustively categorize these comments, I chose to group comments by meaning. Comments were turned into document vectors comprised of the average of all word vectors in the comment. The word vectors were obtained from spaCy, which uses the word vectors from the paper by Levy and Goldberg (2014). [Correction from Matthew Honnibal; spaCy now uses the GloVe vectors by Pennington et al.]

5 Id. After clustering comment categories and removing duplicates, I found that less than 800,000 of the 22M+ comments submitted to the FCC (3-4%) could be considered truly unique. (Emphasis by author)
How did were individuals contacted?

Using QuestionPro, SPL generated six blocks. Each block consisted of approximately 75,000 emails for a total of 456,288 emails.

An email was sent to each block that informed the recipients of the research project, asked recipients one survey question, and provided the first 255 characters of the comment submitted to the FCC associated with that email address.

**Survey question:** *Did you submit the comment quoted below to the FCC, yes or no?*

Out of the six blocks, Block 2 suffered from technical difficulties. Block 2 had formatting issues and some recipients received the same email several times.

The data from Block 2 was kept (and used) based on the assumption that the technical difficulties would only reduce the response rate, but not change the veracity of the survey responses.

<table>
<thead>
<tr>
<th></th>
<th>Block 0</th>
<th>Block 1</th>
<th>Block 2</th>
<th>Block 3</th>
<th>Block 4</th>
<th>Block 5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Date</strong></td>
<td>12/07/2017</td>
<td>12/08/2017</td>
<td>12/8/17</td>
<td>12/8/17</td>
<td>12/09/2017</td>
<td>12/09/2017</td>
</tr>
<tr>
<td><strong>Number of emails</strong></td>
<td>77,363</td>
<td>77,301</td>
<td>77,295</td>
<td>74,749</td>
<td>74,818</td>
<td>74,762</td>
</tr>
</tbody>
</table>

The data collected from recipients included
- Responses to an emailed survey question
- Time of a response
- Bounce rate
- Data related to the recipients’ interaction with the email: open rates and whether the recipient chose to opt-in to receive report results.\(^6\)

---

\(^6\) SPL offered recipients an optional contact form to fill out if they wanted to keep updated on the report itself.
Can you please confirm that you submitted this comment?

- Yes, I submitted the comment above to the FCC
- No, I did not submit the comment above to the FCC

If you would like to provide additional details about your research project or your FCC submission experience, you can click here to fill out a comment form.

Thank you.

The Truth in Public Comments research team at Startup Policy Lab

---

How will my information be used?

We are planning to publicly release aggregate statistics. Your individual responses and any comments you voluntarily submit will remain confidential to Startup Policy Lab, discussed only in our research round-up (as long as they include related to the project) and shared when the project is complete. As part of our research, we may follow up on comments you submit.

How was I selected to receive this email?

Your information was selected randomly from the FCC’s Electronic Comment Filing System (ECFS), which is part of the public record. Your responses and information will only be used once, for this research project, and we will not pass your personal information to any reports or final findings.

Read about how the FCC handles information submitted via the ECFS at https://www.fcc.gov/ecfs and click here comment. You can also search the FCC’s proceedings at https://www.fcc.gov/receivable-proceedings.

How do I learn more?

Learn more about Startup Policy Lab at https://www.startuppolicylab.org. Startup Policy Lab is a nonprofit independent think tank working at the intersection of technology and government. Our mission is to promote innovative, open, and data-driven policymaking.

Starting Policy Lab is a fiscally sponsored of Community Initiatives – a 501c3 nonprofit organization – located at 1604 Market Street in San Francisco, California.

FCC Public Comments in the News

The Washington Post reports that individuals’ identities have been appropriated without their consent or knowledge in public comments on the FCC website: https://www.washingtonpost.com/business/2017/06/15/psc-alpha-and-happy-workday-impersonate-those-people-who-are-filing-the-fcc-in-proprietary-name/

Notice: Please note that this message is not a commercial message and is intended to verify the authenticity of comments posted on the FCC site.

Data SPL collected:

- Did you submit this comment to the FCC? clicked yes/clicked no
- Timestamp of click
- The email failed to deliver: true/false
- Any direct replies to the email (tipc.research@startuppolicylab.org via mail.questionpro.com)
Stolen Identity Estimate

The following table provides specific numbers. There are five fields, from left to right,

- Campaign ID assigned by SPL for the purpose of organization and analysis
- Percentage of respondents indicating they did not submit a public comment to the FCC
- Percentage of emails that bounced back
- Total number of public comments in the Public Campaign
- Total estimated number of stolen identities in the Public Campaign

<table>
<thead>
<tr>
<th>campaign</th>
<th>percent_not_original_commenter</th>
<th>percent_send_failed</th>
<th>num_comments_total</th>
<th>est_stolen_ids</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>0.9577464789</td>
<td>0.0200901607</td>
<td>100800</td>
<td>47300</td>
</tr>
<tr>
<td>16</td>
<td>0.9509803922</td>
<td>0.0190765017</td>
<td>293295</td>
<td>136798</td>
</tr>
<tr>
<td>33</td>
<td>0.92</td>
<td>0.0266170858</td>
<td>105586</td>
<td>47276</td>
</tr>
<tr>
<td>3</td>
<td>0.8985507246</td>
<td>0.0230234153</td>
<td>1371979</td>
<td>602204</td>
</tr>
<tr>
<td>37</td>
<td>0.8874172185</td>
<td>0.0083787997</td>
<td>13744</td>
<td>6047</td>
</tr>
<tr>
<td>4</td>
<td>0.8852459016</td>
<td>0.0123227384</td>
<td>818146</td>
<td>357667</td>
</tr>
<tr>
<td>6</td>
<td>0.8783783784</td>
<td>0.0223004695</td>
<td>1282764</td>
<td>550812</td>
</tr>
<tr>
<td>7</td>
<td>0.8783783784</td>
<td>0.0216858012</td>
<td>532493</td>
<td>228793</td>
</tr>
<tr>
<td>31</td>
<td>0.8773584906</td>
<td>0.0261150235</td>
<td>92886</td>
<td>39683</td>
</tr>
<tr>
<td>53</td>
<td>0.8673469388</td>
<td>0.0339282212</td>
<td>83079</td>
<td>34806</td>
</tr>
<tr>
<td>11</td>
<td>0.8671328671</td>
<td>0.1018409714</td>
<td>317603</td>
<td>123678</td>
</tr>
<tr>
<td>32</td>
<td>0.8571428571</td>
<td>0.0304096998</td>
<td>127511</td>
<td>52985</td>
</tr>
<tr>
<td>9</td>
<td>0.8552631579</td>
<td>0.0295412306</td>
<td>410098</td>
<td>170190</td>
</tr>
<tr>
<td>5</td>
<td>0.8529411765</td>
<td>0.0220810943</td>
<td>957467</td>
<td>399315</td>
</tr>
<tr>
<td>2</td>
<td>0.8461538462</td>
<td>0.0309540084</td>
<td>1384869</td>
<td>567769</td>
</tr>
<tr>
<td>8</td>
<td>0.8153846154</td>
<td>0.1995509128</td>
<td>649299</td>
<td>211890</td>
</tr>
<tr>
<td>55</td>
<td>0.7666666667</td>
<td>0.0124644224</td>
<td>62637</td>
<td>23711</td>
</tr>
</tbody>
</table>

Total: 3600924
**Limited timeframe**

The FCC is voting on Docket FCC-17-108 on December 14, 2017. Our goal was to provide the FCC with independent data before that date regarding potential manipulation of the public comment process. This deadline limited the window of conducting the survey, collecting data, and providing an analysis to less seven days.

**Assumptions**

Given the following assumptions, SPL is disclosing as much as possible regarding the process and noting the assumption in our conclusions.

For the purposes of this study, it was assumed that the people who actually responded at the bottom of the decision tree are representative of the people who did not respond. This is a substantial assumption that SPL does not take lightly. Consequently, SPL treated individuals for whom no bounce back was received as a representative sample of those whose email were real.

By assuming that the people who didn't respond are representative of those who did, this approach effectively eliminate one of the branches of the decision tree, making this a simple Bernoulli distribution.
Send email

- Bounce Back
- Invalid email

Valid Email

- Didn’t Respond
  - Didn’t write comment
  - Wrote comment

- Responded