TPInsights

Tuesday, December 5, 2017

Public Comments in the World of Massively Multiplayer Regulatory Proceedings

Economics Staff

Scott Wallsten, PhD
(202) 828-4405
swallsten@techpolicyinstitute.org

Robert Hahn, PhD
(202) 828-4405
rhahn@techpolicyinstitute.org

Thomas Lenard, PhD
(202) 828-4405
tlenard@techpolicyinstitute.org

Sarah Oh, JD, PhD
(202) 425-7725
soh@techpolicyinstitute.org

Lindsay Poss, MS
(202)-828-4405
lposs@techpolicyinstitute.org

Nathaniel Lovin
(202) 828-4405
nlovin@techpolicyinstitute.org

For Press Inquiries
David Fish
(571) 389-4446
dfish@techpolicyinstitute.org

Technology Policy Institute
409 12th Street, SW
Suite 700
Washington, D.C. 20024

By the time the FCC's ferociously controversial net neutrality draft Order was released on November 22, 2017, more than 22 million comments were submitted to the Commission through its new application programming interface (API). This avalanche of public input is impossible to navigate and interpret using human labor alone. Machine learning tools are uniquely suited to navigating and interpreting such a large amount of information. Their use, however, implies a new set of problems and rules of engagement for regulatory proceedings in a digital world.

When considering this new world, we should keep a couple of facts in mind:

Human judgment still matters. While machine learning tools are necessary to read and interpret such massive amounts of information, the algorithms still must be trained by humans. Analyses that use machine learning must make the judgments and assumptions transparent.

Algorithms must be targeted at solving the problems created by the large amounts of data. Those will generally include the ability to identify fraudulent submissions as they become easier to submit and to provide useful interpretations of the submissions that the agency can address. This final point is likely the biggest challenge in using machine learning to analyze such massive amounts of data.

This post explains how APIs have encouraged public participation in regulatory rulemakings while simultaneously making it more complicated for agencies to absorb the information presented. Additionally, we ran queries on the 22 million comments to investigate fraudulent submissions and applied textual analysis to a random sample of 220,000 comments.

APIs Are a Big Deal

Open APIs, which make it possible for third parties to connect directly with the underlying computing system, have reduced the transaction costs of filing comments in regulatory proceedings. The reduced costs associated with filing comments has two implications. The first is the potential flood of comments agencies may receive as advocacy groups can more easily encourage like-minded people to submit legitimate comments via their websites. We found that 90 percent of comments in a sample of the data could be identified based on 25 unique phrases, implying that they were sent as part of a form letter campaign. A report by Emprata also estimated that more than 90 percent of the 22 million comments appear to be generated by clickable forms on third-party websites.[1]

Computerized submission tools also make it possible to submit comments based on forms that are difficult to detect as forms. For example, some groups made it possible to select combinations of text as if they were playing games of

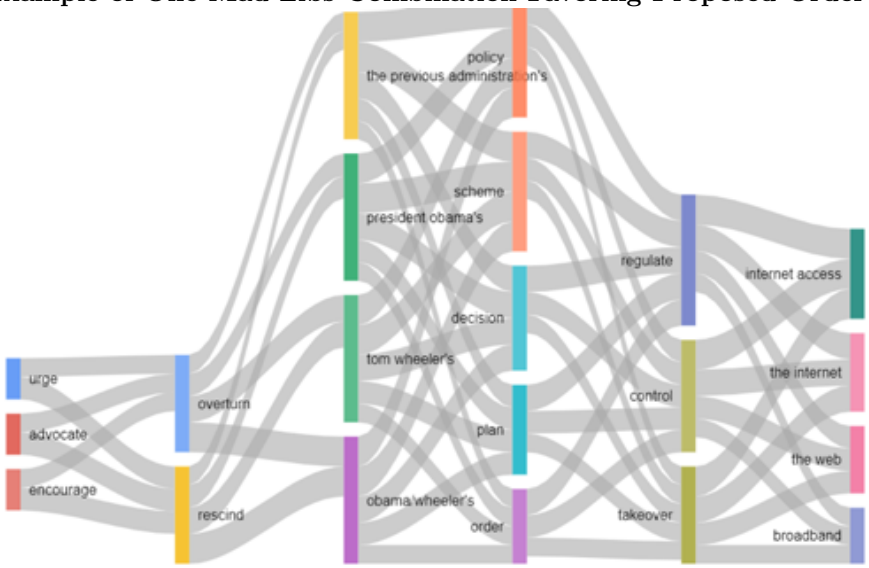
[1] https://www.emprata.com/reports/fcc-restoring-internet-freedom-docket/.

"Mad Libs."

One sentence, for example, favoring the Order was filled with rotating keywords so that a single sentence became more than 1440 combinations of text, all of which support the Order:[2]

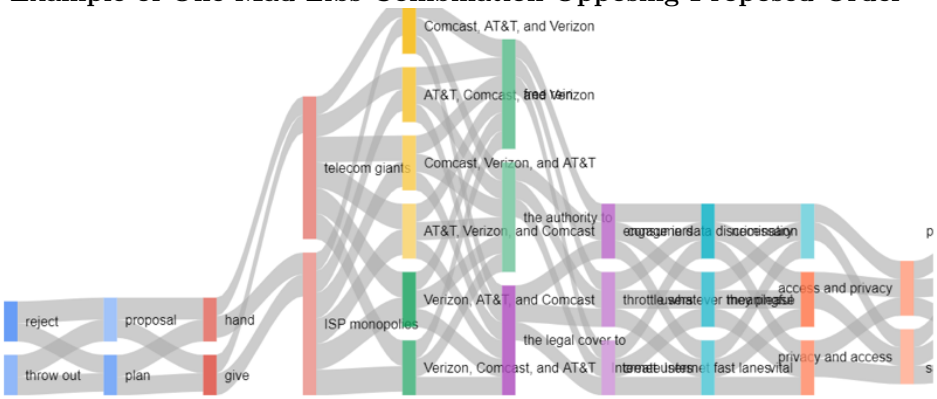
"I (urge) / (advocate) / (encourage) you to (overturn) / (rescind) (the previous administration's) / (president obama's) / (tom wheelers's) / (obama/wheelers's) (order) / (plan) / (decision) / (policy) / (scheme) to (takeover) / (control) / (regulate) (broadband) / (the web) / (the internet) / (internet access)."

Example of One Mad Libs Combination Favoring Proposed Order



Another, even more complicated, Mad Libs opposed the Order, as shown in the figure below.

Example of One Mad Libs Combination Opposing Proposed Order



If these Mad Libs comments appear in the same proportion in the entire population of comments as they do in our sample, then these fill-in-the-blank forms would be responsible for over four percent of comments in the proceeding. A quick search on all 22 million comments shows 325,498 instances of the Mad Libs form letter that opposes the order. Other data scientists have identified these variations of form letter as well[3][4].

Just as with other submissions based on form letters, the Mad Libs comments are not necessarily fraudulent if a real person chose which words to put in each blank and signed their true name to the submission. However, APIs have also made it easier to submit fraudulent comments.

In particular, comments submitted under made-up names and misused identities are always fraudulent. To identify likely candidates for such submissions, we queried the full database of 22 million comments for email domains com-

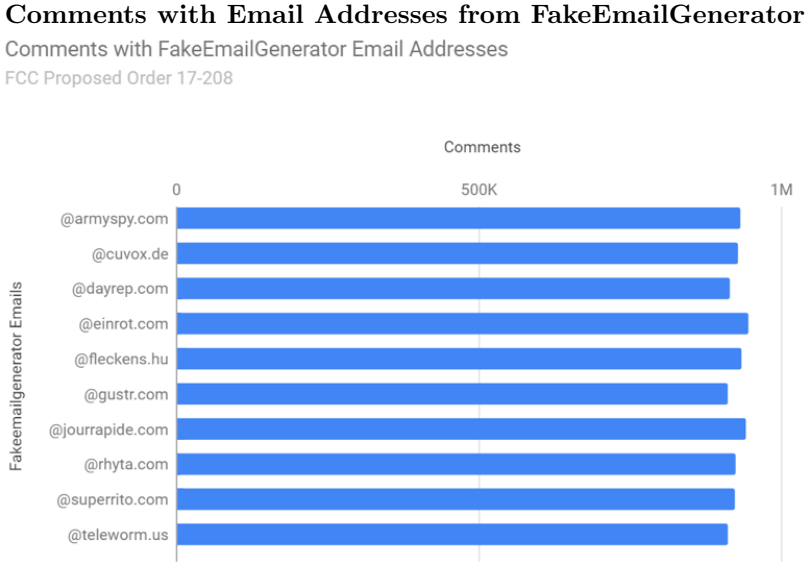
[2] $1,440 = 3 * 2 * 4 * 5 * 3 * 4$ keywords.

[3] https://hackernoon.com/more-than-a-million-pro-repeal-net-neutrality-comments-were-likely-faked-e9f0e3ed36a6.

[4] https://www.wired.com/story/bots-broke-fcc-public-comment-system/ with 419,904 combinations of text from $2 * 2 * 2 * 2 * 6 * 3 * 3 * 3 * 3 * 2 * 3 * 3 * 3$ keywords.

monly identified as spam generated from fakeemailgenerator.com. The figure below shows the number of comments with email addresses from these domains. Whoever submitted comments with the ten domains listed below did so rather systematically. Each email domain was represented in approximately 900,000 comments.

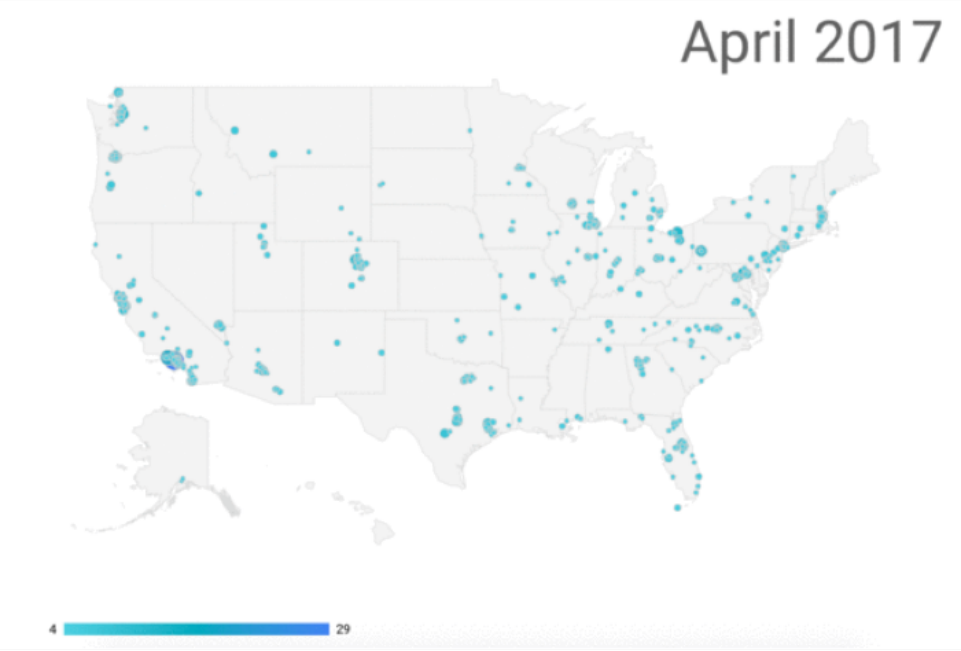
Comments with Email Addresses from FakeEmailGenerator



Consistent with the hypothesis that fraudulent comments came from these domains, their time stamps indicate that they tended to be submitted in large batches more or less simultaneously.

This analysis suggests that 9 million, or 40 percent, of the 22 million submissions came from these domains and are likely to be fraudulent.[5] We tracked the locations of the addresses from these comments, and show their geographic origin over time in the following figure.

Location of Comments with FakeEmailGenerator Email Addresses



Note: Purple dots represent comments with email addresses likely generated by fakeemailgenerator.com. Blue dots represent all comments submitted over the last few months.

[5] https://www.bloomberg.com/news/articles/2017-11-29/fake-views-444-938-russian-emails-among-suspect-comments-to-fcc/.

The comments appear to be distributed across cities without any particular geographic pattern, though the time loop shows the increase in the number of comments from fakeemailgenerator.com in July and August. From our cursory investigation, geolocation alone does not appear to tip off the status of comments as fraudulent or not. Machine learning models will eventually be needed to determine if mailing addresses are useful to verify the likelihood of authentic comments.

What are the Commenters Saying?

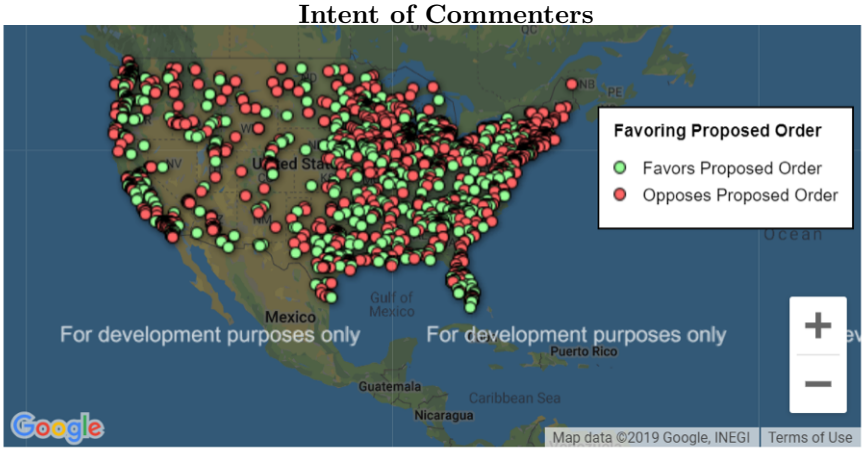
Identifying fraudulent comments, grouping (legitimate) comments submitted by forms, and eliminating duplicate submissions are relatively straightforward tasks, at least conceptually, if not necessarily in practice. Interpreting the comments, however, is more complicated.

We begin the interpretation task by classifying a random sample of 220,000 comments (i.e., about one percent of the total). We generated this sample to start thinking about how to train the algorithm for additional analyses on all 22 million comments and not to draw conclusions, per se. However, analyzing a representative sample is one way to study the overall population, just as statistically representative samples of the nation yield valid information on the makeup of the whole. Thus, even though the sample is intended for training the algorithm for future research into how machine learning can be applied to analyze regulatory proceedings, the results of the analysis of this sample are also likely to reflect the same analyses done on all 22 million comments.

While regulatory proceedings are not supposed to aggregate societal preferences in the same way that elections or legislative bodies do, a basic question is how many of the comments favor the Order and how many oppose it. A machine learning algorithm does not inherently know how to identify the difference between a comment supporting the net neutrality Order and one opposing it. It must first be trained and calibrated.

We manually classified all 220,000 comments as favoring or opposing the Order, identifying phrases associated with each side.[6] This manual labor makes it possible for a machine learning tool to identify all 22 million comments as favoring or opposing the order.

Intent of Commenters

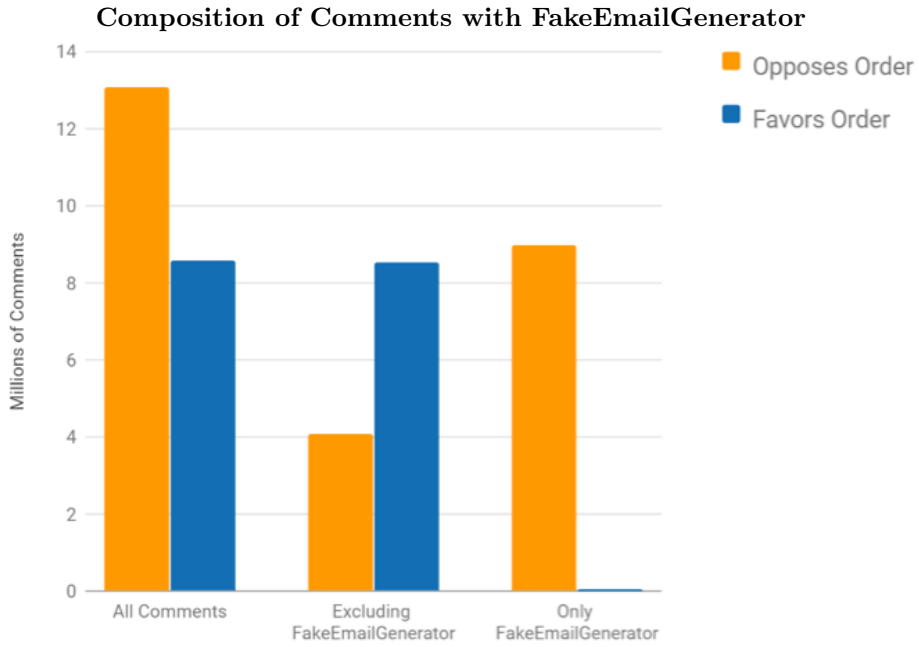


Note: Each dot represents a single comment.

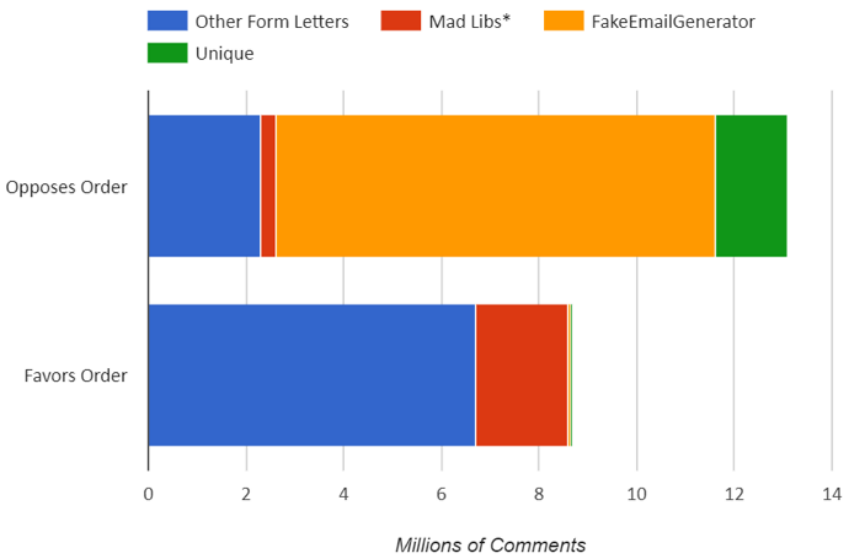
Our categorization approach generally yielded results similar to the Emprata report. Across all comments, including those likely to be fraudulent, we find about 61 percent opposed to the proposed Order and 39 percent in favor of it.

Removing the comments created by FakeEmailGenerator changes the pro/con ratio. Because more than 95 percent of the comments from FakeEmailGenerator opposed the order, removing them yields 68 percent favoring the order and 32 percent opposed.

[6] The Emprata report trained a machine learning model based on a few manually classified phrases (Emprata Report, p. 23).



While the emails from FakeEmailGenerator were almost all opposed to the Order, form letters and Mad Libs generally favored the Order, as did the unique comments.



Manual assignment or classification of comments raises important issues of human judgment. The Emprata researchers claimed an accuracy of 99.8% accuracy in classifying the intent of 95% of the 22 million comments. The last one percent of comments that did not fall into their classification criteria would require closer scrutiny.[7] We made these human judgments on the comment text.[8]The methodological differences between the Emprata report and our analysis highlights that there is not one correct way to analyze the comments. Using multiple approaches can help increase the confidence of the results if they are similar.

How Strongly are the Commenters Saying It?

[7]The Pew Institute conducted a similar analysis and found a larger proportion of unique comments, at 6%: http://www.pewinternet.org/2017/11/29/public-comments-to-the-federal-communications-commission-about-net-neutrality-contain-many-inaccuracies-and-duplicates/.

[8]Our methodology is available here: http://fiscly.com/22_Million_Comments.

An agency may want to know more than whether a commenter is for or against a rule. One of the simpler analyses one can do is identify and acknowledge comments that express strong preferences, rather than weak, indifferent, offensive, or irrelevant preferences.

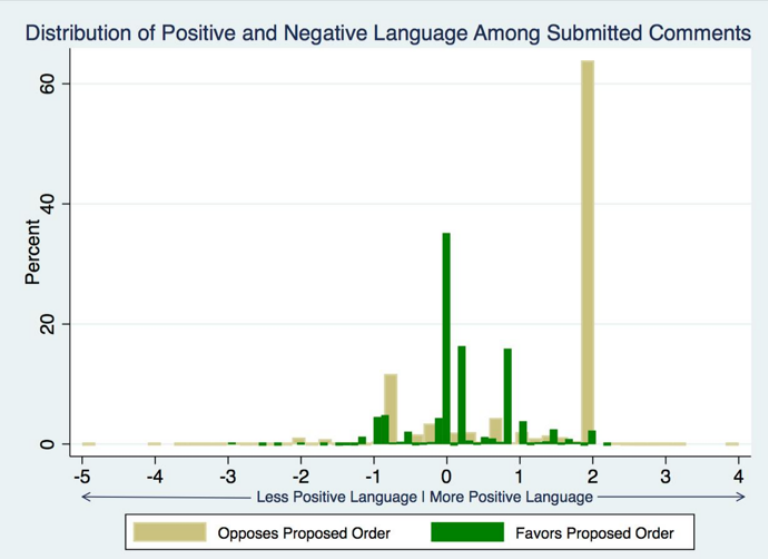
"Sentiment analysis" is a common use of machine learning. We used an off-the-shelf list of words categorized from strongly negative to strongly positive[9] to score each word in the text of each comment. Then, we took the average of these scores to create a comment score.[10]

Intensity of Preferences

This approach has some disadvantages. It does not take into account sentence construction. It cannot identify a negative sentence constructed with "positive" words. Adverbs can yield anomalies: "Breathtakingly stupid" would yield a net positive score because "breathtakingly" is more positive than "stupid" is negative. Nevertheless, a significant amount of spot checking suggests that our approach seems to yield reasonable results on the emotive strength of comments. Much textual analysis done today uses these standard scoring methods. We note, however, that assigning a sentiment score to a comment implies a more accurate level of categorization than is possible for something that is largely normative.

The figure below shows the distribution of the type of language used in comments for and against the proposed order. The figure shows that comments favoring the Order tend to be more neutral, clustered around zero. The majority of comments opposing the Order tend to be positive, with longer tails on both negative and positive sides of the distribution.

Distribution of Sentiments



[9] Words like "superb" and "breathtaking" yield the highest score, while the harshest swear words, like "prick" receive the lowest score.
[10] We use a list developed by Finn Årup Nielsen, described here: http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010.

Conclusion

Machine learning tools can give us insights into massive numbers of comments that would be impossible otherwise. As the use of APIs and bots increases,[11]government agencies must be able to deploy such technology to deal with them and stay true to the Administrative Procedure Act.[12]Nevertheless, human judgment still matters, at least in terms of setting the starting points of machine learning algorithms. Depending on the cost, it may be worth using different training sets to determine how sensitive these results are to human judgment underlying the models.

To explore interactive maps of our 220,000 comment sample, we have made available our data notes at .

Recent TPInsights

Econometrics in the Cloud: Robust Standard Errors in BigQuery ML (Dec 10, 2019)

Econometrics in the Cloud: Extending Google BigQuery ML (Nov 6, 2019)

Economics, Experts, and Federalism in *Mozilla v. FCC* (Oct 4, 2019)

The Law and Economics of *Apple Inc. v. Pepper* (Dec. 20, 2018)

Stay tuned for more economic and legal analysis from Washington, D.C. in ***TPInsights***. Contact Ashley Benjamin at (202) 828-4405 for more information

Technology Policy Institute

The Technology Policy Institute is a non-profit research and educational organization that focuses on the economics of innovation, technological change, and related regulation. More information is available at www.techpolicyinstitute.org.

[12]http://www.realclearpolicy.com/articles/2017/07/07/bots_go_to_washington_110290.html.
[12]http://www.realclearpolicy.com/articles/2017/09/14/the_fcc_should_embrace_artificial_intelligence_110355.html.