



August 30, 2017

VIA ELECTRONIC FILING (ECFS)

Marlene H. Dortch
Secretary
Federal Communications Commission
445 12th Street SW
Washington, D.C. 20554

Re: *In the Matter of Restoring Internet Freedom*
WC Docket No. 17-108

Ms. Dortch:

Broadband for America¹ respectfully files this letter to highlight the results of an analysis of the unprecedented volume of comments in the Restoring Internet Freedom proceeding conducted by Emprata,² an independent expert data analytics firm. Emprata reviewed 21.766 million publicly available comments in the docket and found that more than 20.684 million comments (95 percent) can be attributed to email domains that appear to be artificial, international filings, form letters, and duplicative submissions. As stated in the attached Fact Sheet summarizing key findings of the analysis (Attachment A), once filings from artificial email domains attributed to FakeMailGenerator.com and unverifiable international comments are accounted for, 69.9 percent of comments are in favor of repealing Title II. The full report is also attached.

Please contact me with any questions.

Sincerely,

/s/ Bryan N. Tramont
General Counsel
Broadband for America

¹ Broadband for America is dedicated to protecting a free and open internet for all Americans. We support and promote bipartisan policies that align with the core principles of an open internet, including no blocking, no throttling, and no anticompetitive paid prioritization. Broadband for America believes that smart and modern broadband policies led by Congress that promote innovation and protect consumers are central to our country's economic prosperity. <http://www.broadbandforamerica.com/>

² www.emprata.com.

ATTACHMENT A



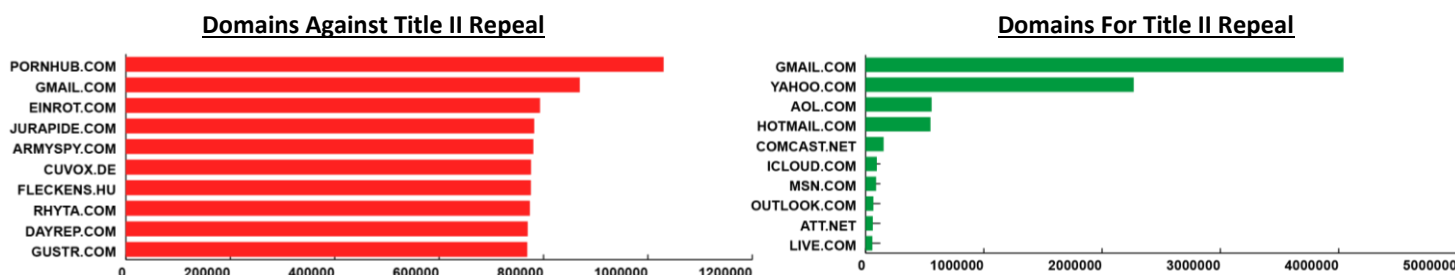
Unprecedented Volume and Clutter in FCC's Restoring Internet Freedom Docket

The Federal Communications Commission (FCC), policymakers, and the media face enormous challenges navigating sentiment about Title II regulations because of the unprecedented volume (over 21 million comments) and clutter in the electronic docket. Broadband for America commissioned an independent analysis based on publicly available data to take a comprehensive look at the comments.

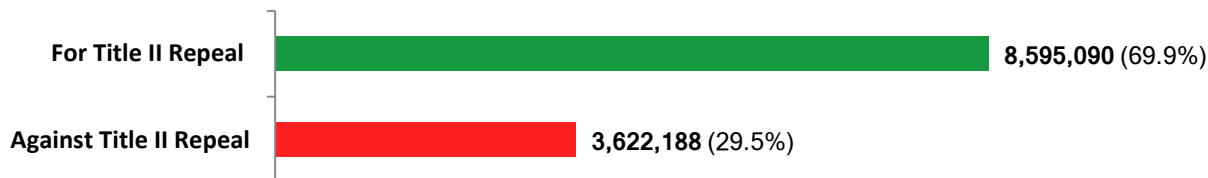
The analysis shows 69.9% of comments are in favor of repealing Title II once accounting for fake and unverifiable international comments.

KEY FINDINGS

- **Millions of Artificial Emails:** Over 7.75 million comments – the largest percentage of any set of comments, 36% of the total comments in the docket – were generated by fictitious email domains attributed to FakeMailGenerator.com.¹ One email domain boasts, “No legitimate email will ever be sent from einrot.com.”



- **Unverifiable International Comments:** 1.72 million unverifiable comments from international addresses in Russia, India, France, and 26 other countries flooded the docket.² Almost all of those comments (99.4%) oppose repealing Title II.
- **Duplicative Comments:** 9.93 million comments were filed using repeated email and/or physical addresses.³ There is rampant duplication in the docket as many people or organizations filed multiple comments.
- **Clear Majority Favors Repealing Title II:** When accounting for obviously fake email domains (7.75 million) and unverifiable international comments (1.72 million), the overall sentiment of comments is 69.9% in favor of repealing Title II.



CONGRESSIONAL ACTION NEEDED

The vast majority of the comments support open internet principles, but the analysis clearly shows there is strong concern about the damage done by Title II. This underscores the need for Congressional action. We stand ready to work with lawmakers on both sides of the aisle to ensure durable net neutrality principles are codified into law.

ABOUT THE ANALYSIS

Broadband for America commissioned Paul Salaszyk, Ph.D., CEO of Emprata, an expert analytics firm based in Washington, D.C., to perform an impartial, data-driven analysis of the comments. The methodology for determining the source and sentiment of comments is explained in detail on pg. 22 of the report available at www.emprata.com/fcc-restoring-internet-freedom-docket.

¹ Emprata FCC Restoring Internet Freedom Docket 17-108 Comment Analysis (pg. 17)

² Ibid. (pg. 12)

³ Ibid. (pg. 14)

ATTACHMENT B

FCC Restoring Internet Freedom Docket 17-108 Comments Analysis

August 30, 2017

Prepared by:



www.emprata.com

info@emprata.com

Table of Contents

1. Executive Summary.....	2
2. Background	5
3. Data Summary	6
3.1 Data Completeness.....	6
3.2 Comments Summary	7
3.3 Address Validation.....	9
4. Results.....	10
4.1 Considering All Comments.....	10
4.2 Duplicative Comments.....	14
4.3 Valid Addresses.....	15
4.4 Unique Comments	16
4.4 Email Domains.....	17
5. Conclusion.....	20
6. Disclaimer	21
APPENDIX A – METHODOLOGY	22

About Emprata

Emprata offers a variety of data analytics and software development services, each tailored to meet the specific needs of our clients. As our founding principle, we endeavor to unleash the power of data and analytics for strategic decision-making. With applications in several domains including health, finance, transportation, communications, and defense, Emprata’s portfolio of clients include those from both the public and private sectors.

1. Executive Summary

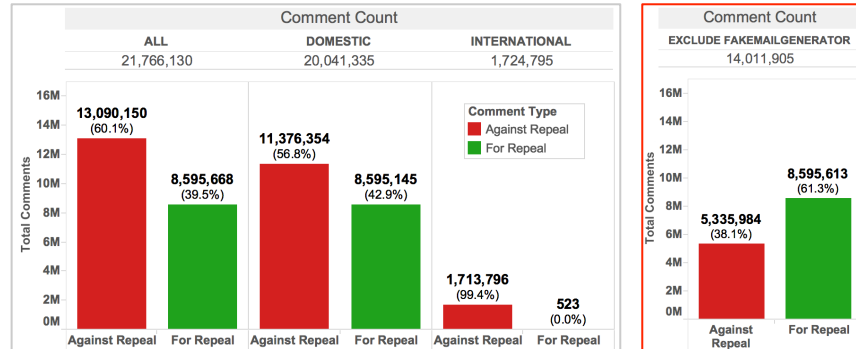
Emprata, LLC, a consulting firm specializing in advanced data analytics, accessed the full dataset of comments submitted to the Federal Communications Commission's (FCC) "Restoring Internet Freedom" docket and performed a comprehensive, independent analysis of the comment data. At the time of our analysis, there were 21.766 million comments in the docket. Our findings are summarized herein.

1. **Forming Conclusions:** The lack of user authentication by the Electronic Comments Filing System (ECFS) makes it difficult to determine "genuine" comment submissions. Emprata was also not able to authenticate the filer, address, email, or comment data used for this analysis, nor the methods used to collect those data elements. As a result, it is very difficult to draw any definitive conclusions from the comments found in the docket. Any conclusions that one might draw from the data would be based on the subset of data that they considered to be "real."
2. **All Comments:** Considering all 21.766 million comments at face value, general sentiment is against repeal of Title II (60% against, 39% for). General sentiment favors repeal of Title II when eliminating duplicative comments and comments from email domains that appear to be illegitimate.
3. **Form Letter Comments:** The overwhelming majority of comments for and against repealing Title II are form letters (pre-generated portions of text) that appear multiple times in the docket. The form letters likely originated from numerous sources organized by groups that were for or against the repeal of Title II. Form letters comprise upwards of 89.8% of comments against Title II repeal and upwards of 99.6% of the comments for Title II repeal.
4. **Unique Comments:** There are significantly more unique comments submitted against Title II repeal (1.77 million) versus for Title II repeal (24k). In addition, there are considerably more "personalized" comments (appearing only once in the docket) against repeal (1.52 million) versus 23k for repeal. Presumably, these comments originated from individuals that took the time to type a personalized comment. Although these comments represent less than 10% of the total, this is a notable difference.
5. **Data Completeness:** More than 81% of the total docket contained complete (i.e. usable) street address, city, state, ZIP code, and email information. 98% of comments in favor of the repeal of Title II contained usable data versus 70% of comments against the repeal of Title II. In addition, based on a 65% sampling of addresses, 84% of addresses for repeal of Title II were found to be valid versus 68% against repeal.
6. **Artificial Email Domains:** More than 7.75 million comments – the largest percentage of any set of comments (36% of the total comments) – appear to have been generated by self-described "temporary" and "disposable" email domains attributed to FakeMailGenerator.com and with nearly identical language. Virtually all of those comments oppose repealing Title II. Assuming that comments submitted from these email domains are illegitimate, sentiment favors repeal of Title II (61% for, 38% against).
7. **International Comments:** An unusually large volume of comments (1.72 million) are attributed to international addresses, which we did not verify. The vast majority of those comments (99.4%) oppose repealing Title II.
8. **Duplicative Comments:** 9.93 million comments were filed from submissions listing the same physical address and email, indicating that many entities filed multiple comments. This was more prevalent in comments against repeal of Title II (accounting for 82% of the total duplicates), with a majority of duplicate comments associated with email domains from FakeMailGenerator.com.

Comment sentiment has been determined using the methodology outlined in detail in the following report. The following illustrations provide a summary of the sentiment for several sub-groupings of the data, as well as a high-level comment summary, email domain summary, and geographic summary.

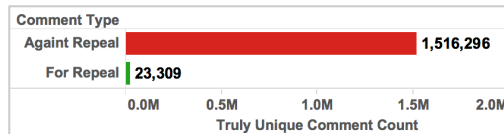
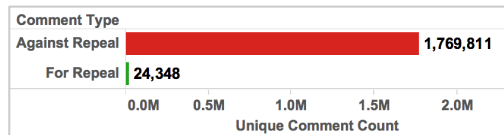
Entire Dataset

- **ALL** – Considers all comments submitted to the FCC
- **DOMESTIC** – Considers only domestic comments
- **INTERNATIONAL** – Considers only international comments
- **EXCLUDE FAKEMAILGENERATOR** – Excludes comments from 10 email domains associated with FakeMailGenerator.com



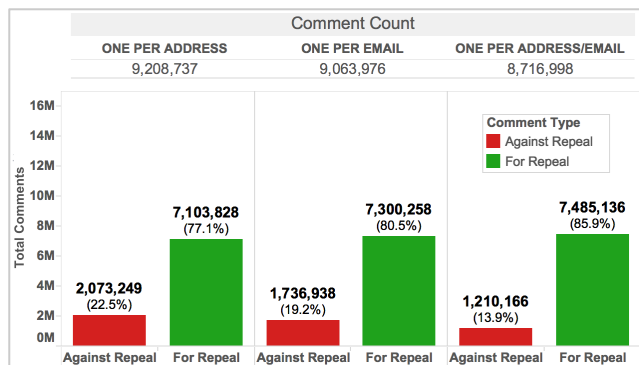
Unique Comments

- **UNIQUE COMMENTS** – Number of unique comments in the docket
- **TRULY UNIQUE COMMENTS** – Number of comments appearing only once in the docket



Eliminating Duplicative Comments

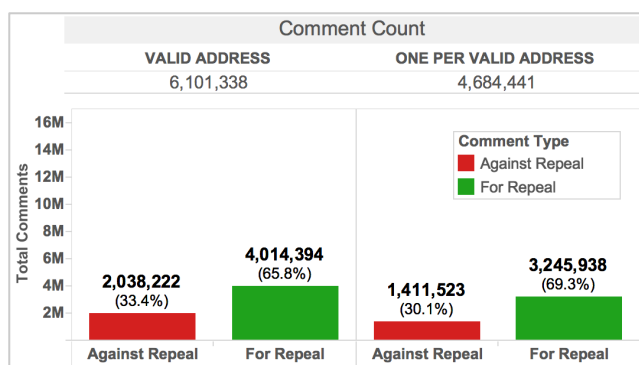
- **ONE PER ADDRESS** – Considers only the first comment for each unique address, city, state, ZIP code combination
- **ONE PER EMAIL** – Considers only the first comment from each email address
- **ONE PER ADDRESS/EMAIL** – Considers only the first comment from each unique address and email combination













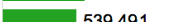
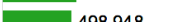









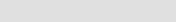


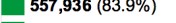
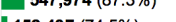
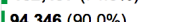
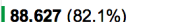
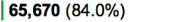
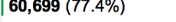
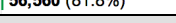
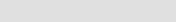
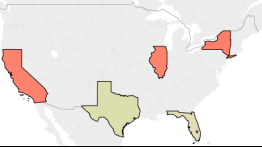
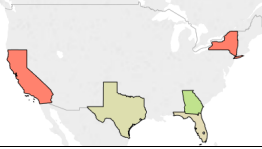
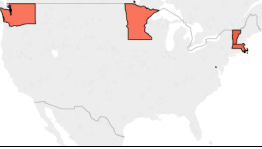
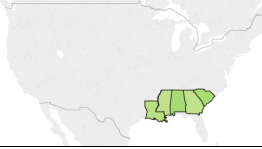




Considering only Valid Addresses

(Considers sample address data through August 4, 2017)

- **VALID ADDRESS** – Considers only comments where an exact address match was found
- **ONE PER VALID ADDRESS** – Considers only the first comment from each valid address



	AGAINST Title II Repeal	FOR Title II Repeal
Comment Summary		
Total	13,090,150 (60.1%)	8,595,668 (39.5%)
Avg. Per Day	129,605	85,106
Unique Comments	1,769,811	24,348
Appear 1x	1,516,296	23,309
Top 7 Days	August 7, 2017  1,261,759 July 24, 2017  1,225,074 July 19, 2017  901,598 July 17, 2017  831,276 July 12, 2017  815,064 July 18, 2017  665,996 July 31, 2017  657,108	August 11, 2017  864,792 August 7, 2017  750,949 August 15, 2017  650,070 August 4, 2017  600,948 July 12, 2017  594,157 May 26, 2017  539,491 May 30, 2017  498,948
Top 3 Comments	>> "I AM IN FAVOR OF STRONG NET NEUTRALITY UNDER TITLE II OF THE TELECOMMUNICATIONS ACT. SINCERELY, NAME" (7,569,190) >> "THE FCCS OPEN INTERNET RULES NET NEUTRALITY RULES ARE EXTREMELY IMPORTANT TO ME. I URGE YOU TO PROTECT THEM. (...)" (445,658) >> "I STRONGLY OPPOSE CHAIRMAN PAI'S PROPOSAL TO REVERSE NET NEUTRALITY PROTECTIONS BECAUSE A FREE AND OPEN INTERNET IS VITAL (...)" (340,205)	>> (VARIATIONS OF) "DEAR FCC, I AM CONCERNED ABOUT NET NEUTRALITY. I STRONGLY URGE THE FCC TO REPEAL (...)" (1,384,893) >> "IN 2015, CHAIRMAN TOM WHEELERS FEDERAL COMMUNICATIONS COMMISSION FCC IMPOSED RESTRICTIVE TITLE II, UTILITY-STYLE (...)" (1,083,324) >> "BEFORE LEAVING OFFICE, THE OBAMA ADMINISTRATION RAMMED THROUGH A MASSIVE SCHEME THAT GAVE THE FEDERAL GOVERNMENT BROAD REGULATORY (...)" (1,047,049)
Email Domain Summary		
Top 10 Domains (comment count – Sentiment %)	PORNHUB.COM  1,030,000 (100.0%) GMAIL.COM  869,467 (17.6%) EINROT.COM  793,133 (100.0%) JOURRAPIDE.COM  782,034 (100.0%) ARMYSPY.COM  780,509 (100.0%) CUVOX.DE  775,897 (100.0%) FLECKENS.HU  775,818 (100.0%) RHYTA.COM  773,743 (100.0%) DAYREP.COM  769,711 (100.0%) GUSTR.COM  768,833 (100.0%)	GMAIL.COM  4,039,880 (82.0%) YAHOO.COM  2,266,653 (83.9%) HOTMAIL.COM  557,936 (83.9%) AOL.COM  547,974 (87.3%) COMCAST.NET  152,437 (74.5%) ICLOUD.COM  94,346 (90.0%) MSN.COM  88,627 (82.1%) OUTLOOK.COM  65,670 (84.0%) ATT.NET  60,699 (77.4%) LIVE.COM  56,560 (81.8%)
Geographic Summary		
Address Match %	68%	84%
Top 5 States (comment count)	California (1,244,220) Texas (761,062) New York (702,899) Florida (558,356) Illinois (499,866) 	Texas (826,653) California (730,606) Florida (582,724) New York (395,731) Georgia (383,770) 
Top 5 States (%)	Wash DC (76.6%) Vermont (76.2%) Massachusetts (76.2%) Minnesota (68.2%) Washington (67.3%) 	Alabama (62.9%) South Carolina (62.2%) Louisiana (58.8%) Georgia (56.6%) Mississippi (56.5%) 
Top 5 Countries (comment count)	Russia (444,925) Germany (433,033) France (194,530) India (6,801) Canada (4,590) 	Canada (34) Australia (29) Germany (21) United Kingdom (18) Spain (12) 

2. Background

On April 27, 2017, the FCC, through a Notice of Proposed Rulemaking (WC Docket No.17-108), began a new proceeding on “Restoring Internet Freedom,” laying out a process to review existing regulations and consider repealing the Title II classification of Internet Service Providers (ISPs).

The FCC elicited feedback from the public during a comment period that began on April 27, 2017, and ran through August 30, 2017. Individuals wishing to express their opinion on this rulemaking were able to do so through the FCC’s Electronic Comment Filing System (ECFS), a web-based form. In addition, the FCC made an application programming interface (API) available to the public for those individuals or organizations wanting to submit comments from the public in bulk.

As of August 22, 2017, 21.766 million comments have been logged in the “Restoring Internet Freedom” proceeding. The analysis of such a large, unstructured dataset required the application of text mining and natural language processing techniques. More specifically, through the application of keyword searches and machine learning, measuring the volume of sentiment in favor of and against the proposed repeal of Title II in the full dataset was possible.

Emprata accessed the full dataset and performed a comprehensive, independent analysis of the data. We did not eliminate or discount comments that seemed artificially generated, duplicated, or submitted by actors who may have intended to influence the final sentiment tally. However, since there was an unusually large subset of comments that met aforementioned criteria, we categorized those comments in the dataset and in this analysis.

3. Data Summary

When submitting comments to ECFS, users had the opportunity to include their name, street address, city, state, ZIP code (including international address), email address, and a narrative text stating their position on the docket. This information is not validated by the FCC’s ECFS, meaning that the system accepts essentially all data submitted by users. Distinguishing between “genuine” comments and those comments submitted solely with the intent of influencing the final sentiment tally is difficult.

We were able to extract the full dataset of comments related to the “Restoring Internet Freedom” docket using the FCC’s ECFS API. User submitted data consists of free form text, so as a first step, we performed data cleansing to remove special characters and standardized the various data fields. Not all commenters completed all data fields. After data cleansing, more that 19% of the total docket contained unusable street address, city, state, ZIP code, or email information.

Although 21.766 million comments were submitted to ECFS on the proposed rulemaking, there were far fewer unique comments (1.797million). The vast majority of comments were form letters (i.e. portions of pre-generated text). Users submitted the most prevalent comment in favor of and against Title II repeal 1.385 million and 7.569 million times, respectively. The top 10 and 100 most prevalent comments accounted for 66% and 89% of the total comments, respectively.

3.1 Data Completeness

When submitting comments to ECFS, in addition to the narrative text stating their position on the proposed rulemaking, users had the opportunity to include their name, street address, city, state, ZIP code, international address (for international commenters), and email address. We performed rudimentary data cleansing and transformation to remove special characters and to standardize the address, email, and comment fields. Null or invalid values for remaining fields were deemed “unusable”.

Attribute	Usability Criteria	AGAINST Title II Repeal (Usable %)	FOR Title II Repeal (Usable %)
filer	Non-NULL	99.9%	99.9%
address	Non-NULL	83.9%	99.9%
city	Non-NULL	81.2%	99.7%
state	State Abbreviation or State Name	81.1%	99.6%
ZIP code	5 digits or ZIP + 4	86.7%	98.0%
contact email	Contains “@”	85.4%	98.1%
address, city, state, ZIP code	All Criteria Above	78.6%	98.0%
address, city, state, ZIP code, contact email	All Criteria Above	69.7%	97.9%
international address	Country name exists	95.3%	51.2%

Table 1 – Data attribute usability

After data cleansing, more than 81% of the total docket contained decipherable (i.e. “usable”) street address, city, state, ZIP code, and email. Address and email information associated with comments for repeal were much more complete than against repeal. 98% of comments for the repeal of Title II were usable versus 70% of comments against the repeal of Title II. Any analysis performed on the address/email information alone would retain the vast majority of comments for the repeal of Title II, but would discount 30% of the comments against repeal.

We also performed rudimentary data cleansing on the comment field. We marked null or indecipherable comments (i.e. comments with a length less than 15 characters, containing random character or number sequences, etc.) as “Invalid.” Invalid comments accounted for upwards of 0.3% of the total.

3.2 Comments Summary

Although 21.766 million comments were submitted to the FCC on the proposed rulemaking, there were significantly fewer unique comments (1.797 million). A vast majority of the comments submitted originated from form letters with exact or similar phrasing. These comments likely originated from numerous sources organized by groups that were for or against the repeal of Title II. The top 5, 50, and 500 most prevalent comments made up 55%, 87%, and 89% of the total comments, respectively – meaning that the total overall sentiment of comment submissions could be extracted from just a relatively few number of comments.

The graph below shows the cumulative percentage of comments made up by the most prevalent comments in favor of and against Title II repeal. The top 25 most prevalent comments comprise essentially all (98%) of comments received in favor of Title II repeal, whereas the top 25 most prevalent comments make up roughly 80% of comments against repeal. This tells us that significantly more unique/non-form letter comments have been submitted against Title II repeal. It also tells us that if form letters are taken away, it seems that very few commenters took the time to manually enter a comment in favor of Title II repeal.

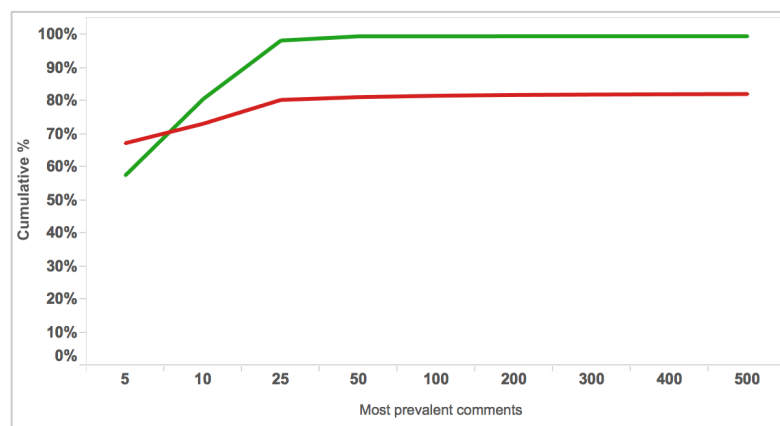


Figure 1 – Cumulative distribution of most prevalent comments

The general sentiment of comments was negative in tone regardless of whether the submitter was for or against repeal of Title II. Those arguing for the preservation of Title II maintained that repeal would eliminate a free and open Internet and would allow ISPs to prioritize access. Those in favor of repeal were

critical of the previous FCC's decision to classify ISPs under Title II and claimed overly burdensome regulations and government overreach were hindering Internet progress.

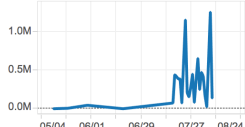
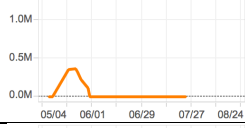
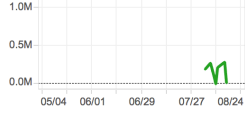
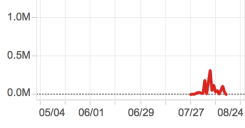
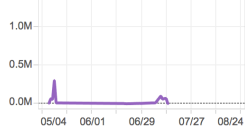



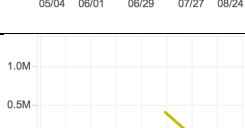
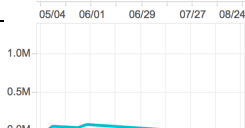
Comment	Count	Sentiment	% of Total	Count by date
(1) "I AM IN FAVOR OF STRONG NET NEUTRALITY UNDER TITLE II OF THE TELECOMMUNICATIONS ACT. SINCERELY, NAME"	7,569,190	Against Title II Repeal	34.8%	
(2) (VARIATIONS OF) "DEAR FCC, I AM CONCERNED ABOUT NET NEUTRALITY. I STRONGLY URGE THE FCC TO REPEAL (...)"	1,384,893	For Title II Repeal	6.4%	
(3) "IN 2015, CHAIRMAN TOM WHEELER'S FEDERAL COMMUNICATIONS COMMISSION FCC IMPOSED RESTRICTIVE TITLE II, UTILITY-STYLE REGULATIONS UNDER THE GUISE OF AN OPEN INTERNET. NOT ONLY HAVE THESE REGULATIONS INHIBITED INNOVATION IN THE INTERNET ECOSYSTEM, THEY HURT TAXPAYER (...)"	1,083,324	For Title II Repeal	5.0%	
(4) "BEFORE LEAVING OFFICE, THE OBAMA ADMINISTRATION RAMMED THROUGH A MASSIVE SCHEME THAT GAVE THE FEDERAL GOVERNMENT BROAD REGULATORY CONTROL OVER THE INTERNET. THAT MISGUIDED POLICY DECISION IS THREATENING INNOVATION AND HURTING BROADBAND INVESTMENT (...)"	1,047,049	For Title II Repeal	4.8%	
(5) "THE UNPRECEDENTED REGULATORY POWER THE OBAMA ADMINISTRATION IMPOSED ON THE INTERNET IS SMOOTHERING INNOVATION, DAMAGING THE AMERICAN ECONOMY AND OBSTRUCTING JOB CREATION. I URGE THE FEDERAL COMMUNICATIONS COMMISSION TO END THE BUREAUCRATIC (...)"	818,331	For Title II Repeal	3.8%	
(6) "THE CURRENT FCC REGULATORY SCHEME KNOWN AS TITLE II REPRESENTS AN UNPRECEDENTED INCREASE IN GOVERNMENT CONTROL OVER THE INTERNET. SUCH OVER-REGULATION IS HURTING OUR ECONOMY AND SUFFOCATING INNOVATION. I SUPPORT CHAIRMAN PAIS' PLAN TO RETURN TO A COMMON SENSE (...)"	615,151	For Title II Repeal	2.8%	
(7) "THE OBAMA-ERA FCC REGULATIONS KNOWN AS TITLE II ENABLE THE FEDERAL GOVERNMENT TO EXERT AN EXTRAORDINARY AND UNNECESSARY AMOUNT OF REGULATORY CONTROL OVER THE INTERNET. THIS BUREAUCRATIC OVERREACH IMPEDES INNOVATION, STIFLES INVESTMENT AND CONTINUES TO CREAT (...)"	562,280	For Title II Repeal	2.4%	
(8) "THE FCC'S OPEN INTERNET RULES NET NEUTRALITY RULES ARE EXTREMELY IMPORTANT TO ME. I URGE YOU TO PROTECT THEM. I DON'T WANT ISPs TO HAVE THE POWER TO BLOCK WEBSITES, SLOW THEM DOWN, GIVE SOME SITES AN ADVANTAGE OVER OTHERS, OR SPLIT THE INTERNET INTO FAST LANE (...)"	445,658	Against Title II Repeal	2.1%	
(9) "AS A CONCERNED TAXPAYER AND CONSUMER, I AM WRITING TO URGE THE FCC TO SET THE INTERNET FREE AND REMOVE THE INAPPROPRIATE, UNNECESSARY AND OVERLY VAST REGULATIONS CURRENTLY HOLDING BACK THE FULL POTENTIAL OF THE INTERNET (...)"	430,190	For Title II Repeal	2.0%	
(10) "OBAMA'S FEDERAL COMMUNICATIONS COMMISSION FCC FORCED REGULATIONS ON THE INTERNET THAT PUT THE GOVERNMENT, AND UNACCOUNTABLE BUREAUCRATS, IN CONTROL. THESE RULES HAVE COST TAXPAYERS, SLOWED DOWN BROADBAND INFRASTRUCTURE INVESTMENT, AND (...)"	410,097	For Title II Repeal	1.9%	

Table 2 – Top 10 most prevalent comments

3.3 Address Validation

As part of our data cleansing and quality assurance efforts, we performed address geocoding on 5.985 million unique addresses (a sample of approximately 65% of the total addresses found in comments through August 4, 2017) using Texas A&M University's batch geocoding service.¹ Using this service, considered at face value, we were able to determine the validity of physical addresses provided in the dataset. We treated addresses with an "AddressPoint," "Parcel," and "StreetSegmentInterpolation" geocode quality type as valid (i.e. real) addresses. Address data associated with comments for Title II repeal were more valid than those against.

- **Valid Addresses For Title II Repeal:**
 - 84% – when counting each unique address once
 - 83.8% – when counting addresses one or more times (based on the number of comments submitted from each address)
- **Valid Addresses Against Title II Repeal:**
 - 68% – when counting each unique address once
 - 23.1% – when counting addresses one or more times (based on the number of comments submitted from each address)

¹ <http://geoservices.tamu.edu/Services/Geocode/About/>

4. Results

The lack of user authentication by ECFS makes it difficult to determine “genuine” comment submissions. Emprata was also not able to authenticate the filer, address, email, or comment data used for this analysis, nor the methods used to collect those data elements. As a result, it is very difficult to draw any definitive conclusions from the comments found in the docket. Any conclusions that one might draw from the data would be based on the subset of data that they considered to be “real.”

Emprata used a hybrid text mining approach consisting of (1) manual sentiment assignment, (2) keyword/phrase matching, and (3) natural language processing (NLP) in its analysis (further described in the Appendix of this report). Using this approach, we assigned sentiment in favor of repealing Title II or against repealing Title II to each of the 21.766 million comments submitted to the FCC. A third categorization of comments, representing “neutral” and “invalid” comments, were also considered but were not detailed in this analysis (as they comprised a very small percentage of the overall dataset at 0.4%).

When considering all 21.766 million comments at face value in their entirety, domestically and internationally, the predominant sentiment was against repealing Title II. Also, there are significantly more unique comments against repeal that appear to have been “personalized” (i.e. not originating from a form letter). However, overall sentiment changes considerably when setting aside comments from suspicious email domains, and filtering (sub-grouping) the data by address and email, as is demonstrated in the following sections.

4.1 Considering All Comments

When considering all 21.766 million comments in their entirety, domestically and internationally, the predominant sentiment was against the repeal of Title II. Again, this portion of our analysis assumes that all comments submitted to the FCC are “genuine.”

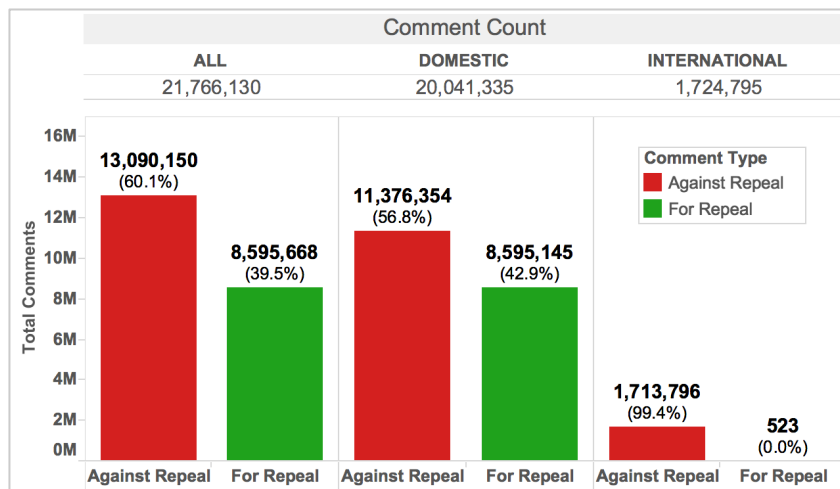


Figure 2 – Bar chart of sentiment when considering all data

There were three periods where significant quantities of comments were submitted, (1) mid-late May and (2) mid-late July, and (3) early-mid August. During the initial peak of comments in May, the overall sentiment of the comments fluctuated in sentiment. After this initial peak, during June and July, the sentiment of the comments received was predominantly against Title II repeal. The maximum number of comments received in a single day occurred on August 7 with just over 2.01 million comments.

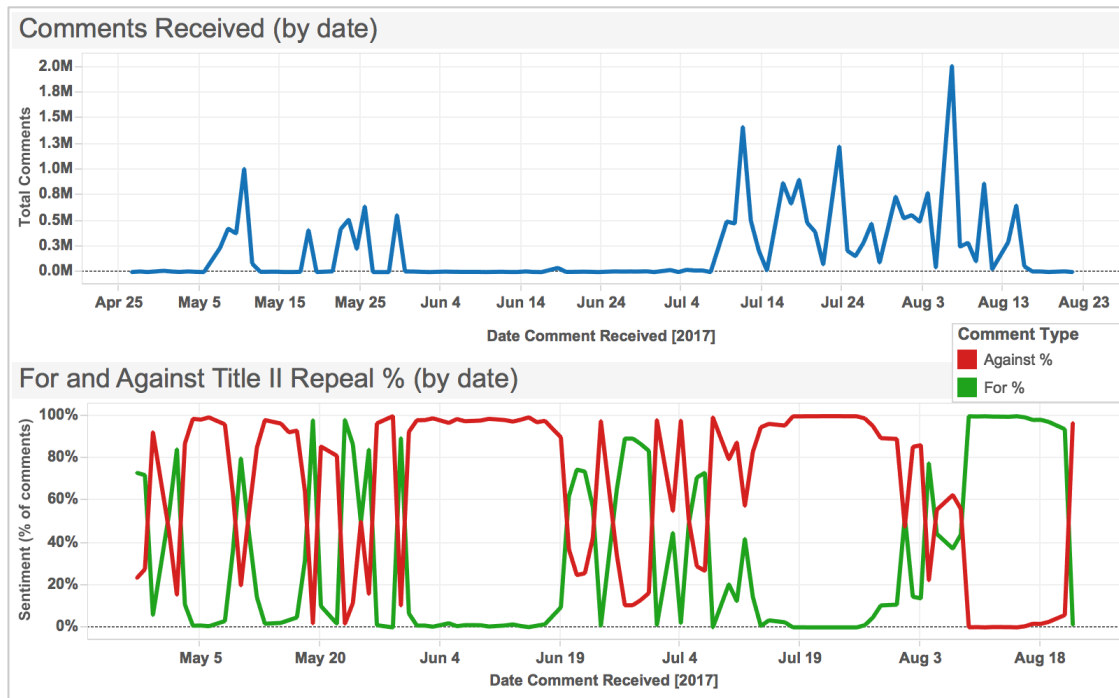


Figure 3 – Total comments received and sentiment percentage by date

The majority of international comments were received on July 12 (562k), which coincided with the highly publicized internet-wide Day of Action for Net Neutrality, with a smaller peak on May 25 (108k).

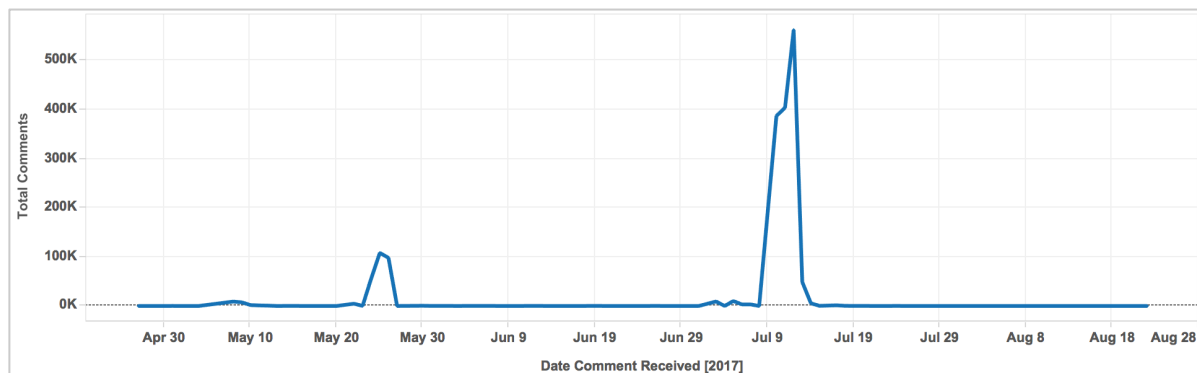


Figure 4 – Total international comments received by date

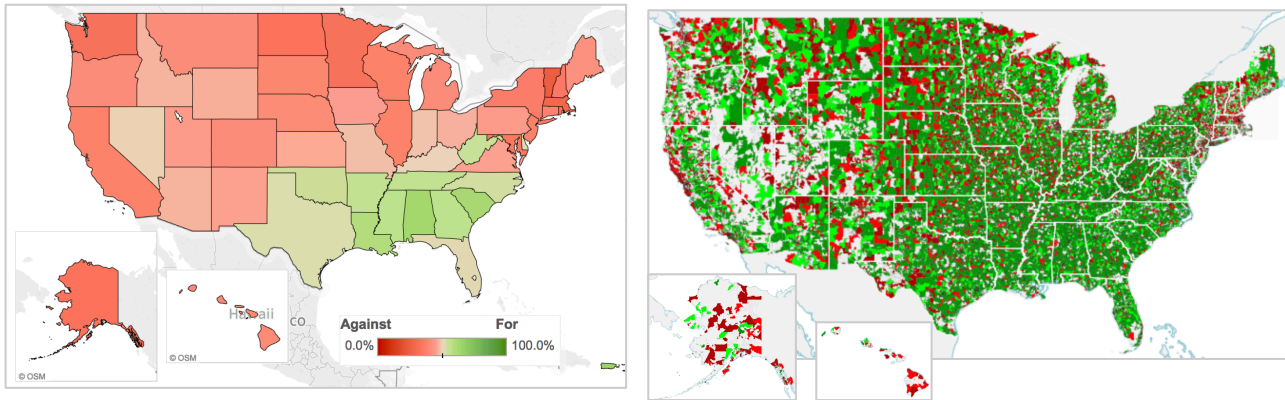


Figure 5 – Sentiment maps by state and zip code when considering all data

Comments pertaining to Title II were submitted from all U.S. states and territories. As is illustrated in the state (left) and ZIP code (right) maps above, the northeast, west coast, and urban population centers of the U.S. areas had the largest ratios of sentiment against Title II (shown in **RED**), whereas the south and southeastern regions were more in favor of Title II repeal (shown in **GREEN**). Not surprisingly, the majority of the U.S. comments originated from the states with the highest population including California, Texas, Florida, New York, and Illinois.

State	Total	Against Count	Against %	For Count	For %
UNKNOWN	2,522,301	2,469,549	97.9%	30,473	1.2%
California	1,982,480	1,244,220	62.8%	730,606	36.9%
Texas	1,593,220	761,062	47.8%	826,653	51.9%
Florida	1,146,374	558,356	48.7%	582,724	50.8%
New York	1,101,877	702,899	63.8%	395,731	35.9%
Illinois	793,022	499,866	63.0%	290,605	36.6%
Pennsylvania	791,742	443,727	56.0%	346,559	43.8%
Ohio	725,055	386,247	53.3%	337,392	46.5%
Georgia	678,477	292,528	43.1%	383,770	56.6%
Michigan	644,614	383,873	59.6%	259,226	40.2%
North Carolina	621,231	289,145	46.5%	330,654	53.2%
New Jersey	506,187	320,837	63.4%	183,881	36.3%
Virginia	453,317	246,992	54.5%	205,061	45.2%
Massachusetts	427,422	310,301	72.6%	115,944	27.1%
Missouri	424,111	219,841	51.8%	203,062	47.9%
Tennessee	423,341	187,585	44.3%	235,030	55.5%
Indiana	421,970	217,144	51.5%	203,918	48.3%
Washington	414,272	278,914	67.3%	133,645	32.3%
Alabama	360,015	132,901	36.9%	226,370	62.9%
Louisiana	339,527	139,196	41.0%	199,537	58.8%
Arizona	330,541	173,802	52.6%	155,430	47.0%
Colorado	330,133	197,019	59.7%	131,911	40.0%
Minnesota	327,103	223,165	68.2%	103,320	31.6%
Wisconsin	323,735	205,278	63.4%	117,620	36.3%
Maryland	323,050	213,326	66.0%	108,860	33.7%
South Carolina	316,141	118,805	37.6%	196,562	62.2%
Kentucky	305,721	153,976	50.4%	151,222	49.5%
Oklahoma	273,214	124,884	45.7%	147,807	54.1%

State	Total	Against Count	Against %	For Count	For %
Mississippi	262,461	113,623	43.3%	148,380	56.5%
Oregon	261,663	153,911	58.8%	105,586	40.4%
Arkansas	221,502	98,276	44.4%	122,836	55.5%
Kansas	206,681	111,664	54.0%	94,591	45.8%
Connecticut	205,177	127,851	62.3%	76,731	37.4%
Iowa	174,735	96,938	55.5%	77,501	44.4%
Utah	172,434	96,515	56.0%	75,490	43.8%
Nevada	156,056	78,416	50.2%	77,007	49.3%
Nebraska	131,835	79,554	60.3%	52,054	39.5%
New Mexico	118,244	64,824	54.8%	53,089	44.9%
West Virginia	113,174	51,239	45.3%	61,732	54.5%
Idaho	95,579	50,657	53.0%	44,717	46.8%
New Hampshire	80,147	52,229	65.2%	27,667	34.5%
Montana	76,384	45,015	58.9%	30,911	40.5%
Maine	75,199	44,180	58.8%	30,848	41.0%
Rhode Island	65,855	44,108	67.0%	21,633	32.8%
Hawaii	60,680	36,275	59.8%	23,936	39.4%
South Dakota	59,834	37,090	62.0%	22,665	37.9%
North Dakota	58,326	39,228	67.3%	18,966	32.5%
Delaware	57,169	27,732	48.5%	29,320	51.3%
District of Columbia	56,880	43,573	76.6%	12,865	22.6%
Alaska	53,175	35,587	66.9%	17,443	32.8%
Vermont	48,498	36,954	76.2%	11,410	23.5%
Wyoming	43,637	23,432	53.7%	20,162	46.2%
Puerto Rico	6,679	2,302	34.5%	4,337	64.9%
US Virgin Islands	2,386	2,087	87.5%	156	6.5%
Mariana Islands	1,138	1,133	99.6%	4	0.4%
Guam	248	195	78.6%	44	17.7%
American Samoa	161	124	77.0%	14	8.7%

Table 3 – Sentiment by state and U.S. territory when considering all data

Roughly 1.72 million international comments, representing 169 countries, were submitted, virtually all of which oppose repealing Title II. This includes roughly 526K comments where the international address entered was “United States.” All of those comments had a sentiment against Title II and were submitted using a “mail.ru” email domain. The majority of the international based comments originated from Russia, Germany, France, India, and Canada.

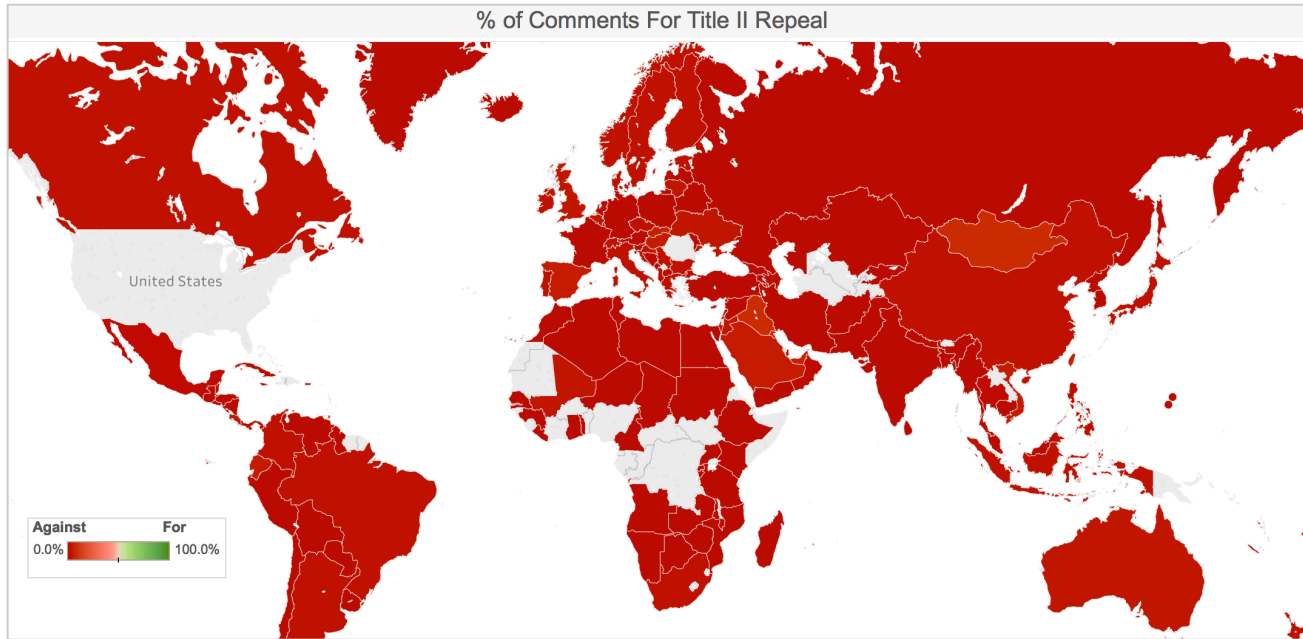


Figure 6 – Sentiment map by country for international comments

Country	Total	Against Count	Against %	For Count	For %
RUSSIA	444,938	444,925	100.0%	4	0.0%
GERMANY	433,092	433,033	100.0%	21	0.0%
FRANCE	194,554	194,530	100.0%	6	0.0%
UNKNOWN	91,221	80,968	88.8%	255	0.3%
INDIA	6,819	6,801	99.7%	6	0.1%
CANADA	4,684	4,590	98.0%	34	0.7%
JERSEY	2,346	2,345	100.0%	1	0.0%
GEORGIA	1,726	1,725	99.9%	1	0.1%
AUSTRALIA	1,707	1,649	96.6%	29	1.7%
TURKEY	1,118	1,116	99.8%	1	0.1%
UNITED KINGDOM	1,059	1,019	96.2%	18	1.7%
NETHERLANDS	997	970	97.3%	5	0.5%
OMAN	969	957	98.8%	5	0.5%
SWEDEN	897	868	96.8%	9	1.0%
MEXICO	715	707	98.9%	4	0.6%
IRELAND	659	645	97.9%	5	0.8%
NORWAY	649	635	97.8%	6	0.9%
NEW ZEALAND	612	594	97.1%	7	1.1%
DENMARK	596	583	97.8%	5	0.8%
PANAMA	568	567	99.8%	0	0.0%
SPAIN	415	392	94.5%	12	2.9%
SWITZERLAND	402	394	98.0%	3	0.7%
BELGIUM	389	377	96.9%	7	1.8%
FINLAND	377	364	96.6%	4	1.1%
BRAZIL	373	364	97.6%	3	0.8%
PORTUGAL	361	348	96.4%	3	0.8%
ISRAEL	345	333	96.5%	8	2.3%
AUSTRIA	321	317	98.8%	1	0.3%
SOUTH AFRICA	311	303	97.4%	4	1.3%
ITALY	307	296	96.4%	5	1.6%
POLAND	239	233	97.5%	1	0.4%

Table 4 – Sentiment by country for international comments

4.2 Duplicative Comments

Filtering (sub-grouping) the data by address and/or email resulted in a sentiment predominantly in favor of Title II repeal. This is due to the fact that (1) a relatively large portion of the comments against Title II did not contain complete address and/or email information and (2) many comments were filed from an entity listing the same email and/or physical address, indicating that many entities filed multiple comments. The following graphs illustrate the sentiment for and against repealing Title II for each data sub-grouping.

- **ONE PER ADDRESS** – Considers only the first comment for each street address, city, state, ZIP code combination
- **ONE PER EMAIL** – Considers only the first comment from each email address
- **ONE PER ADDRESS/EMAIL** – Considers only the first comment from each unique address and email combination

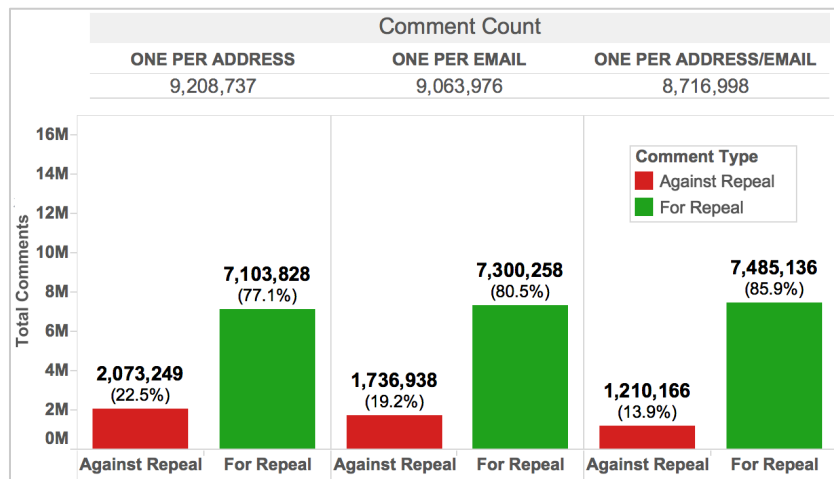


Figure 7 – Bar chart of sentiment when eliminating duplicate comments by address and email

9.93 million comments were filed from entities listing the same email and/or physical address, indicating that many entities filed multiple comments. This was more prevalent in comments against repeal of Title II (accounting for 82% of the total duplicates). The majority of these duplicates were associated with comments associated with [FakeMailGenerator.com](https://www.fakemailgenerator.com) email domains. After taking this into account, general sentiment favors repeal of Title II nationwide (as shown below).

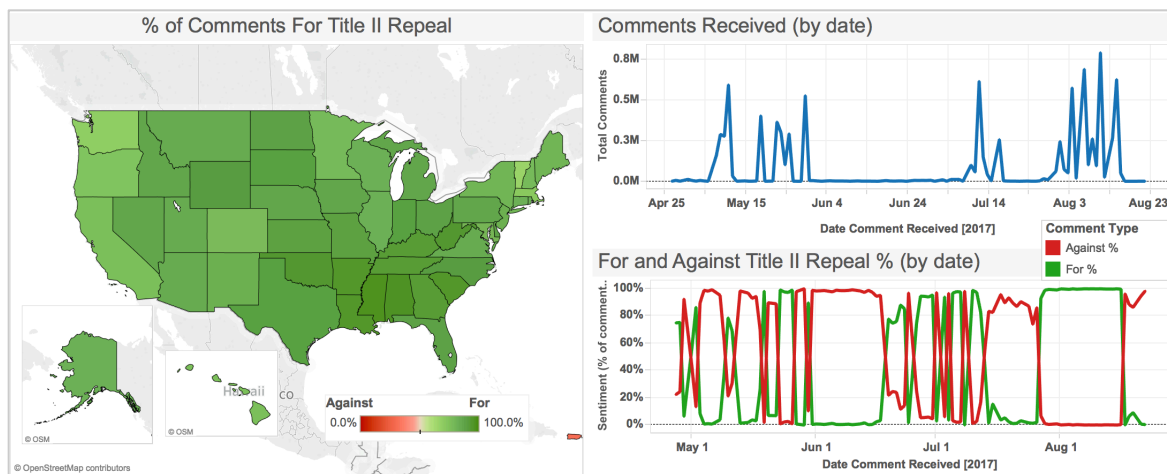


Figure 8 – Map and time series of sentiment for “One per Address/Email” case

4.3 Valid Addresses

As was stated previously, 84% of addresses for Title II repeal were found to be valid versus 68% of addresses against. When considering duplicate address entries, the percentage of valid comments against repeal drops considerably to 23%. When considering only those comments with valid addresses, sentiment is predominantly for Title II repeal (as is shown below). Although the vast majority of states were for Title II repeal using this criteria, a handful of states did remain against Title II repeal (Washington, Vermont, Oregon, and Massachusetts).

- **VALID ADDRESS** – Considers only comments where an exact address match was found
- **ONE PER VALID ADDRESS** - Considers only the first comment from each valid address

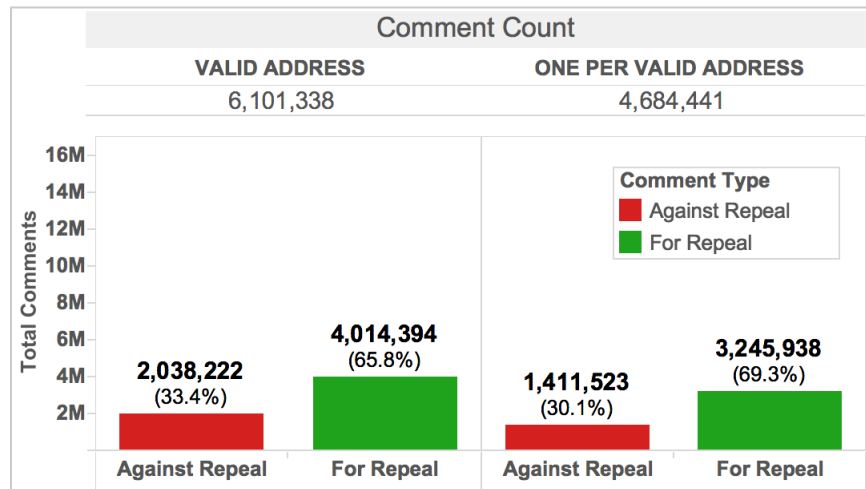


Figure 9 – Bar chart of sentiment when eliminating duplicate comments by address and email

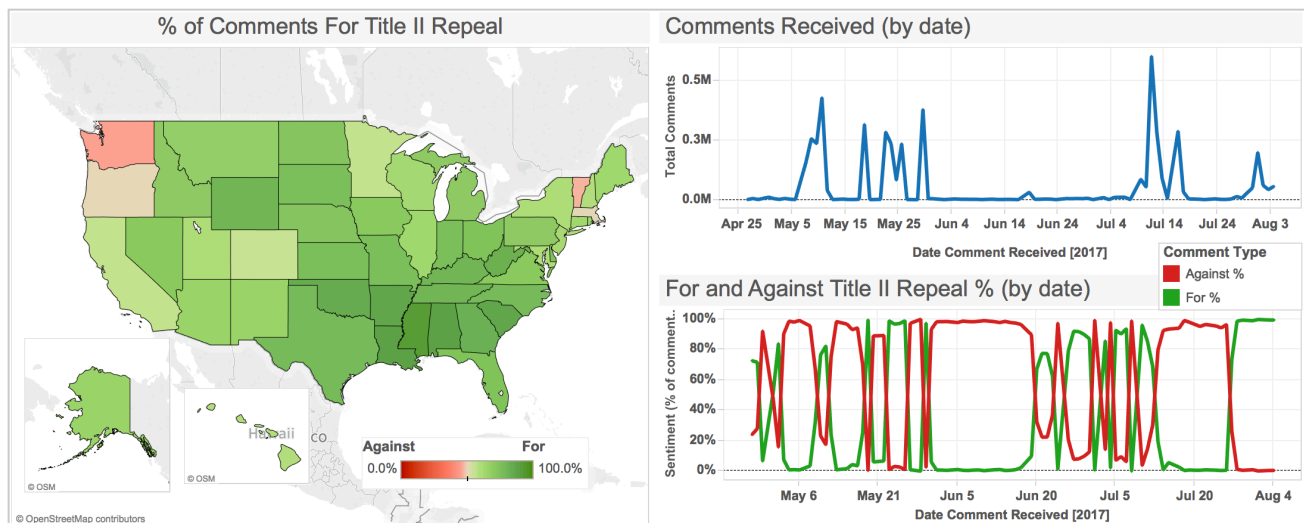


Figure 10 – Map and time series of sentiment for “One per Valid Address” case (data through 8/4/17)

4.4 Unique Comments

The overwhelming majority of comments for and against repealing Title II are form letters (i.e. pre-generated portions of text) that appear multiple times in the docket. The form letters likely originated from numerous sources organized by groups that were for or against the repeal of Title II. As part of our work, we identified 26 form letters as well as numerous other portions of repeated text. Form letters make up upwards of 89.6% of comments against Title II repeal and upwards of 99.6% of the comments for Title II repeal.

Notwithstanding the presence of form letter comments, there are significantly more unique comments submitted against Title II repeal (1.77 million) versus for Title II repeal (24k). Naturally, comments submitted through form letters (or automated methods) would appear in the comments multiple times. Comment text appearing only once would have a much smaller chance of originating from a form letter (or automated method). Presumably, these comments originated from individuals that took the time to type out their own personalized message. There are considerably more “truly unique” comments (appearing only once in the docket) against repeal (1.52 million) versus 23k for repeal. Although these comments account for less than 10% of the total, the differences are significant.

	AGAINST Title II Repeal	FOR Title II Repeal
Total Comments	13,090,150	8,595,668
Unique Comments	1,769,811	24,348
Truly Unique Comments (Comment appears only once)	1,516,296	23,309

Table 5 – Unique comments by sentiment

To help identify outliers and abnormalities in the comment text, we measured the length of each comment and placed them into comment length “bins” (for example, 25-50, 50-75, 75-100, etc.). We would expect this distribution to resemble a bell shaped curve, with a peak at some reasonable comment length and downward sloping curves on their side of that peak. Large spikes in this distribution would be indicative of form letter (or automated method) comments, as they would all have the same (or similar) lengths. We made the following observations (illustrated on the graphs below):

1. Comments against Title II repeal have a much more natural distribution than those for Title II repeal, especially between comment lengths of 25-425 characters.
2. There are virtually no comments with a length between 25-225 characters submitted by those for Title II repeal, which one might consider to be a realistic comment length submitted by a “real” person.
3. There are three significant outliers in the comments against Title II repeal and eight for Title II repeal. These represent comments associated with form letters.
4. Upon further investigation of comments with outlier lengths, form letters and repeat comments were identified for both cases, but were more prevalent in comments for Title II repeal.

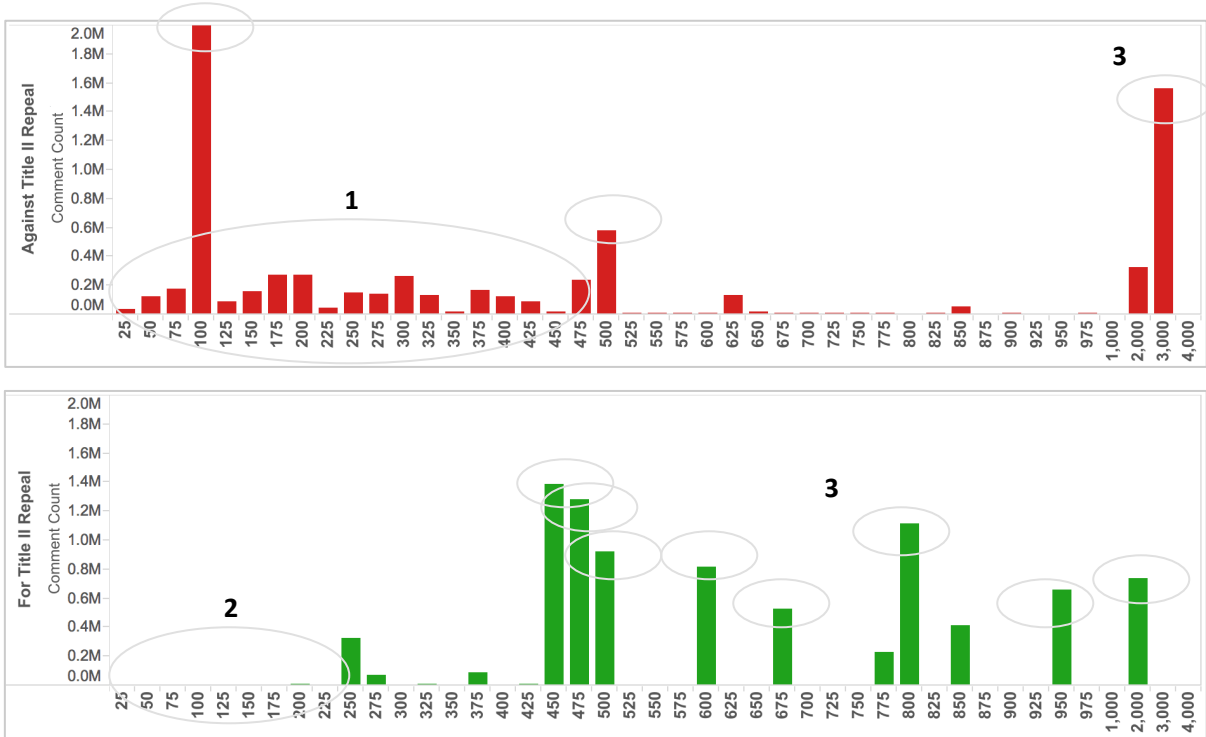


Figure 11 – Distribution of comment length by sentiment

4.4 Email Domains

Users submitted comments from approximately 119,000 email domains. Email addresses are self-entered in ECFS (i.e. ECFS does not in any way verify email addresses or domains). As a result, a large volume of comments were submitted under domains sourced to FakeMailGenerator.com,² including einrot.com, which for example, submitted 793,136 comments. The website *einrot.com* makes explicit that: “This service allows anyone to create a temporary email address that is only capable of receiving email. No legitimate email will ever be sent from einrot.com.”³

In addition, this grouping of 10 domains also contained a very similar comment count (~775k), which is irregular. Collectively, these comments were used in 7.75 million submissions (36%). Virtually all of those comments oppose repealing Title II. Assuming that comments submitted from these email domains are illegitimate, and excluding them from the total count, sentiment favors repeal of Title II (61% for, 39% against). Some subset of these comments could be “genuine,” but that is difficult to determine. For example, it should be noted that 107,920 of physical addresses associated with these 10 email domains (based on our sampling of comments through August 4, 2017) were found to be valid.

There are numerous instances where sentiment received from domains was either 100% against or for Title II repeal. However, the sentiment of comments listing more recognizable and widely used email domains, such as Gmail, Yahoo, and Hotmail, were generally in support of repealing Title II.

² Domains associated with FakeMailGenerator.com include: einrot.com, jourrapide.com, armyspy.com, fleckens.hu, cuvox.de, rhyta.com, gustr.com, superrito.com, teleworm.com

³ What is einrot.com? (<http://www.einrot.com>)

Email Domain	Total	Against Count	Against %	For Count	For %
EMAIL.COM	4,929,839	869,467	17.6%	4,039,880	82.0%
YAHOO.COM	2,703,019	429,069	15.9%	2,266,653	83.9%
PORNHUB.COM	1,030,000	1,030,000	100.0%	0	0.0%
EINROT.COM	793,136	793,133	100.0%	3	0.0%
JOORRAPIDE.COM	782,039	782,034	100.0%	5	0.0%
ARMYSPY.COM	780,515	780,509	100.0%	5	0.0%
CUVOX.DE	775,903	775,897	100.0%	6	0.0%
FLECKENS.HU	775,820	775,818	100.0%	2	0.0%
RHYTA.COM	773,755	773,743	100.0%	10	0.0%
DAYREP.COM	769,719	769,711	100.0%	8	0.0%
GUSTR.COM	768,838	768,833	100.0%	5	0.0%
SUPERRITO.COM	767,485	767,482	100.0%	3	0.0%
TELEWORM.US	765,106	765,097	100.0%	8	0.0%
HOTMAIL.COM	665,063	104,086	15.7%	557,936	83.9%
AOL.COM	627,573	77,342	12.3%	547,974	87.3%
HURRA.DE	363,357	363,357	100.0%	0	0.0%
COMCAST.NET	204,756	51,219	25.0%	152,437	74.5%
MSN.COM	108,016	18,748	17.4%	88,627	82.1%
ICLOUD.COM	104,855	10,075	9.6%	94,346	90.0%
ATT.NET	78,386	16,317	20.8%	60,699	77.4%
OUTLOOK.COM	78,221	12,164	15.6%	65,670	84.0%
LIVE.COM	69,148	12,161	17.6%	56,560	81.8%
SBCGLOBAL.NET	68,368	22,087	32.3%	45,450	66.5%
MAIL.RU	60,605	60,430	99.7%	165	0.3%
YMAIL.COM	44,496	4,009	9.0%	40,350	90.7%
VERIZON.NET	40,674	16,712	41.1%	23,562	57.9%
BELLSOUTH.NET	39,690	6,631	16.7%	32,495	81.9%
COX.NET	39,367	9,666	24.6%	29,517	75.0%
ME.COM	25,401	13,109	51.6%	11,830	46.6%
CHARTER.NET	23,877	5,926	24.8%	17,868	74.8%

Email Domain	Total	Against Count	Against %	For Count	For %
EARTHLINK.NET	17,150	9,980	58.2%	7,006	40.9%
ROCKETMAIL.COM	16,851	1,816	10.8%	14,992	89.0%
MAIL.COM	16,097	1,841	11.4%	14,221	88.4%
MAC.COM	14,230	11,503	80.8%	2,521	17.7%
WINDSTREAM.NET	13,291	1,158	8.7%	12,089	91.0%
EXAMPLE.COM	12,426	12,073	97.2%	79	0.6%
JUNO.COM	9,684	3,433	35.5%	6,181	63.8%
GMX.COM	7,451	735	9.9%	6,699	89.9%
FRONTIER.COM	6,861	1,975	28.8%	4,843	70.6%
ROADRUNNER.CO..	6,086	2,341	38.5%	3,707	60.9%
OPTONLINE.NET	5,843	4,761	81.5%	995	17.0%
AIM.COM	5,633	1,656	29.4%	3,918	69.6%
CENTURYLINK.NET	5,159	1,259	24.4%	3,855	74.7%
SUDDENLINK.NET	5,124	712	13.9%	4,399	85.9%
EMBARQMAIL.COM	4,414	979	22.2%	3,402	77.1%
NETSCAPE.NET	3,893	1,232	31.7%	2,606	66.9%
NETZERO.NET	3,519	787	22.4%	2,706	76.9%
REDIFFMAIL.COM	3,165	29	0.9%	3,127	98.8%
CABLEONE.NET	3,151	522	16.6%	2,616	83.0%
TAMPABAY.RR.CO..	3,140	939	29.9%	2,160	68.8%
ATT.COM	3,126	318	10.2%	321	10.3%
PACBELL.NET	2,993	2,214	74.0%	675	22.6%
Q.COM	2,965	1,917	64.7%	1,013	34.2%
CFL.RR.COM	2,867	801	27.9%	2,022	70.5%
TWC.COM	2,808	850	30.3%	1,944	69.2%
CENTURYTEL.NET	2,803	737	26.3%	2,047	73.0%
NETZERO.COM	2,642	447	16.9%	2,186	82.7%
MCHSI.COM	2,589	719	27.8%	1,863	72.0%
TDS.NET	2,544	609	23.9%	1,927	75.8%
FRONTIERNET.NET	2,524	905	35.9%	1,607	63.7%

Table 6 – Comment count and sentiment by email domain

Although comments were submitted from approximately 119k domains, 72.3% of those domains only contributed a single comment and 94% contributed less than 5 comments. A total of 19 domains submitted more than 100,000 comments.

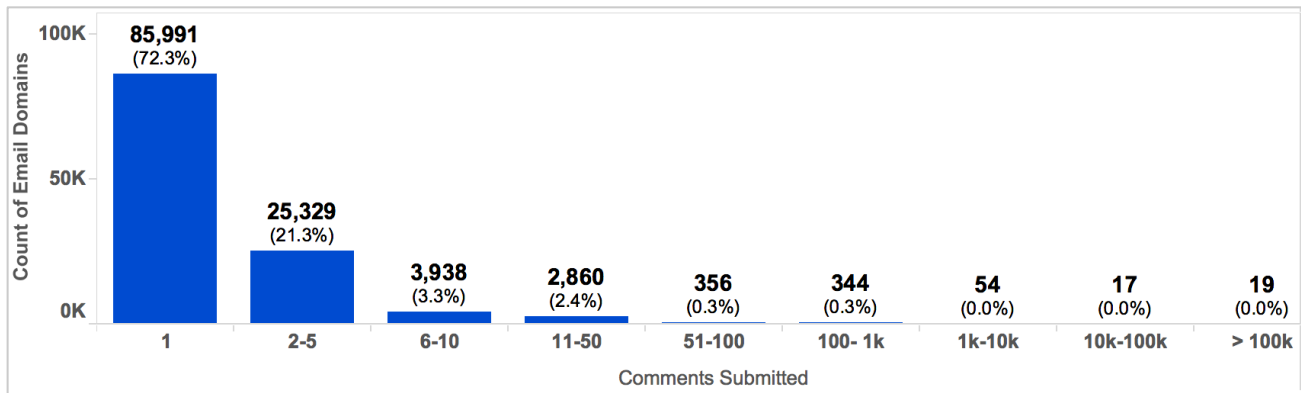
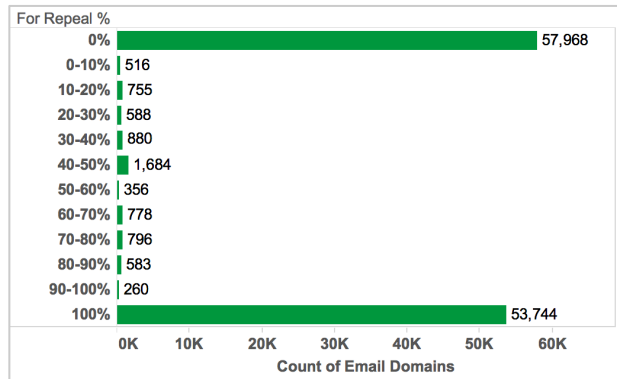
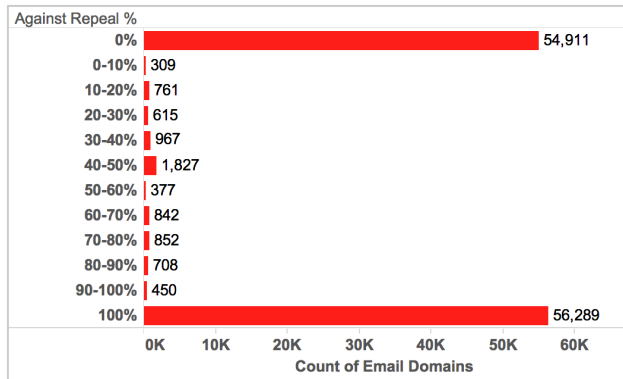


Figure 12 – Number of email domains by comments submitted

The bar charts below show the frequency of domains for a given sentiment % range. For example, comments were submitted from 1,827 email domains where the percentage of overall sentiment against Title II repeal was between 40-50%. Interestingly, comments submitted from 56,968 and 53,744 domains were 100% against and for repeal, respectively. When only considering domains with at least 100 comment submissions, the distribution changes significantly (sentiment % is spread more evenly in the ranges between 0 and 100%).

All Domains



Domains with at least 100 comments

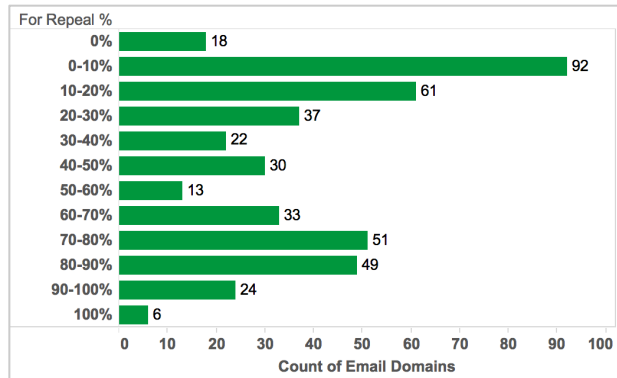
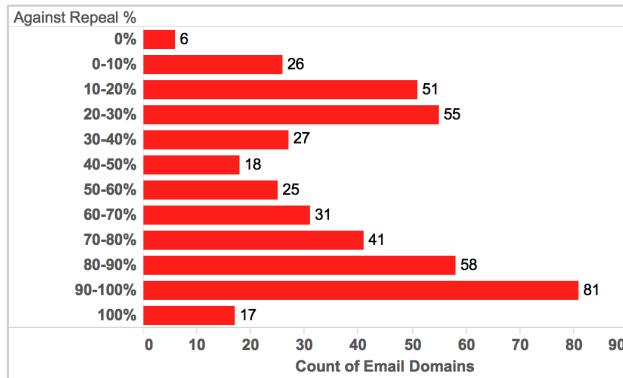


Figure 13 – Number of email domains by sentiment %

5. Conclusion

Any conclusions that one might draw from the data would be based on the subset of data that they considered to be “genuine.” Determining “genuine” comments in the “Restoring Internet Freedom” docket is challenging due to the fact that the FCC’s ECFS does not authenticate users. As a result, Emprata was not able to authenticate the filer, address, email, or comment data used for this analysis, nor the methods used to collect those data elements. Nevertheless, the following observations can be made:

- Considering all comments at face value, general sentiment is against Title II repeal. When eliminating duplicative comments and those originating from seemingly “fake” email domains, general sentiment changes to for Title II repeal.
- The vast majority of comments originate from form letters (i.e. portions of pre-generated text).
- There are significantly more unique/non-form letter comments submitted against Title II repeal than for Title II repeal.
- The physical address and email information provided by commenters for Title II repeal is much more complete (i.e. usable) than those provided against Title II repeal. This also holds true when it comes to the validity of physical addresses.
- Comments from email domains associated with FakeMailGenerator.com comprise a large percentage of comments against Title II repeal and the docket as a whole.
- Essentially all international comments are against Title II repeal, but the majority originate from irregular email domains and addresses that we did not validate.
- The majority of duplicative comments (i.e. multiple comments submitted from the same address/email) are from commenters that are against Title II repeal (primarily associated with comments from FakeMailGenerator.com).

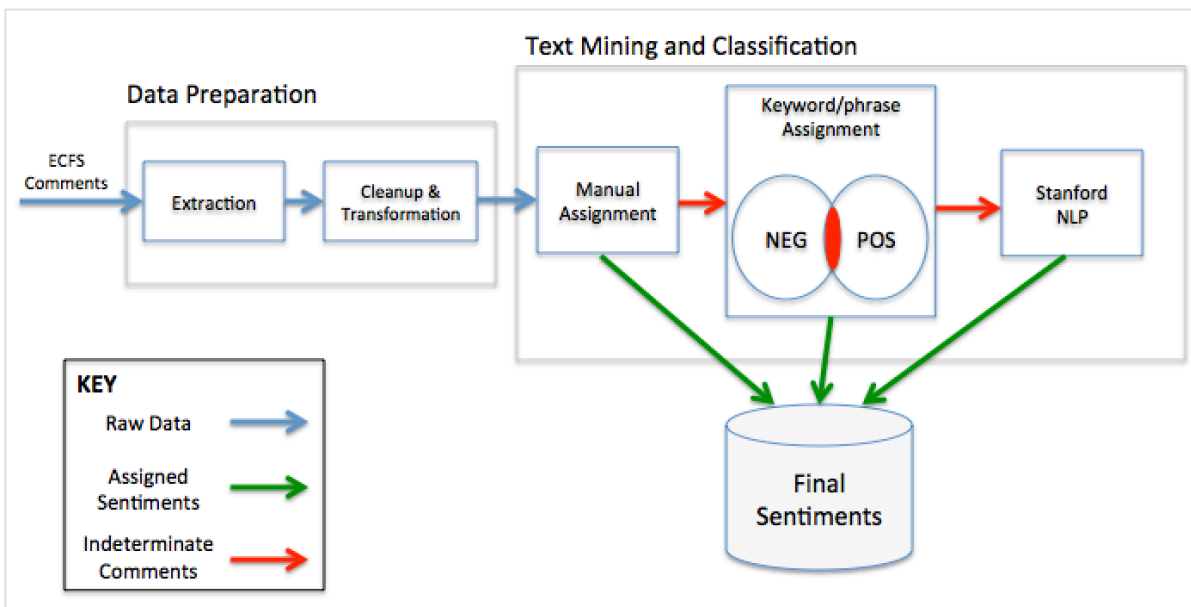
6. Disclaimer

Emprata was contracted by *Broadband for America* to perform an independent and unbiased analysis of the comment data received by the FCC in response to their Notice of Proposed Rulemaking (WC Docket No.17-108) on the “Restoring Internet Freedom” proceeding. Emprata does not have a vested interest in whether Title II is repealed or not. This analysis, completed during the course of two weeks, was performed to obtain an understanding of the comment sentiments, both for and against Title II repeal, in response to the docket. We are releasing our report for policymakers and the public to review and come to their own conclusions. This analysis is by no means complete; further analysis by other researchers is appropriate and could uncover other insights.

APPENDIX A – METHODOLOGY

Sentiment Analysis takes unstructured data and finds ways to convert that data into a more structured format in order to obtain deeper insights into the “sentiment” of the text. Sentiment Analysis is accomplished by identifying words that can determine a person’s sentiment (using a dictionary of words known to be positive or negative). For each word in each sentence, the sentiment is drawn and scored. By aggregating the scores, one can categorize narrative text as being for Title II repeal, against Title II repeal, or neutral.

We explored a number of Natural Language Processing (NLP) and machine learning methods for Sentiment Analysis, but found their resulting classifications to be inconsistent and unreliable. This was due to the fact that *essentially all* comments had a negative sentiment, regardless of whether the comment was for or against repealing Title II. As a result, an alternate methodology was selected for this analysis (described below).



Sentiment assignment methodology

Data Extraction and Transformation

We extracted the full comments dataset using the FCC’s public ECFS API (URL: <https://publicapi.fcc.gov/ecfs>) and ingested the data into a relational database. To prepare the data for analysis, we performed various data cleanup/transformation activities, including the following:

- Removal of special characters, tabs, new lines, leading spaces, and trailing spaces
- Transformation of text fields to upper case
- Standardization of state field to two character State Abbreviation
- Standardization of ZIP code field to five digits, including padding of leading zeros
- Extraction of country name from the “international address” field
- Removal of duplicate comment (allowing only one per submission ID)

Text Mining and Classification

We employed a hybrid text mining approach consisting of (1) manual sentiment assignment, (2) keyword/phrase matching, and (3) natural language processing (NLP) in its analysis. Given the cumulative distribution of the comments (where a small number of comments accounted for a large percentage of total comments) and prevalent keywords/phrases, this hybrid approach garnered the most accurate results.

Step 1: Manual Sentiment Assignment: We assigned for/against Title II repeal sentiment to the 500 most prevalent comments and known form letters. These comments accounted for 95% of the total). To ensure their correctness, the manual assignments were verified by a second resource.

Step 2: Keyword/phrase matching: We assigned for/against Title II repeal sentiment for remaining data based on prevalent keywords/phrases encountered during manual sentiment assignment. This step resulted in a sentiment assignment for an additional 4% of the total comments received. Sentiment phrase examples included:

- Against Title II Repeal: “Protect Net Neutrality,” “Preserve Title II,” “Support Net Neutrality”
- For Title II Repeal: “Repeal Title II,” “Stifle Innovation,” “Government Overreach”

Comments including phrases from both for/against Title II repeal keyword/phrase list were “conflicting” and were therefore set aside (i.e. not assigned a sentiment in this step).

To validate the results of the matching, we manually verified the sentiment assignments against a random sample of 1,000 records. The overall accuracy of the keyword/phrase matching was 99.8%.

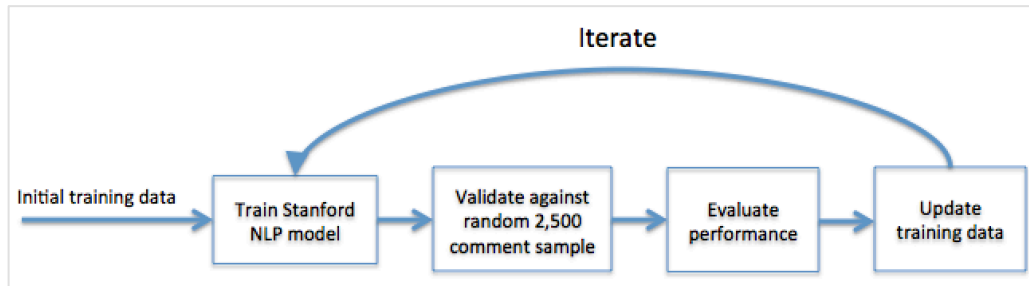
	Prediction: Against Title II Repeal	Prediction: For Title II Repeal
Actual: Against Title II Repeal	488 (48.8%)	0 (0%)
Actual: For Title II Repeal	2 (0.2%)	510 (51.0%)

Confusion matrix for keyword/phrase matching

Step 3: Natural Language Processing (NLP): We assigned a sentiment to the remaining less than 1% of comments using the Stanford CoreNLP Recursive Neural Network (a supervised machine learning technique). Using training sets of both known for/against Title II repeal sentiment to build data features, this tool was able to classify comments based on similarities to each training set. In other words, the algorithm looks at each comment and determines “does this comment look more like the set of known comments for repealing Title II or the set of comments against repealing Title II?”

Stanford NLP splits text into sentences and scores each word in every sentence based on their association to text found in a training dataset. The average score of each sentence and ultimately the overall text is determined. These scores are determined by how frequently a given phrase occurs near a set of known words from the trained data sets. Based on the score, a sentiment can be assigned.

We used a sampling of known sentiments for and against repeal from the Manual Sentiment Assignment step to train the Stanford NLP model (40 comments total). To ensure a proper sampling across the entire spectrum of comments, we performed a text association analysis. This analysis eliminated multiple comments with similar phrasing from being included in the training set.



Stanford NLP training

To fine-tune the supervised learning model, we implemented an iterative technique consisting of training, validating, and updating the training data used in the NLP model. At each iteration, we measured performance against a random sampling of 2,500 records. After several iterations, we selected the final set of models to use in this analysis.

Due to the variety of comments, we utilized three models in our classification approach:

- **Naïve Bayes:** This is the main model used. This model focuses on one common text categorization task, sentiment analysis, the sentiment analysis extraction of sentiment, the positive or negative orientation that a writer expresses toward Title II repeal.
- **Maximum Entropy:** The Maximum Entropy (MaxEnt) classifier is closely related to a Naive Bayes classifier. This was used occasionally against a variety of comments that were misclassified.
- **Support Vector Machine (SVM):** We took the set of comments that could not be classified correctly from either Naive Bayes or the Entropy and re-ran them using the SVM classification technique.

Using a random sample of 875 comments, the NLP model's overall accuracy was 95.2%. The model correctly classified upwards of 98.6% of comments against repealing Title II. In addition, the vast majority of the comments the NLP was asked to categorize were against Title II repeal.

	Prediction: Against Title II Repeal	Prediction: For Title II Repeal
Actual: Against Title II Repeal	833 (95.2%)	12 (1.4%)
Actual: For Title II Repeal	28 (3.2%)	2 (0.2%)

Confusion matrix for Stanford NLP

The model did have relatively greater difficulty classifying comments in favor of repealing Title II. We presume this was due to the fact that we were forced to use form letter comments to train the model (given the lack of non-form letter comments for Title II repeal in the docket). Form letter comments are not as “genuine” as when someone manually writes a comment, which caused difficulty in the assignment of sentiment for “genuine” comments for Title II repeal. It should also be noted that there were very few comments for repeal in the NLP classification set. This is consistent with the cumulative distribution function and lack of unique comments for Title II repeal.

The performance of the NLP model surely could have been improved with more time (especially when classifying comments for Title II repeal), but given that the model was asked to classify less than 1% of the data, we did not spend much time improving its accuracy.