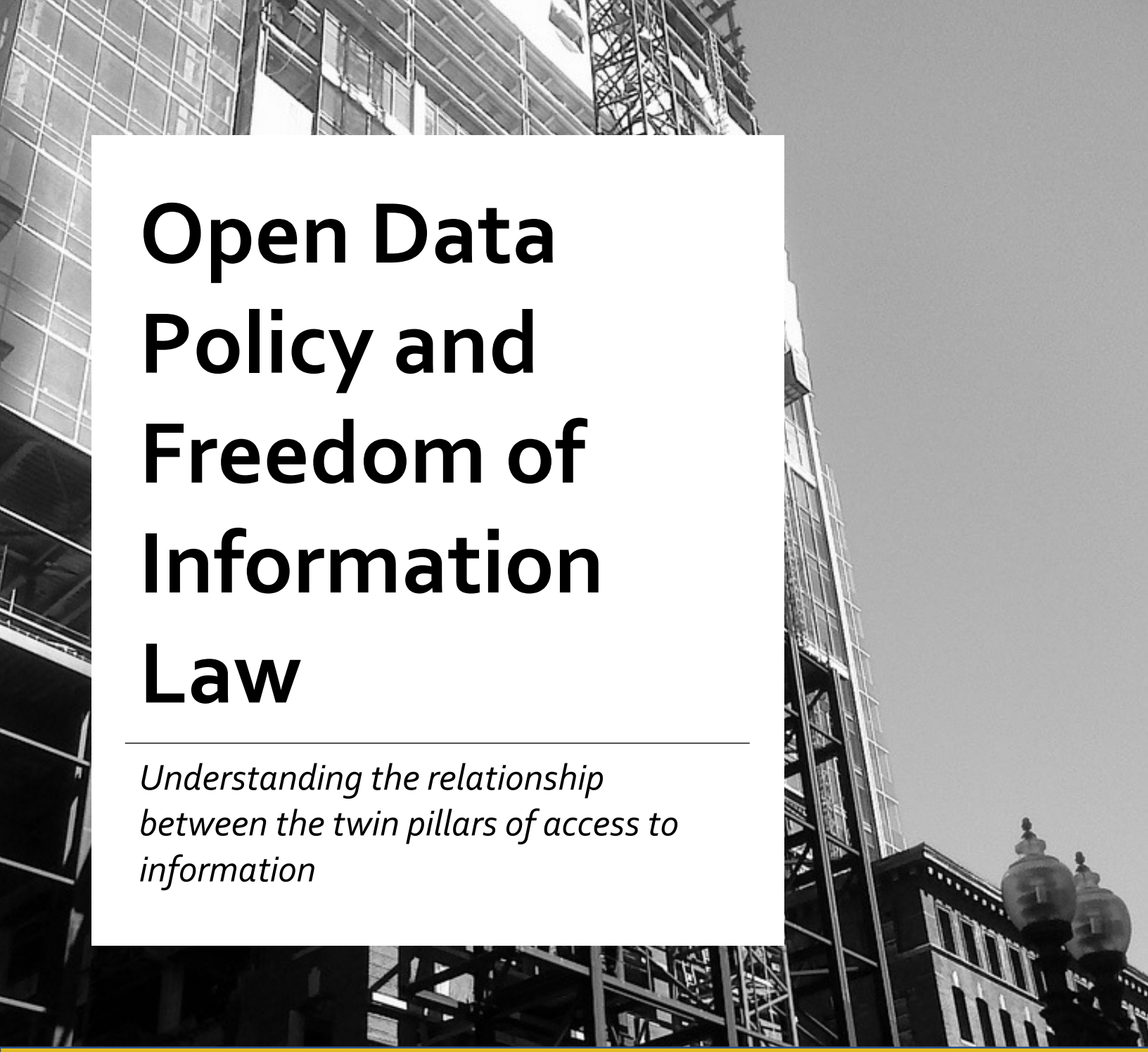


Open Data Policy and Freedom of Information Law

Understanding the relationship between the twin pillars of access to information

Alena Stern | Sunlight Foundation Fellow

October 2018

SUNLIGHT FOUNDATION

Table of Contents

Acknowledgements 3
Abstract 4
Background 5
Methodology 7
Results and Recommendations 14
Conclusion 25
Annexes 26
Annex A: Full List of Cities Included in Study 26
Annex B: Detailed Methodology 29
Annex C: Study Design 37
Annex D: City-specific PRR Volume Plots 40
Annex E: Detailed Results 45

Acknowledgements

I would like to thank the following individuals and institutions for their guidance and support: Daniel Maliniak, Nathan Zencey, Nicholas Bell, David Christensen, Bruce Desmarais, Bob Gradeck, and Tom Johnson for providing feedback on the pre-analysis plan and research methodology. I would also like to thank Reed Duecy-Gibbs at Next Request for providing valuable background information regarding online public record request systems. In addition, I'd like the cities included in this study for responding to my public record requests to provide historical request data and answering my questions regarding the data. Finally, I'd like to thank the Sunlight Foundation's Open Cities team and Bloomberg Philanthropies' What Works Cities program for supporting this research.

Abstract

Freedom of Information (FOI) laws have long been a bulwark protecting the rights of citizens to access government information. Yet as a growing number of cities across the United States choose to proactively disclose information through open data programs, understanding the relationship between open data and FOI has become critical to cities wanting to use their limited open government resources most effectively. We adopted a mixed-methods approach to evaluate the relationship between open data and FOI; we designed a panel data study to assess how adopting open data programs affects the volume of public record requests (PRRs) and the diversity of requestors. We used Latent Dirichlet Allocation (LDA) model to understand what types of data are requested via PRR, and whether this changes after cities adopt an open data program. We found that adopting an open data program significantly decreases the volume of PRRs received by cities and that this effect grows over time. Additionally, we found that more robust open data programs are associated with greater decreases in PRR volume. Our LDA analysis revealed that data on parcel records, permits, and plans as well as police incident reports are the most demanded by citizens, though there is significant variation in demand across cities. Our results regarding the relationship between open data programs and the diversity of requesters and the subject of data requested were inconclusive and should be the focus of future research efforts.

Background

Since the passage of the Freedom of Information Act (FOIA) in 1967, Freedom of Information Laws (FOI) have been critical tools in ensuring public access to government information as a fundamental, democratic right. But the momentum is shifting from a reactive model of disclosing government data in response to public record requests (PRRs) to more systematic efforts to proactively publish open data as cities "set the default to open" through the passage of open data policies. In 2017 alone, 28 U.S. cities published open data policies, bringing the total of local governments with open data policies to more than 105. The Sunlight Foundation's Open Cities Team has played a central role in supporting this growing body of open data policies, either working directly with or providing resources to 60 of the local and state governments that have adopted open data policies as part of the What Works Cities initiative.

Therefore, as more and more cities adopt open data policies it is important for city governments and the organizations like the Sunlight Foundation that work with them to understand the relationship between open data and FOI to ensure that cities use limited open government resources to respond to citizen information needs most effectively. In this paper, we explore this relationship to answer the question: are the channels of open data policy and FOI law competitors or complements?

Specifically, we aim to understand how adopting an open data program (defined in this paper as adopting an open data policy, launching an open data portal, or both) affects PRRs in three main dimensions: volume, requesters, and content.

If cities can understand the change in volume of PRRs, they stand to save significant time and money. If proactive disclosure reduces PRR volume, the potential cost savings could be significant - the US Federal Government spent $448,961,678 processing FOIA requests in 2018 and Yakima, Washington (population 93,986) spends $500,000 annually responding to PRRs. Yet evidence on the effect of open data on PRR volume is limited and mixed: a 2016 Yale Law Journal article notes that Federal FOIA requests increased by over 16% since passing the US

Open Data Policy. On the other hand, a 2014 Reinvent Albany report estimated that proactive disclosure could reduce New York City FOIA requests by 20%, saving the city $3.5 million annually. Socrata and the City of Chicago cite a 50% reduction in FOIA requests since launching its open data portal.

Understanding who requesters are and what they need is necessary for city officials to use limited resources effectively and efficiently for better government transparency. Open data can can provide significant value to a wide group of stakeholders - including advocates and journalists, non-profit organizations, local businesses, researchers, government agencies, and concerned citizens. Cities adopting an open data program must publish the right data that meets the needs of these user groups. publish the right data. Previous analysis has suggested that cities use public records requests as a roadmap to understand citizen demand, yet there is no systematic research on what data is being requested via PRR.

Methodology

Research Questions

As the first multi-city study of the relationship between open data and FOI, our research aimed to supplement other research on this topic, which has so far provided mixed signals about whether passing an open data policy increases or decreases the number of PRRs cities receive. So, the first primary question of our study was to analyze the effect of passing an open data policy on PRR volume.

We also aimed to understand whether open data effectively creates a "big tent" with room for all of the many different constituencies city governments want to reach. Open data policy could popularize the possibility of civic engagement and inspire new residents to seek information from their local governments. Alternately, if the types of requesters reduced in variety over time, it may signal that open data is meeting the needs of some types of requesters, but that there are unique frequent requestors with idiosyncratic needs.

Finally, we aimed to understand how the content of PRRs changes as cities release data demanded by citizens. If the type of data people are requesting doesn't change, cities may need to better prioritize highly-requested data for publication as open data, improve the quality and comprehensiveness of published data, or to invest in raising citizen awareness of available data resources.

We designed a study to evaluate the impact of a city choosing to adopt an open data policy on multiple different facets of PRRs. Specifically, we aimed to answer the following questions:

1. Does adopting an open data policy affect the volume of PRRs a city receives?
2. Does adopting an open data policy affect the variety of requestors submitting PRRs?
3. Does adopting an open data policy affect the time it takes cities to complete PRRs?

4. Does the robustness of a city's open data program affect the magnitude of its effect on the outcomes above?

We were also interested in understanding what types of information citizens are requesting via PRR, and how this is affected by the adoption of an open data program (referring to the combination of a city's open data policy and open data portal). We investigated this relationship through the following questions:

5. What types of information are most requested via PRR?
6. Do we see changes in the types of information requested by citizens after the adoption of an open data program?

Data Collection and Sources

The main data source for this research was PRR data from 52 medium-sized cities across the United States that operate standardized, online public records request platforms, which ensured a relatively consistent level of ease and accessibility of submitting a request in our sample.[1] We included What Works Cities that maintain an online PRR platform or other cities that we were able to identify through internet searches of keywords. We obtained the data from these platforms using a variety of methods:

1. Exporting the full archive of PRRs hosted on the online portal as a .csv file
2. Scraping the full history of PRR data from portals which publish previous requests, but do not offer an export option
3. Downloading PRR data that has been published on city's open data portal
4. Submitting a public record request to obtain the archive of PRR data

[1] We limit our analysis to cities with a population between 10,000 and 1 million residents in 2016 to align with the What Works Cities definition of a medium-sized city. See Table 1 in the Annex for a full list of cities included in the study and the method used to access their PRR data.

Note that in all cases in which we identified a city with an online PRR platform but the city did not proactively publish PRR data, we submitted a PRR request for the archive of past PRRs.[2] We did not ultimately include a city in our analysis if the city was unable to produce responsive information to our request or required payment for staff time to complete the request. While this is not a nationally representative or randomized sample, we believe the lessons learned from this large and diverse group of cities will provide broadly applicable insights for cities.

For each city, we downloaded or requested data from the date when the city implemented its public record request portal through the end of May 2018, the last full month prior to the start of data collection. Our sample includes a total of 236,616 public record requests dating from October 2009 - June 2018. Only 33 of the 52 cities in our sample publish the PRR text, representing 110,063 PRRs.[3]

To isolate the effect of adopting an open data policy on PRRs, we included several demographic and political variables in our analysis that have been shown by previous research to be correlated with citizen demand for government data. See Table 2 in Annex A for a full list of covariates considered for inclusion in the final models with data sources.

Analysis Methodology

To answer questions 1-4, we designed a statistical analysis to evaluate the effect of adopting an open data program on our outcomes of interest. For questions 1-3, our treatment variable is a binary variable indicating whether a city had an open data policy in place in a given month.

[2] We removed our own PRRs from the sample prior to analysis.

[3] Cities have the discretion not to publish public record requests containing sensitive or personally identifying information (ex. information on a crime committed against a minor or a requestor's social security number). Often, cities redact sensitive information. In rare cases the city will release a summary of the request produced by city staff. We included these in our count but we will have limited information regarding the substance of the request.

For question 4, we wanted to assess whether the scope of a city's open data program impacts our outcomes of interest. To do this, we considered binary treatment variables indicating whether a city had the following in place in a given month: 1) open data policy, 2) open data portal, 3) robust open data policy, and 4) robust open data portal. We also considered a treatment variable that is an aggregate score reflecting which of the four open data program elements a city has in place.

We considered an open data policy robust if it received a score a majority of possible points in a previous study that evaluated policies for compliance with the Sunlight Foundation's 31 open data policy guidelines.[4] We considered an open data portal robust if it has data on at least two of the following three topics: budget, crime, and spending. Past research indicates these data are highly demanded and all the governments examined here would likely manage at least these datasets. We applied the robustness designation to all months that the portal was in existence, as we did not always know exactly when datasets were first added to a portal.

Our data for this analysis was an unbalanced panel where the unit of observation is a city and the unit of time is a month and year. We therefore used a panel statistical model. There were several possible panel statistical models that we could have used (pooled OLS, random effects, and fixed effects) which rely on different assumptions about the variance and randomness of the error in our observations to produce valid results. We ran each model and tested whether the relevant assumptions hold and found that random effects would be the appropriate model for analysis, likely because the time invariant city-specific effects vary randomly across cities. See Annex B for more detail.

We included covariates in our model to control for other factors besides adoption of an open data program which likely influence the PRRs a city receives as well as year dummy variables to control for year-specific time effects: population, percent of the population with bachelor's

[4] A policy received 1 point for full compliance with each guideline, 0.5 points for near-compliance, and -1 points for enacting the opposite of the guideline.

degrees, whether the political lean of the state is Republican, whether the city has a Mayor-Council governance structure, percentage of population that is white, median age, year dummy variables for 2015-2018, months the PRR portal has been in place, percentage of the population that is male, percentage of the population that is between the ages of 25-34. Our basic model was as follows:

$$R_{it} = \beta_0 + \beta_1 T_{it} + \beta_2 C_i + \beta_3 D + e_{it}$$

Where:
R_{it} is the outcome of interest for city i at month and year (time) t,
T_{it} is a dummy variable indicating whether city i has been assigned to the treatment group (e.g. has passed an open data policy) in time t,
C_i is a vector of time-invariant socio-economic, government structure, and budget variables at the city level (see footnote 10),
D is a vector of year dummy variables (2015-2018), and
e_{it} is an i.i.d disturbance term.

We also ran a version of our model that included interaction terms between our treatment variable and selected covariates to understand if the marginal effect of the treatment varies across time and city characteristics. Our three outcomes of interest were: 1) the number of PRRs received per 10,000 residents, 2) the ratio of unique requesters to total requesters, and 3) the average length of time in days between the PRR submission and completion dates (each measured for a given city, month, and year).[5,6] We chose number of PRRs per 10,000 residents to normalize our dependent variable for the effect of population, as cities with larger populations are both more likely to be treated (because they may have larger budgets and government capacity to adopt an open data program) as well as have a higher volume of PRRs due to the larger pool of potential requestors.

We used several robustness checks of our findings: 1) taking the log of the dependent variable, 2) lagging the treatment variable, 3) robust standard errors, and 4) inverse propensity score

[5] 11 cities in our sample either publish the email address or name of the user. We treated each unique email address/name as a unique user (though we recognize that the same user may use multiple names or addresses or that two individuals with the same first and last name may be different people). Open data published as a result of this study will redact this information.
[6] Only 6 cities in our sample publish request creation and completion dates.

weighting. The inverse propensity score weighting weighted each observation by the inverse of its likelihood of being in the treatment group (the propensity score) so that observations in the treatment group and control group that are more similar to each other were given more weight in fitting the model. This controlled for pre-treatment differences between our treatment and control group observations that may affect both the likelihood of an observation to be treated, and the outcomes of interest.

To answer questions 5 and 6 about the subject of PRRs, our primary challenge was figuring out how to group the unstructured raw text data of PRRs into a finite number of coherent categories as none of the cities in our sample publish categorizations.[7] To do this grouping and categorization, we used Latent Dirichlet Allocation (LDA), which is a generative statistical model within natural language processing. LDA groups documents (in our case, PRR text) into a fixed number of topics assigned by the researcher by first assigning each word in a document randomly to a topic, and through many iterations identifies which words appear more frequently together, eventually converging into the fixed number of topics where each topic is comprised of a cluster of associated words. Prior to training the LDA model on our data, we performed extensive cleaning of the data - including stemming words and removing punctuation and stop words - to improve the performance of the model. The full data cleaning process is described in Annex B.

We then assigned topic composition scores to each PRR which represented the proportion of the content of that PRR that belonged to one or more topics, with the maximum score of 1.0 representing a PRR perfectly fitting in a single topic. Similarly, if the content of a PRR was evenly split across four topics, it would have received a score of 0.25 for each of the four topics. We then added up these topic composition scores across PRRs to identify the most popular topics of information requested by citizens. We tested a number of different methods for calculating topic popularity, which are outlined further in Annex C.

[7] A small subset did publish the department to which they routed the request.

Our final scoring methodology applied a dampened popularity metric to mitigate the amount a single city affects the overall rating by taking the natural log of the topic popularity scores for each city and adding across cities. This is critical because the number of observations per city varied significantly (see Table 1 in Annex A). We then mapped the 60 topics that the model produced to the type of data that the city would provide to satisfy the request. For example, the model clustered requests for police reports on automobile thefts into a separate category from other police report requests. Yet in both cases, the responsive data would be a police report, so we group them together. This mapping process grouped the 60 topics into 19 data type clusters, which can further be grouped into 8 larger subject categories. The full results at all levels of aggregation can be found in the results section of the Annex.

We answered question 6 (regarding the changes in the type of information requested after adopting an open data program) qualitatively by assessing overall trends and city-specific results. We compared the proportion of requests represented by each data type in months with an open data policy in place versus months without an open data policy for those cities in our sample that have observations before and after the passage of an open data policy. For our city-specific analysis, we examined the results in Greensboro, North Carolina because it has the largest number of observations and pre-treatment months of data.

Results and Recommendations

CITIES THAT ADOPT OPEN DATA POLICIES RECEIVE FEWER PUBLIC RECORD REQUESTS.

We regressed the number of monthly public record requests cities receive per 10,000 residents against the treatment variable of adopting an open data policy and the list of covariates described above. The regression results are provided below:

```
Unbalanced Panel: n = 52, T = 2-96, N = 1472

Coefficients:
                 Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)   -7.5943e+01  9.7206e+01 -0.7813 0.4347741
policy        -6.4802e+00  2.0708e+00 -3.1293 0.0017870 **
population    -2.8392e-05  1.2366e-05 -2.2959 0.0218208 *
pct_bachelor  -2.8606e+01  5.1408e+01 -0.5565 0.5779874
rep            5.1032e+00  8.2904e+00  0.6156 0.5382820
pct_white     -1.1131e+01  1.8349e+01 -0.6066 0.5441970
mayor_council  1.1345e+01  5.8649e+00  1.9344 0.0532608 .
median_age     5.9682e-01  7.0180e-01  0.8504 0.3952401
X2015         -1.7420e+00  1.8973e+00 -0.9181 0.3587070
X2016          1.9351e+00  2.2157e+00  0.8733 0.3826314
X2017          5.5015e+00  2.6841e+00  2.0497 0.0405728 *
X2018          1.1023e+01  3.0903e+00  3.5671 0.0003727 ***
months         6.1402e-02  4.7964e-02  1.2802 0.2006923
pct_male       1.3989e+02  1.8174e+02  0.7697 0.4415764
pct_25_34      4.2831e+01  1.1228e+02  0.3815 0.7029142
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    303140
Residual Sum of Squares: 277150
R-Squared:      0.085748
Adj. R-Squared: 0.076963
F-statistic: 9.76083 on 14 and 1457 DF, p-value: < 2.22e-16
```

Adopting an open data policy is associated with a significant decrease in the number of PRRs that a city receives. Adopting an open data policy has a statistically significant effect on our outcome variable at the 1% confidence level; the variable coefficient suggests that, on average, adopting an open data policy is associated with a decrease of 6.48 public record requests received per 10,000 residents per month. For the average city in our sample, with a population just over 220,000, this would mean approximately 143 fewer PRRs per month. This

represents a decrease in PRR volume of approximately 30% relative to the average monthly volume in 2017-2018 for our control group.[8]

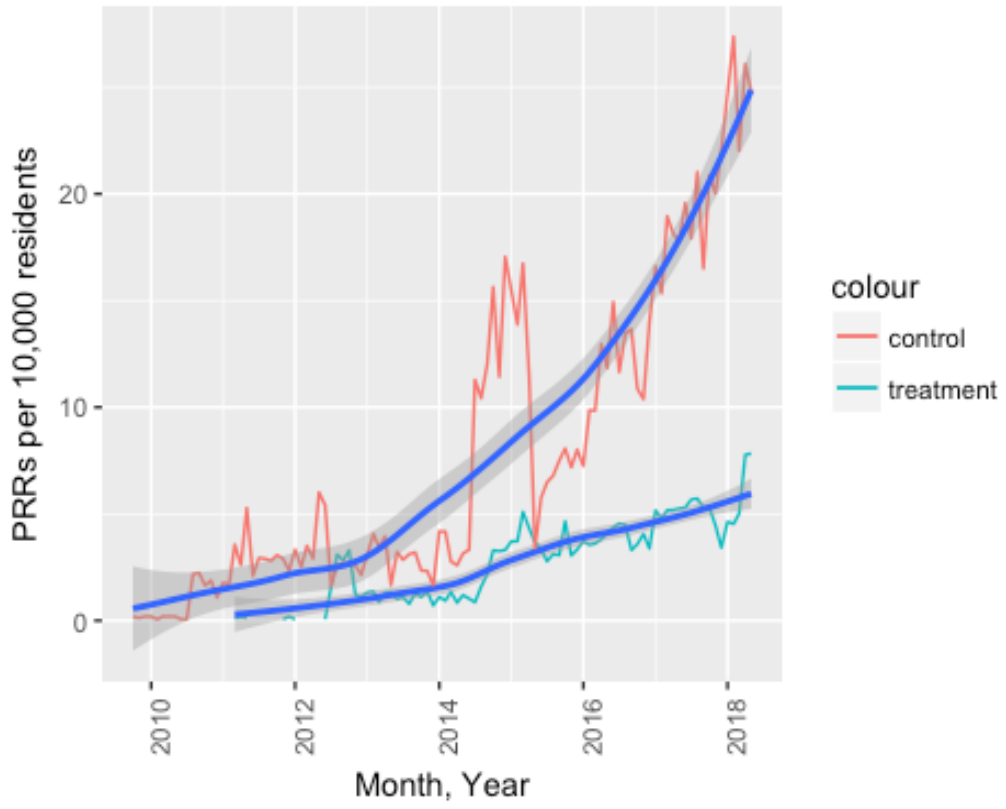
Population has a significant, negative effect on the number of PRRs received per 10,000 people but a positive and significant effect on the raw count, meaning that population growth eventually outpaces PRR growth. Taking the negative and significant population term in the population-adjusted dependent variable model with the result that the population term was positive and highly significant in the regression model with raw count as the dependent variable (see Annex C), suggests that the relationship between population and PRR volume is positive but nonlinear; as population increases so does PRR volume, but the increase in PRR volume does not keep pace with population size for very large cities. This could be the result of citizen demotivation caused by longer response times in very large cities, or because some types of demand expressed via PRR do not scale proportionally with population.

The Mayor-Council variable was significant at the 10% level, with cities that have a Mayor-Council government type experiencing an average of 11.3 more PRRs per month per 10,000 residents holding all else equal.[9] This may be because the Mayor-Council government structure generates interest in the activity of government, or because of differences between cities that adopt Mayor-Council governments and other cities illustrated in Annex A Table 4.

Figure 1: Average PRR Volume Over Time for Treatment and Control Cities

[8] Average volume of PRRs for a city in our sample that did not adopt an open data policy.

[9] The Mayor-Council government system is characterized by having a mayor who is elected by voters. This is distinct from the Council-Manager system of government, which has a city manager as the executive. These are the two most common municipal government structures in the US.



In addition, we found that the magnitude and significance of the effect of adopting an open data policy increases over time, with only the coefficients on the 2017 and 2018 year dummy interaction terms being statistically significant. This finding is corroborated by the fact that average PRR volume significantly increases over time, as shown in Table 1. Possible explanations include steady population growth, growing interest in government data, and/or increasing tech-savviness of the population. Figure 1 illustrates the average PRR count per 10,000 residents of treatment and control cities over time; adopting an open data policy does not decrease the number of PRRs that cities receive, but rather *slows the rate of growth in PRR volume*. This trend held for individual cities and a trimmed sample dropping the largest and smallest cities (see Figure 1 in Annex A).

Therefore, adopting a policy of proactive disclosure of public information appears to be one of the best ways for cities to effectively respond to growing demand. The demonstrated cost and

time savings from displaced PRRs is one tool that open data advocates can use to encourage governments to adopt open data policies.

ADOPTING ROBUST OPEN DATA PROGRAMS YIELDS A GREATER REDUCTION IN PRRS

To understand whether the robustness of an open data program affects the magnitude of the change in PRRs cities experience, we ran the random effects regression model above replacing the policy treatment with a variety of different treatments: 1) portal, 2) robust portal, 3) robust policy, and 4) an aggregate treatment score of the level of treatment in each city.[10]

We found that adopting a robust open data portal reduces PRR volume significantly more than adopting a portal alone and adopting a robust open data policy produces a greater reduction in request volume than a policy alone, though the difference is not statistically significant. The model that included portal as the treatment variable showed that launching a portal did not have a statistically significant effect on PRR volume, while launching a robust portal had a significant and negative effect with a coefficient of -7.15.[11] We ran a significance test on the two coefficients and found that the difference is significant at the 5% level (see Annex C for more detail). Our regression model with robust policy as the treatment had a statistically significant coefficient of -9.59.[12] Extending our analysis above, this means that adopting a robust open data policy is associated with a 44% decrease in the volume of PRRs received relative to the average monthly PRR volume in 2017-2018 for a city in our sample that did not adopt an open data policy. However, the difference of coefficients test between the policy and robust policy coefficients found that the two coefficients are not significantly different. It is clear that as cities invest in more robust open data portals, the magnitude of the effect on PRRs increases. There is also some initial evidence that cities adopting more robust open data

[10] 1 point is assigned to a city for whether they have a policy, robust policy, portal, and robust portal in place. The treatment score ranges from 0-4.

[11] All but one of the cities (Albuquerque, NM) with a robust portal also have adopted an open data policy.

[12] We dropped the following 5 cities from our analysis for which the robustness of the city's open data policy was not included in the previous research upon which we based the robust policy variable: Cape Coral, Fort Worth, Greensboro, Riverside, Palo Alto.

policies will increase the magnitude of the effect on PRRs, though this finding should be interpreted cautiously given the null result of the difference of coefficients significance test.

Launching an open data portal will only affect the volume of PRRs once the city has released sufficient data on the portal to displace the need for certain PRRs.

THE EFFECT OF ADOPTING AN OPEN DATA PROGRAM INCREASES OVER TIME.

We ran models with interaction terms to understand how the impact of adopting an open data policy changes based on city characteristics and time. The marginal effect of adopting an open data policy grew with each year. In fact, when we interacted the policy term with dummy variables for 2015-2018, the effect of adopting an open data policy only became significant in 2018, with a coefficient twice as large as that in 2017 (-17.46 in 2018 versus -8.65 in 2017).

For each additional month the policy is in place, a city can expect to get about 4 fewer PRRs than the month before, holding all else equal. When we ran the interacted model with the addition of a variable that measures the number of months the policy has been in place with raw PRR count as the DV, the coefficient on the interaction term between policy and this variable was statistically significant with a coefficient of -3.99.

There is clear evidence of a maturation effect of open data programs. It takes time for citizens to change their behavior to use open data resources.

PROPERTY AND CRIME/PUBLIC SAFETY INFORMATION ARE MOST REQUESTED

After grouping the 60 topics generated by our final LDA model into 19 data type groups, we found that the most frequently requested data types were in the categories of crime and public safety and property as illustrated in Figure 2 and Table 6 below:

Figure 2: PRRs by Category

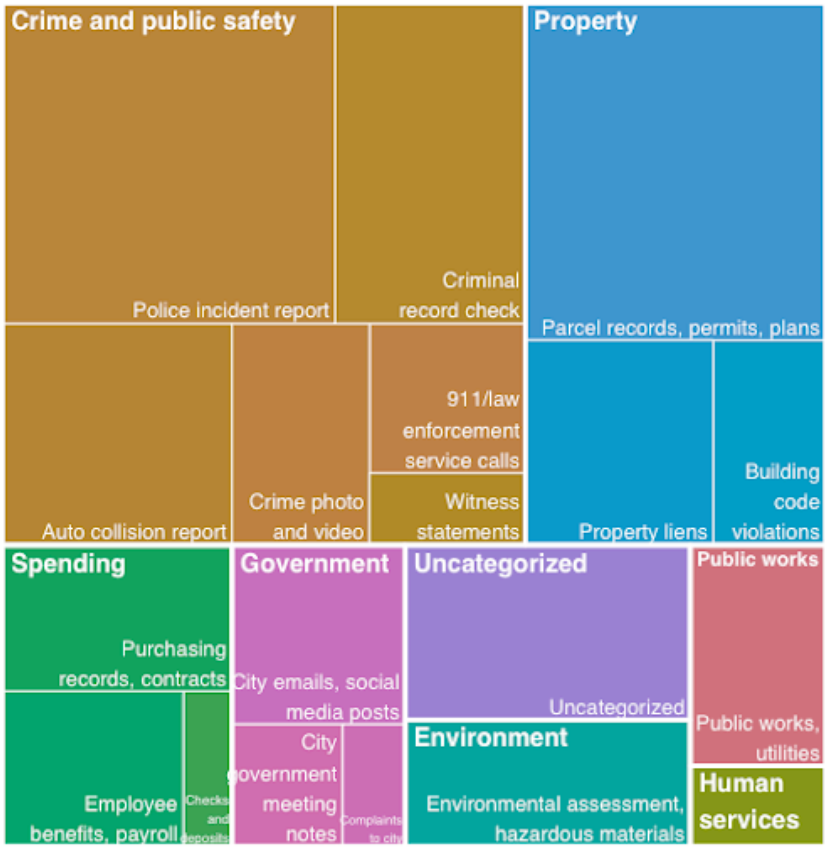


Table 2: Top PRR Data Types by Rank

Rank	Data Type	Demand
1	Police incident report	Highest

2	Parcel records, permits, plans	Highest
3	Criminal record check	High
4	Auto collision report	High
5	Uncategorized	High
6	Property liens	Medium
7	Environmental assessment, hazardous materials	Medium
8	Purchasing records, contracts	Medium
9	City emails, social media posts	Medium
10	Crime photo and video	Medium
11	Public works, utilities	Medium
12	Employee benefits, payroll	Medium
13	911/law enforcement service calls	Medium
14	Building code violations	Medium
15	City government meeting notes	Low
16	Witness statements	Low
17	Human services cases	Low
18	Complaints to city	Low
19	Checks and deposits	Low

We caution against placing significant emphasis on the relative rank of adjacent data types given the noise in the numeric scores produced by the model. However, there were clear popularity tiers where we saw a steep drop-off in total score between data types. These are given by the "Highest", "High", "Medium", and "Low" designations in Table 2. We see that parcel records, permits and plans and police incident reports were the most frequently requested types of data. Furthermore, crime and public safety and property data were the overall most common categories (crime and public safety was the most frequently requested category).

CITIES SHOULD INVEST IN RELEASING PARCEL RECORDS, PERMITS, AND PLANS DATA AND POLICE INCIDENT REPORT INFORMATION AS OPEN DATA.[13]

It is important to note that the dampened popularity approach to mitigate city-specific impacts on the overall rankings did have a clear effect on both topic and data type popularity (full results in Annex C). Indeed, we found that the relative popularity of different topics varies considerably across cities. Seven of the 19 different data types was the most requested data type in at least one of the 33 different cities in our sample. 14 different topics finished in the top spot across our sample.

One particularly striking example of city variation was the demand for auto accident reports. The most popular single topic of the 60 was requests by insurance companies for auto accident reports for insurance claims. Interestingly, this result was driven by the overwhelming volume of requests on this topic from five cities in our sample in Washington state (Everett, Arlington, Redmond, Kirkland, and Pullman). In fact, when we used the dampened popularity method, this topic fell to 34th in the rankings. Figure 3 above illustrates the disparity in demand across cities by showing the deviation of the percentage of total requests comprised by this topic for each city in our sample from the sample mean. The large volume of auto accident report requests in Washington state could be the result of a variety of legal and/or logistical factors, but certainly suggests the opportunity for significant efficiency gains via a coordinated statewide effort to reduce barriers to auto accident report.[14]

[13] One example of such an investment can be seen in the city of San Francisco which is currently developing a new system for releasing police incident data. Key improvements include reducing lag time from 2 weeks to 1 day and including homicide incidents.

[14] The public record request contacts in Washington state that we consulted offered a number of possible explanations, including Washington State's transparent disclosure laws (RCW 42.56), requestors that may not be aware of accident report-specific sites (such as this site for Olympia) or do not want to pay the $10.50 fee to access a report via the Washington State Patrol site, or wish to obtain all accident reports in a given time period rather than an individual accident report. Given that accident reports do not require redaction (unlike some police reports), we think that there is opportunity for efficiency gains through coordinated statewide proactive disclosure of accident reports.

CITIES SEEKING TO EFFECTIVELY PRIORITIZE OPEN DATA RELEASE SHOULD INVEST IN UNDERSTANDING DEMAND IN A LOCAL CONTEXT.

Perhaps because of this city-specific demand or public information, we did not see a clear difference in the types of data requested with and without an open data policy in place, as shown in Table 3 below. This may be because most of our cities are either in the treatment group or the control group for the full-time series, making it difficult to qualitatively assess how adopting an open data policy affects the subject of PRRs across cities.

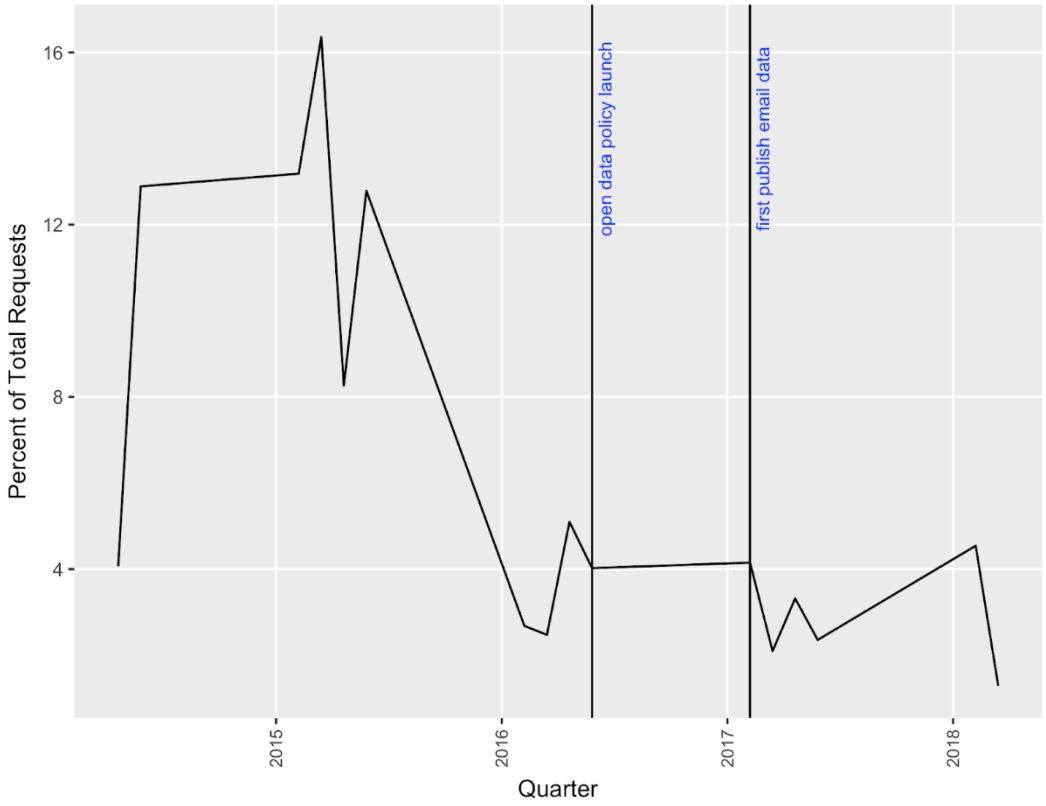
Table 3: Difference in Types of Data Requested in Cities Lacking Open Data Policy

Data Type	**% Treatment**	**% Control**	**Difference**
Auto Collision Report	5.38	8.23	-2.85
City Emails, Social Media	4.09	4.85	-0.76
Building Code Violations	3.50	4.13	-0.63
Crime Photo and Video	4.21	4.71	-0.50
Purchasing Records, Contracts	5.35	5.81	-0.46

City Government Meeting Notes	1.77	2.19	-0.42
Parcel Records, Permits, Plans	16.01	16.42	-0.42
Witness Statements	1.04	1.35	-0.31
Property Liens	5.15	5.44	-0.29
Human Services Cases	1.46	1.56	-0.11
Public Works, Utilities	3.66	3.58	0.07
Police Incident Report	15.13	14.83	0.29
Employee Benefits, Payroll	4.44	3.94	0.50
Environmental Assessment, Hazardous Materials	6.40	5.80	0.60
Checks and Deposits	1.64	1.03	0.61
Criminal Record Check	8.44	7.77	0.67
Uncategorized	5.26	4.11	1.15
Complaints to City	2.14	0.86	1.28
911/Law Enforcement Service Calls	4.94	3.37	1.56

To better understand the relationship between adopting an open data program and the topic of public record requests at a single-city level, we further explored the results in Greensboro, North Carolina. This analysis provided similarly inclusive results, with the full findings given in Annex C. Figure 4 below, which explores the change in the proportion of PRRs for city emails and social media posts over time, shows some initial evidence of a reduction in demand for a type of data after the city of Greensboro began publishing it as open data. However, these results were not conclusive and this remains an open question for future research.

Figure 4: Percent of Quarterly Requests for City Emails, Social Media Posts Over Time



Conclusion

We found clear evidence that adopting an open data program reduces the volume of PRRs that a city receives at a significant magnitude - a 30% decrease on average compared to the PRR volume received by cities that did not adopt an open data policy in 2017 and 2018. This finding makes a clear argument for open data advocates that adopting an open data program may be one of the best ways to save cities time and money in the face of growing demand for PRRs over time. Moreover, we found that cities can expect to see a return on investment in a more robust open data program with an even greater reduction in PRR volume.

Cities seeking to improve the responsiveness of their open data programs to citizen demand can use our findings on the most commonly requested data via PRR. Our results suggest that all cities could see significant reductions in PRRs by investing in releasing police incident reports and parcel records, permits, and plans as open data. Given the variation in demand that we found across cities, we would encourage city officials to replicate our methodology on the corpus of PRRs in their city to understand local demand.

Our analysis also revealed a number of areas for further research. We did not see a significant effect of the adoption of an open data policy on the diversity of requesters, average time to complete requests, or the type of information requested. In all cases, our analysis was hamstrung by a lack of available data; very few cities publish information on the PRR requester (and even when they do it is limited to their name or email) or the request completion dates, and there were also very few cities in our sample for which we had the content of PRRs received before and after adoption of an open data policy. We would encourage future researchers or city officials with greater access to PRR data to replicate our methodology with more robust data.

Annexes

Annex A: Full List of Cities Included in Study

City	State	Population	PR Portal Provider/Link	Data Access[15]	PRR Content	# Raw PRRs[16]	# Months	# Clean PRRs[17]
Albuquerque	NM	559,277	Next Request	Web scrape	N	12,236	34	n/a
Alexandria	VA		GovQA	PRR	N	10,938	94	n/a
Arlington	WA	19,112	GovQA	Download	Y	942	27	887
Asheville	NC	89,121	Seamless Gov	PRR	N	75	3	n/a
Bainbridge Island	WA	24,404	Next Request	Web scrape	Y	712	25	633
Belleville	IL	41,906	GovQA	PRR	N	1,455	64	n/a
Bellevue	WA		GovQA	PRR	N	459	6	n/a
Boulder County	CO	319372	GovQA	PRR	Y	99	8	97
Cape Coral	FL	179,804	GovQA	PRR	N	2,500	2	n/a
Cathedral City	CA	54,056	GovQA	Download	Y	337	9	295
Clark County	WA	459,495	GovQA	PRR	N	9,175	44	n/a
Clearwater	FL	114,361	GovQA	PRR	Y	23,722	45	15,409
Corona	CA		Custom	Download	N	21,151	28	n/a

[15] This refers to how the author obtained the raw PRR data from the portal.

[16] This indicates whether the accessible PRR data included sufficient information on the content of the PRR to be used in the topic analysis portion of the research.

[17] Refers to the number of PRRs used for text analysis after the data cleaning protocol removed records with insufficient information.

Dayton	OH		Custom	PRR	Y	637	20	316
Denton	TX		GovQA	PRR	Y	2580	17	2407
El Dorado County	CA	184,452	GovQA	Download	N	88	21	n/a
Everett	WA	109,043	GovQA	PRR	Y	10,258	15	9,159
Fort Collins	CO		Custom	PRR	Y	116	52	107
Fort Worth	TX		GovQA	PRR	N	36,097	57	n/a
Galveston	TX	50,550	GovQA	PRR	N	3,514	40	n/a
Greensboro	NC		Custom	PRR	Y	4,186	46	1,933
Hayward	CA		Custom	PRR	Y	442	96	431
Joliet	IL	148,262	GovQA	Download	N	1,033	11	n/a
Kirkland	WA	87,701	GovQA	PRR	Y	9,011	32	7,440
Laredo	TX		GovQA	PRR	N	6,825	71	n/a
Las Cruces	NM	101,759	Next Request	Web scrape	Y	679	11	667
Las Vegas	NV	632,912	GovQA	Download	N	7,852	16	n/a
Lynnwood	WA	38,092	GovQA	Download	Y	260	6	219
Mercer Island	WA	25,134	Next Request	Web scrape	Y	421	9	395
Miami	FL	453,579	Next Request	Web scrape	Y	3,092	38	3,010
Middleborough	MA	24,782	Next Request	Web scrape	Y	12	5	11
New Orleans	LA	391,495	Next Request	Web scrape	Y	3,428	23	3,372

Oakland	CA	420,005	Next Request	Web scrape	Y	11,814	23	7,793
Oklahoma City	OK		Custom	PRR	Y	9,076	22	377
Olympia	WA	51,202	GovQA	PRR	Y	7,700	50	6,540
Palo Alto	CA	67,024	GovQA	PRR	Y	1,045	34	981
Peoria	AZ		Custom	PRR	Y	862	11	794
Providence	RI	179,219	Next Request	Web scrape	N	2,252	35	n/a
Pullman	WA	33,282	GovQA	Download	Y	3,150	36	2,352
Rancho Cucamonga	CA	176,534	GovQA	Download	Y	32	6	31
Redmond	WA	62,458	GovQA	PRR	Y	6,528	33	6,119
Renton	WA	100,953	GovQA	PRR	Y	596	13	579
Riverside	CA	324722	GovQA	Download	N	1,589	29	n/a
Sacramento	CA	495,234	GovQA	Download	Y	1,028	17	890
Salt Lake City	UT	193,744	GovQA	Download	N	1,203	39	n/a
San Francisco	CA	870,887	Next Request	Web scrape	Y	1,681	7	1,598
San Mateo	CA	103,959	GovQA	Download	N	37	12	n/a
Tukwila	WA	20,033	GovQA	PRR	Y	4,190	17	3,696
Vallejo	CA	121,299	Next Request	Web scrape	N	350	25	338
Washington	DC	693,972	Custom	Download	N	8,074	31	n/a
West Sacramento	CA	52,981	Next Request	Web scrape	Y	898	39	859

Winchester	VA	27,516	GovQA	PRR	Y	179	18	174

ANNEX B: DETAILED METHODOLOGY

Research Questions 1-4 Methods: Panel Data Analysis

For research questions 1-4 outlined in the Research Questions section above, we use a panel model for our analysis. We consider three different panel model specifications to determine which to use for our analysis:

1. Pooled OLS
2. Random Effects
3. Fixed Effects

We treat pooled OLS as the baseline model as it relies on the assumption that there is no correlation between the city-specific unobserved error and the dependent variable, while we suspect such correlation occurs. The results from the pooled OLS model are given below:

```
## Pooling Model
##
## Call:
## plm(formula = count ~ policy, data = data, model = "pooling",
##     index = c("city_x", "month_year"))
##
## Unbalanced Panel: n = 52, T = 2-96, N = 1472
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -162.5889 -127.1680  -75.1680    7.5163 1109.8320
##
## Coefficients:
##             Estimate Std. Error t-value Pr(>|t|)
## (Intercept) 133.1680     6.1393 21.6912  < 2e-16 ***
## policy       30.4208    12.1311  2.5077  0.01226 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    60928000
## Residual Sum of Squares: 60669000
## R-Squared:      0.0042596
## Adj. R-Squared: 0.0035822
## F-statistic: 6.28841 on 1 and 1470 DF, p-value: 0.01226
```

We then performed the Breusch and Pagan (1980) Lagrange Multiplier test which assesses the viability of imposing the assumption of equal variance in the city error terms, testing the null hypothesis that the data can be pooled in the OLS model specification.[18] We see that the p-value causes us to reject the null hypothesis, concluding that pooled OLS will yield biased estimates.

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan) for unbalanced panels
##
## data:  count ~ policy
## chisq = 12638, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Both random effects and fixed effects address the issue of serial correlation among the city-specific error term. Fixed effects uses within-group deviations to control for the time-invariant city-specific effects while random effects does not assume that the effects of the time-

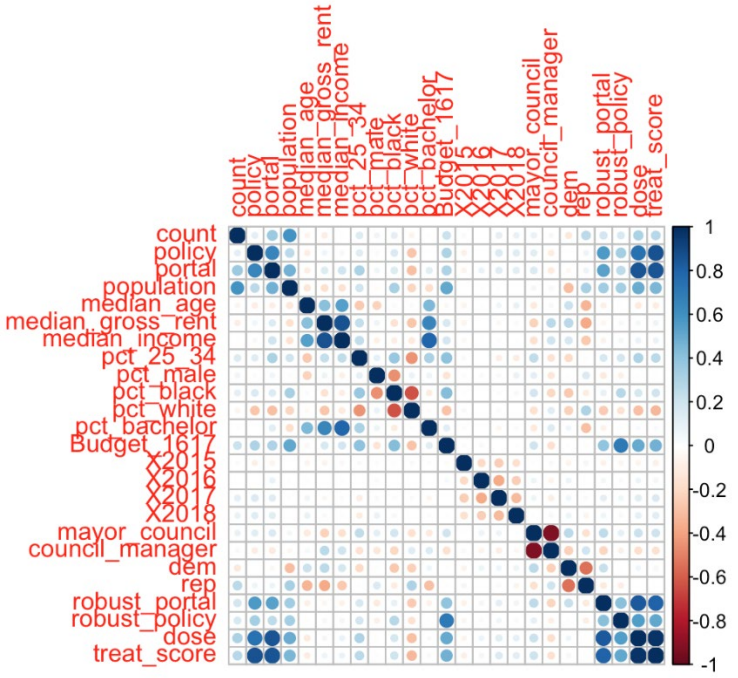
[18] Breusch, T. S. and Pagan, A. R., (1980), The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics, Review of Economic Studies, vol. 47, n1, pp. 239- 53.

invariant variables are the same and allows them to have their own starting values. Random effects relies on the assumption that the city-specific effects are independent of all explanatory variables in all time periods. Note that the Fixed Effects model will not enable us to use any of the city-specific covariates as they are time-invariant as well as forcing us to lose a number of cities from our sample in which there is no variation in the treatment status during the time period for which we have PRR data.

We ran the model with both random and fixed effects and then performed the Hausman test to formally test this assumption by identifying if there are statistically significant differences in the coefficients on the time-varying explanatory variables.[19] Given the p-value of 0.194 in the Hausman test, we fail to reject the null hypothesis and therefore use the Random Effects model.

[19] Woolridge, J. M. (2012). Introductory econometrics: A modern approach. Thomson/ South-Western. Cengage Learning5.

```
##  Hausman Test
##
## data:  count ~ policy
## chisq = 1.6869, df = 1, p-value = 0.194
## alternative hypothesis: one model is inconsistent
```



To determine which covariates to include in the final random effects model, we first calculate the pairwise correlation between all of the potential covariates in our analysis, as shown at left. We identify several groups of covariates that have strong correlations and measure a shared phenomenon:

- population, Budget_1617
- median_income, pct_bachelor, median_gross_rent
- dem, rep
- pct_black, pct_white
- mayor_council, council_manager

To avoid collinearity, we use one covariate from each group in our final model. To identify which covariate in each set to use, we try all possible combinations of covariates and pick the model with the highest R-squared. The best model (henceforth referred to as 'main model') is

given in the results section above. The results of our models are presented in Annex C: Detailed Results.

Research Questions 5-6 Methods: Text Analysis

To answer questions 5-6, we use Latent Dirichlet Allocation (LDA), a generative statistical model within natural language processing that groups documents (in our case PRRs) into a number of topics determined by the user. Before the raw documents can be effectively grouped into categories by the LDA model, the raw data must first be cleaned to remove noise, group like words as the same, and account for context-specific meaning. The full steps of our data cleaning protocol are outlined below and can be found in our iPython Notebook.

1. Remove common phrases in PRRs (such as 'public record request' or 'thank you')
2. Replace meaningful acronyms and number sequences (such as police department names or 311) with full words.
3. Remove all digits
4. Replace hyphens and slashes with spaces (separate hyphenates into component words)
5. Remove all punctuation
6. Turn raw text for each PRR into list of words
7. Remove words that are found in the proper names dictionary[20]
8. Convert all words to lowercase
9. Stem words by replacing words with root (eg. arrested and arresting would both be replaced with "arrest"). We improved performance of this process by using a function that is sensitive to the part of speech of the word.
10. Remove empty entries in word list
11. Remove stop words such as 'the' or 'an'
12. Remove all whitespace characters such as '\n', the new line character

[20] The proper names dictionary was created by combining the lists of the 1000 most popular baby names by year provided by the Social Security Administration from 1950-2017 with a list of common last names from the U.S. Census Bureau, compiled by FiveThirtyEight and accessed via data.world.

13. Remove noise words such as numeric suffixes, street abbreviations, state abbreviations, state names, numbers, single letters, month abbreviations, city names, common record request words (eg. 'please') etc.
14. Create a list of all bigrams (two-word phrases made up of sequential words) from each list of words. Identify common two-word phrases (such as 'police_report') and include those phrases in the final list of words that will be fed to the algorithm
15. Remove entries that have a word list of length 0

At the end of this cleaning process, we had dropped from 110,063 initial observations to 89,145 clean observations. This reduction represents the removal of observations that did not have enough information to be processed by the LDA model. This may be a PRR that simply contains a person's name or that just says "public record request."

There are several key parameters of the LDA model that affect the results, most critically the number of topics to create and the number of passes the algorithm performs to select the topics.

We started off with a baseline of 57 topics, which represents the average number of categories created by the small number of cities that published their own categorization, typically corresponding to the department that the request was routed to. To determine the optimal number of passes, test a 40, 60, and 80 topic models (centering on 60 given our baseline) with 20, 40, and 60 passes each. We want to find the smallest number of passes such that nearly all cases converge by the last pass of the model - achieving near-total convergence as efficiently as possible. After running this test, we find that all of our numbers of topics converge nearly completely by the end of 20 passes. We therefore run our remaining model tests with 30 passes to be safe. We then run the LDA model with 30 passes on a variety of numbers of topics ranging from 20-80. Ultimately, the 60 topic model seems to have the best performance judged qualitatively based upon the similarity of the PRRs it groups together and the coherence of the topics. We therefore use 60 topics for our LDA models.

Even after all of the data cleaning, we still have a number of observations that have a very small number of words for training. Our goal is to exclude observations with a small number of words that are not meaningful (such as a rare proper name that was not removed earlier) but keep observations with a small number of common words that can still be accurately categorized (such as ['police', 'police_report']. We try a couple of other restrictions on the observations included in model training to accomplish this goal and improve the performance of our model:

1. Remove all words that have an overall corpus frequency count (OCFC) of 1[21]
2. Remove all words that have an OCFC less than 10
3. Remove cities with average mash length less than 4 (Greensboro, Dayton, and Oklahoma City)
4. Remove all words with OCFC of 1 and all observations with mash length 2 or less
5. Remove all words with OCFC under 10 and all observations with mash length 2 or less
6. Remove all words with OCFC of 1 and all observations with mash length 3 or less
7. Remove all words with OCFC under 10 and all observations with mash length 3 or less
8. Remove all observations with mash length 3 or less and a total OCFC across all words of 100 or less
9. Remove all observations with mash length 3 or less and a total OCFC across all words of 1000 or less
10. Remove all observations with mash length 3 or less and a total OCFC across all words of 2000 or less
11. Remove all observations with mash length 3 or less, a total OCFC across all words of 1000 or less, and an average OCFC across all words of 500 or less

[21] This is the number of times a given word appears across all 89,145 cleaned PRRs.

Ultimately, we select option 10 because it struck the best balance of preserving the most observations while delivering strong results. With this last round of data cleaning, our final model uses 79,990 PRRs.

Robustness Checks

We apply a number of robustness checks to our main model to assess whether our fundamental results change. First, we run our main model with the raw count as the dependent variable to understand the relationships between our independent variables and raw PRR count. Second, we run our main model with several different subsamples that remove outliers do determine if these outliers are driving our observed results:

- Trimming Cape Coral (which had two months of data)
- Trimming all cities with under 10 months of data
- Trimming all cities with over 90 months of data
- Trimming all observations prior to 2011
- Trimming all cities with under 10 or over 90 months of data

Finally, we apply a number of further robustness checks to our main model. First, we test the robustness of our main model specification by using clustering to obtain robust standard errors and test statistics. Second, we will also run our model specifications using a log of the outcome variable to control for the fact that some cities receive more PRRs than others for reasons beyond adopting an open data program. We will also run each of our model specifications with a lagged explanatory treatment variable to test whether our findings are robust to the hypothesis that there is a lag between when an open data policy is implemented and when citizens would change their behavior (eg. looking for data on an open data portal rather than submitting a public record request). We will run each of our model specifications with three and six-month lags on the treatment variable. Finally, we will run our main model with inverse propensity score weighted observations to control for differences between the cities in treatment and control groups.

ANNEX C: STUDY DESIGN

Covariates Selection

The following covariates were considered for inclusion in our analysis due to their likely correlation with both a city's decision to adopt an open data policy, and the number and topic of PRRs that cities receive (for example, cities with larger populations generally receive more PRRs than cities with smaller populations). However, because many of these variables are also strongly correlated with each other (eg. the Mayor-Council and Council-Manager variables are mutually exclusive, or perfectly negatively correlated), we had to drop some variables to avoid multicollinearity in our model. We identified groups of collinear variables which generally measure the same phenomenon (e.g., population and budget are both capture the size of a city) and then tested all combinations of covariates that include one covariate from each group (plus all non-collinear covariates and the treatment variable) and selected the model that produced the best fit. The "final model" column captures whether this variable was included in the final model.

Covariate	**Source**	**Final Model**
Population	ACS[22]	Y
Median age	ACS	Y
Median gross rent	ACS	N
Median income	ACS	N

[22] American Community Survey 5-year Data (2009-2016) accessed via Census API https://www.census.gov/data/developers/data-sets/acs-5year.html

% of population between 25 and 34	ACS	Y
% of population that is male	ACS	Y
% of population that is black	ACS	N
% of population that is white	ACS	Y
% of population with a bachelor's degree or higher	ACS	Y
FY 2016-2017 budget	WWC[23], City Websites	N
Year dummy variables (2015-2018)	Calculated	Y
Mayor-Council government structure dummy variable	WWC, City Websites	Y
Council-Manager government structure dummy variable	WWC, City Websites	N
State leans Democratic or is strong Democratic dummy variable	WWC	N
State leans Republican or is strong Republican dummy variable	WWC	Y

All of these variables are from the 2016 5-year American Community Survey (ACS). Because the most recent ACS data available is from 2016, we treat these estimates as time-invariant and use the 2016 data for all time periods in our study. Given the relatively small time-window considered in our study (35 of the 46 cities in our sample have data beginning in 2015 or later) and the general consistency of these measures over short time periods, we feel that treating these variables as city-specific and time-invariant makes sense.

We also include data on the city government structure and 2016-2017 budget to capture the effects of city governance and capacity on public record requests.[24] This data is drawn from the

[23] Information published by Bloomberg What Works Cities program on member cities

[24] Mayor-Council, Mayor-Commission, Council-Manager or Commissioner-Manager

What Works Cities (WWC) program as well as the websites of cities that are not included in the WWC cohort. Finally, we will include year dummy variables to control for general trends in demand for government data over time, or exogenous events that may increase demand across all cities (such as the 2016 elections).

Difference of Means of Covariates for Treatment and Control Groups

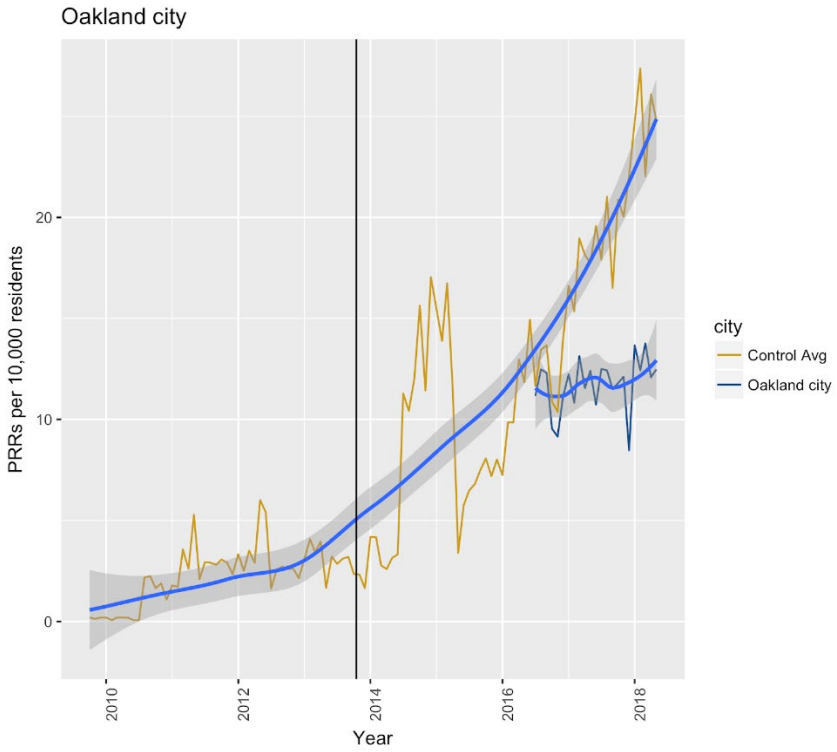
	Control	**Treatment**
population	142476.84	316906.43
median_age	37.22	34.90
median_gross_rent	1114.00	1170.19
median_income	31200.68	30951.10
pct_25_34	14.73	16.62
pct_male	49.24	48.98
pct_black	9.42	13.73
pct_white	71.40	64.09
pct_bachelor	14.38	15.45
mayor_council	0.32	0.29
council_manager	0.52	0.67
dem	0.68	0.48
rep	0.06	0.19

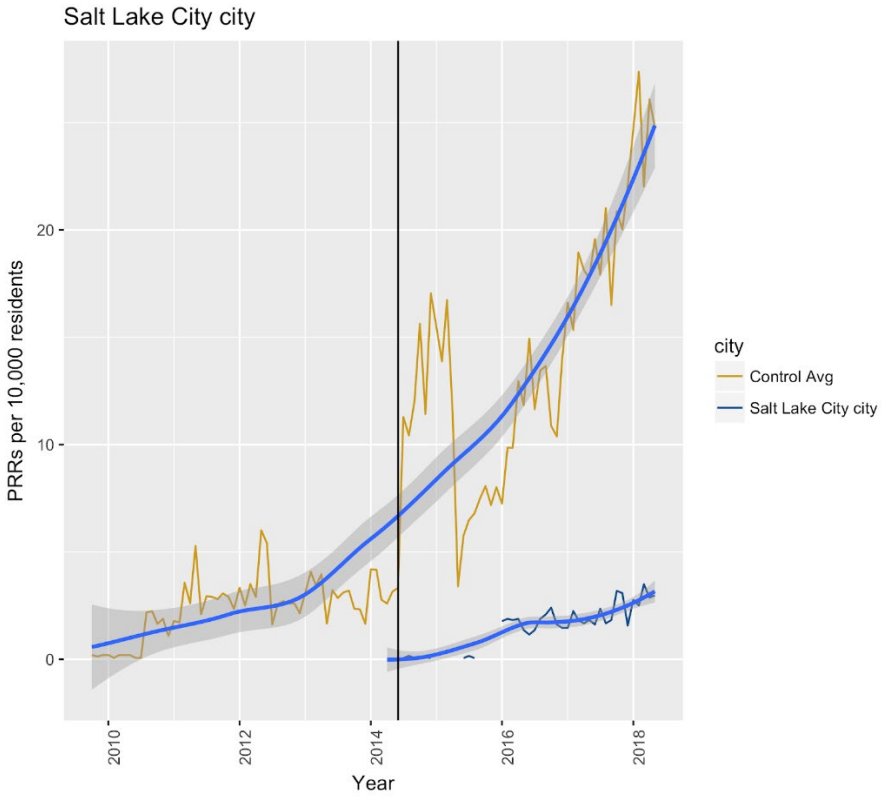
Difference of Means of Covariates for Mayor-Council Cities

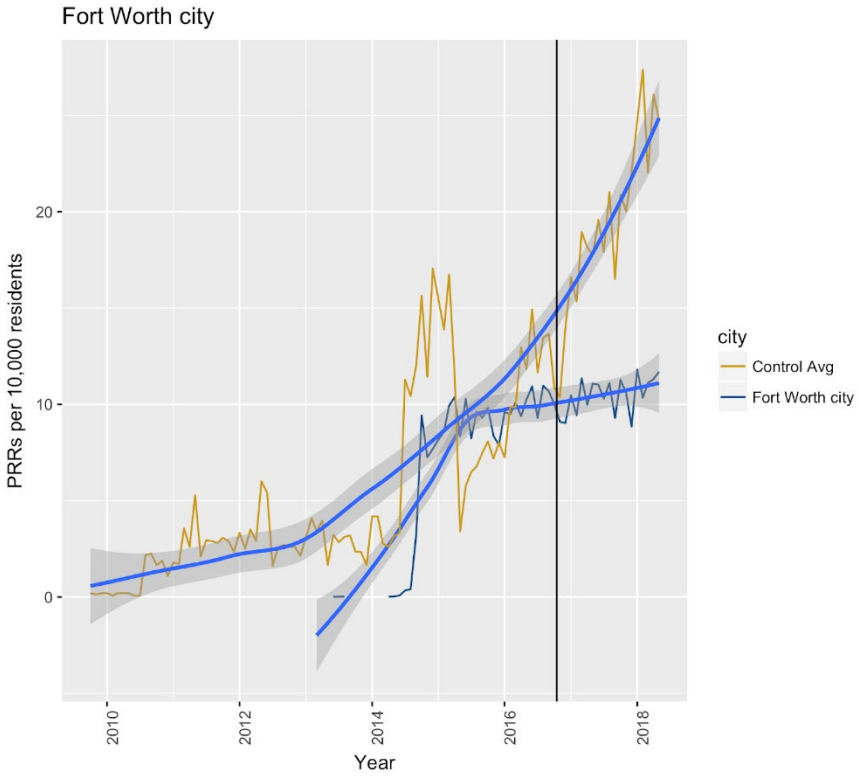
	Other	**Mayor Council**
population	194251.28	254923.19
median_age	36.98	34.73
median_gross_rent	1163.97	1075.31

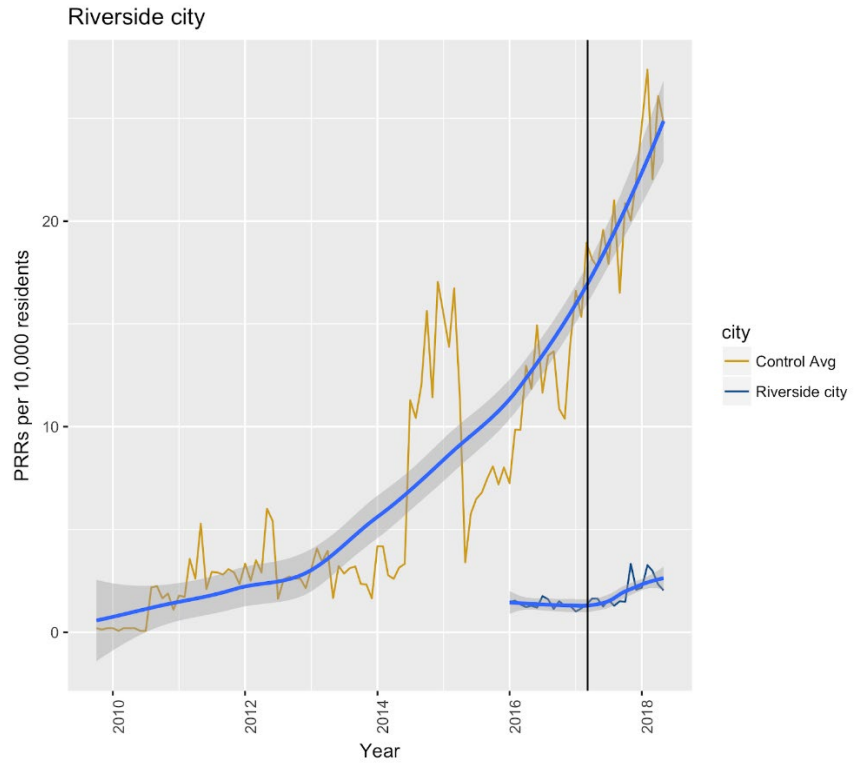
median_income	31996.75	29081.94
pct_25_34	14.48	17.77
pct_male	49.02	49.40
pct_black	9.16	15.66
pct_white	72.36	59.64
pct_bachelor	15.00	14.40
dem	0.50	0.81
rep	0.14	0.06

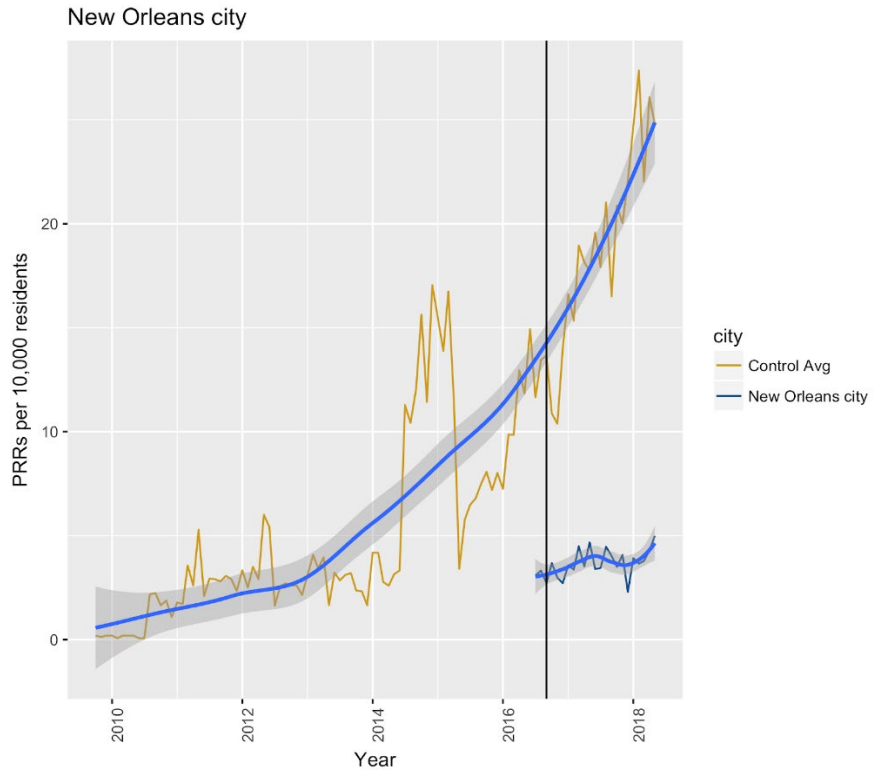
ANNEX D: CITY-SPECIFIC PRR VOLUME PLOTS

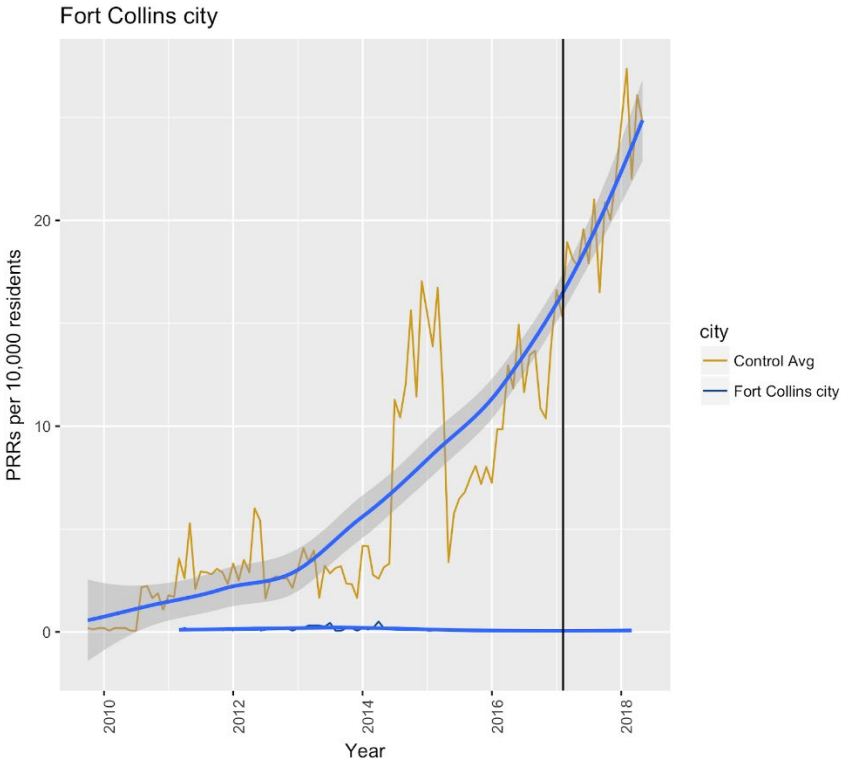

Oakland city
PRRs per 10,000 residents
20
10
0
2010
2012
2014
2016
2018
Year
city
Control Avg
Oakland city


Salt Lake City city
PRRs per 10,000 residents
20
10
0
2010
2012
2014
2016
2018
Year
city
Control Avg
Salt Lake City city


Fort Worth city
PRRs per 10,000 residents
20
10
0
2010
2012
2014
2016
2018
Year
city
Control Avg
Fort Worth city


Riverside city
PRRs per 10,000 residents
20
10
0
2010
2012
2014
2016
2018
Year
city
Control Avg
Riverside city


New Orleans city
PRRs per 10,000 residents
20
10
0
2010
2012
2014
2016
2018
Year
city
Control Avg
New Orleans city


Fort Collins city
PRRs per 10,000 residents
20
10
0
2010
2012
2014
2016
2018
Year
city
Control Avg
Fort Collins city

ANNEX E: DETAILED RESULTS

To understand how the marginal impact of adopting an open data policy varies across our covariates, we run both the main model and the raw count DV model with all of the covariates interacted with the policy variables. In the main model, the only interaction term that is statistically significant is the interaction between policy and the 2018 dummy, which suggests that the marginal impact of adopting a policy is significantly greater in 2018, the last year in our time-series. In the raw count model, a number of interaction terms are significant. The significance of the 2017 and 2018 interaction terms reinforces the result about how the effect of adopting a policy grows over time. We also see that the interaction term between policy and policy_months is also significant and negative, indicating that for each additional month a policy is in place, a city will receive approximately 4 fewer PRRs on average.

Interacted Model with population-adjusted count DV

```
Coefficients:
                      Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)         -70.282490  118.185221 -0.5947   0.55215
policy              -35.926425  111.365005 -0.3226   0.74704
X2015                -1.874995    1.994745 -0.9400   0.34739
X2016                 2.283004    2.380041  0.9592   0.33760
X2017                 7.328062    2.891627  2.5342   0.01137 *
X2018                16.109031    3.369507  4.7808 1.924e-06 ***
mayor_council        14.826231    7.287297  2.0345   0.04208 *
pct_bachelor         23.777937   65.894805  0.3608   0.71827
median_age            0.036113    0.856189  0.0422   0.96636
pct_25_34             4.541664  144.066146  0.0315   0.97486
months                0.045325    0.051046  0.8879   0.37473
rep                   5.266371    9.910806  0.5314   0.59524
pct_white            -9.141829   22.227132 -0.4113   0.68092
pct_male            146.836126  226.511891  0.6482   0.51693
policy:X2015          1.121180    6.736791  0.1664   0.86784
policy:X2016         -3.052757    6.806822 -0.4485   0.65387
policy:X2017         -8.651033    7.609051 -1.1369   0.25575
policy:X2018        -17.460567    8.399582 -2.0787   0.03782 *
policy:mayor_council  -7.150261   11.153437 -0.6411   0.52157
policy:pct_bachelor -116.877181  101.747922 -1.1487   0.25087
policy:median_age      2.044758    1.464092  1.3966   0.16275
policy:pct_25_34      42.134037  298.497227  0.1412   0.88777
policy:months          0.062042    0.135465  0.4580   0.64702
policy:rep            -3.931964    8.606193 -0.4569   0.64783
policy:pct_white      20.477222   30.311079  0.6756   0.49942
policy:pct_male      -71.482775  226.088251 -0.3162   0.75192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    299280
Residual Sum of Squares: 268430
R-Squared:      0.10309
Adj. R-Squared: 0.087581
F-statistic: 6.64724 on 25 and 1446 DF, p-value: < 2.22e-16
```

To answer the research question of whether the intensity of a city's open data program affects the observed effect on PRR volume, we run our main model with the following treatment variables: 1) portal, 2) robust portal, 3) robust policy, and 4) an aggregate treatment score of the level of treatment in each city.[25] The regression output is given below:

I. Regression Results: Portal Treatment Variable

```
Coefficients:
                Estimate Std. Error t-value Pr(>|t|)
(Intercept)   -91.037826 102.286576 -0.8900 0.373598
portal          0.882439   1.723975  0.5119 0.608824
pct_bachelor  -24.546032  53.642973 -0.4576 0.647321
rep            -0.562465   8.496050 -0.0662 0.947225
pct_white       0.738784  18.982646  0.0389 0.968960
mayor_council  12.374445   6.177884  2.0030 0.045360 *
median_age      0.431739   0.732021  0.5898 0.555423
X2015          -2.208651   1.936224 -1.1407 0.254182
X2016           1.004373   2.290176  0.4386 0.661047
X2017           3.716206   2.742019  1.3553 0.175538
X2018           8.983004   3.151127  2.8507 0.004423 **
months          0.063434   0.048600  1.3052 0.192023
pct_male      175.697173 191.208693  0.9189 0.358312
pct_25_34     -34.733002 115.843907 -0.2998 0.764353
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    301810
Residual Sum of Squares: 279110
R-Squared:      0.075244
Adj. R-Squared: 0.066998
F-statistic: 9.12262 on 13 and 1458 DF, p-value: < 2.22e-16
```

II. Regression Results: Robust Portal Treatment Variable

[25] 1 point is assigned to a city for whether they have a policy, robust policy, portal, and robust portal in place. The treatment score ranges from 0-4.

```
Coefficients:
                Estimate Std. Error t-value Pr(>|t|)
(Intercept)   -67.733178 102.071964 -0.6636 0.507062
robust_portal  -7.149945   3.920234 -1.8239 0.068378 .
pct_bachelor  -23.009429  53.119874 -0.4332 0.664962
rep            -0.789879   8.416158 -0.0939 0.925239
pct_white      -1.174214  18.828449 -0.0624 0.950282
mayor_council  12.359851   6.120530  2.0194 0.043627 *
median_age      0.412768   0.725514  0.5689 0.569490
X2015          -2.130861   1.905818 -1.1181 0.263716
X2016           1.596812   2.227666  0.7168 0.473606
X2017           4.490706   2.672299  1.6805 0.093081 .
X2018           9.836670   3.076804  3.1970 0.001418 **
months          0.058972   0.048529  1.2152 0.224490
pct_male      124.185128 191.248410  0.6493 0.516221
pct_25_34      -2.043668 115.545481 -0.0177 0.985891
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    302030
Residual Sum of Squares: 278710
R-Squared:      0.077217
Adj. R-Squared: 0.068989
F-statistic: 9.38445 on 13 and 1458 DF, p-value: < 2.22e-16
```

III. Regression Results: Robust Policy Treatment Variable

```
Coefficients:
                Estimate Std. Error t-value Pr(>|t|)
(Intercept)   -80.547829 104.110052 -0.7737 0.439246
robust_policy  -9.231914   5.394604 -1.7113 0.087234 .
pct_bachelor  -27.097785  54.525308 -0.4970 0.619281
rep            -1.347598   8.654403 -0.1557 0.876281
pct_white       1.935974  19.310481  0.1003 0.920156
mayor_council  11.722577   6.288622  1.8641 0.062509 .
median_age      0.475809   0.744819  0.6388 0.523037
X2015          -1.830215   1.911470 -0.9575 0.338478
X2016           1.614379   2.234565  0.7225 0.470129
X2017           4.483778   2.681518  1.6721 0.094718 .
X2018           9.861239   3.089314  3.1920 0.001443 **
months          0.062898   0.048625  1.2935 0.196031
pct_male      141.036586 195.091550  0.7229 0.469842
pct_25_34      -2.221371 118.553272 -0.0187 0.985053
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    301350
Residual Sum of Squares: 278210
R-Squared:      0.076801
Adj. R-Squared: 0.06857
F-statistic: 9.32906 on 13 and 1458 DF, p-value: < 2.22e-16
```

IV. Regression Results: Treatment Score Treatment Variable

```
Coefficients:
                Estimate Std. Error t-value  Pr(>|t|)
(Intercept)   -73.758776 100.308233 -0.7353 0.4622621
treat_score    -2.041421   0.984858 -2.0728 0.0383656 *
pct_bachelor  -18.220069  52.472440 -0.3472 0.7284677
rep            -0.045594   8.300846 -0.0055 0.9956182
pct_white      -1.568156  18.583629 -0.0844 0.9327629
mayor_council  11.768230   6.043559  1.9472 0.0516983 .
median_age      0.405703   0.715870  0.5667 0.5709869
X2015          -1.523062   1.920223 -0.7932 0.4278082
X2016           2.311388   2.273620  1.0166 0.3095071
X2017           5.508927   2.749966  2.0033 0.0453330 *
X2018          10.984565   3.161738  3.4742 0.0005274 ***
months          0.064174   0.048303  1.3286 0.1841956
pct_male      134.833910 187.813656  0.7179 0.4729258
pct_25_34       0.288897 113.965177  0.0025 0.9979777
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    302390
Residual Sum of Squares: 278830
R-Squared:      0.077919
Adj. R-Squared: 0.069698
F-statistic: 9.47699 on 13 and 1458 DF, p-value: < 2.22e-16
```

We run significance tests on the difference in the magnitude of the policy and robust policy coefficients and the portal and robust portal coefficients. For policy vs. robust policy, one-tailed and two-tailed p-value is 0.3126 and 0.6252 respectively. This means we fail to reject the null hypothesis that the difference between the coefficient on policy and the coefficient on robust policy is zero. On the other hand, for portal vs robust portal, the one-tailed p-value and two-tailed p-value is 0.0304 and 0.0608 respectively. This means we reject the null hypothesis that the effect of robust portal is greater than portal at the 5% confidence level and the null hypothesis that the effect is the same at the 10% level.

As discussed above, we run several robustness checks to verify our results. First, in addition to running our main model given above with the population-adjusted dependent variable, we also run the model with raw PRR count as the dependent variable. The results of this model are given below. As expected, the coefficient on population is positive and highly significant. The policy term becomes insignificant, perhaps because of the confounding effect of population (as shown in Annex A Table 3, cities in the treatment group have higher populations).

Robustness Check 1: Raw Count Dependent Variable

```
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = count ~ policy + population + pct_bachelor + rep +
    pct_white + mayor_council + median_age + X2015 + X2016 +
    X2017 + X2018 + months + pct_male + pct_25_34, data = data,
    model = "random", index = c("city_x", "month_year"))

Unbalanced Panel: n = 52, T = 2-96, N = 1472

Effects:
                 var std.dev share
idiosyncratic 11654.1   108.0 0.346
individual    22003.9   148.3 0.654
theta:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.5424  0.8560  0.8842  0.8739  0.9041  0.9259

Residuals:
    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-523.23  -38.07   -0.20    0.91   38.60  716.18

Coefficients:
               Estimate  Std. Error t-value  Pr(>|t|)
(Intercept)  -7.9707e+02  8.8371e+02 -0.9020  0.367226
policy        1.9977e+01  1.6519e+01  1.2093  0.226742
population    3.4064e-04  1.1253e-04  3.0271  0.002512 **
pct_bachelor -7.7314e+02  4.6674e+02 -1.6565  0.097841 .
rep           6.6768e+01  7.5623e+01  0.8829  0.377435
pct_white     2.0728e+02  1.6670e+02  1.2434  0.213907
mayor_council 3.8040e+01  5.3385e+01  0.7126  0.476235
median_age    1.2296e+01  6.3833e+00  1.9262  0.054275 .
X2015         3.4387e+01  1.5089e+01  2.2789  0.022819 *
X2016         8.0269e+01  1.7742e+01  4.5241 6.557e-06 ***
X2017         1.2738e+02  2.1563e+01  5.9076 4.315e-09 ***
X2018         1.5455e+02  2.4846e+01  6.2202 6.474e-10 ***
months        1.3817e-01  3.8704e-01  0.3570  0.721159
pct_male      1.1147e+02  1.6500e+03  0.0676  0.946148
pct_25_34     1.2961e+03  1.0198e+03  1.2709  0.203952
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    20479000
Residual Sum of Squares: 17121000
R-Squared:      0.16407
Adj. R-Squared: 0.15604
F-statistic: 20.4165 on 14 and 1457 DF, p-value: < 2.22e-16
```

We also tested methods of trimming outliers to identify whether this changes our overall results. Specifically, we ran the main model with the population-adjusted DV and raw DV as follows:

- Trimming Cape Coral (which had two months of data)
- Trimming all cities with under 10 months of data
- Trimming all cities with over 90 months of data
- Trimming all observations prior to 2011
- Trimming all cities with under 10 or over 90 months of data

Ultimately, none of the options above significantly improved the fit of our model or changed our results.[26] Therefore, we chose to proceed with the full data set for all models.

Finally, we run our main model using several different checks to ensure the robustness of our findings: 1) logged count, 2) clustered standard errors, 3) treatment variables (policy/portal) lagged by 3 and 6 months, and 4) inverse propensity score weighting. The regression output is

[26] The model output can be found on our github page.

provided on github, which in all cases confirms our overall result that adopting an open data policy yields a significant decrease in the number of PRRs a city receives per 10,000 residents. We do find that in some cases the coefficient is lower than our main model providing conservative estimate of the average treatment effect.

Text Analysis

We tested a number of different methods to calculate the popularity of the 60 different topics generated by the LDA algorithm (given below). Those methods are:

1. **Winner Take All (WTA):** Only the topic with the highest share of a PRR (or topic composition score) receives "points." The value of the topic composition score for the winning topic is added to its total score.
2. **Proportional Assignment with Threshold (Prop):** Every topic that is contained in a PRR receives "points" provided that the topic composition score is above a given threshold. The value of the topic composition score for each topic above the threshold is added to its total score. We tested low (.2) and high (.5) thresholds.
3. **Dampened Popularity with Threshold (DP)**: The total scores for each city are calculated with either the winner take all (DP W) or proportional assignment (DP P) method. We then take the log of topic totals for each city, and sum the logged totals for each topic across cities to calculate the total score for each topic. We apply the low (.2) and high (.5) thresholds for both the DP W and DP P methods.

Ultimately, we choose the Dampened Popularity with proportional assignment and the .2 threshold as the final method because the dampened popularity reduces the amount a single city influences the overall results. Proportional assignment better reflects reality (a PRR may address multiple topics) and makes sure that data types that are associated with a large number of topics (such as police incident reports) aren't penalized in calculating data type popularity as the topic composition score will likely be split across a large number of topics.

The final list of topics generated by the LDA algorithm with the corresponding data type and category is given in the table below:

#	Topic Key Words	Data Type	W	P .2	P .5	DP W .2	DP W .5	DP P .2	DP P .5
0	[('complaint', 0.1985935), ('log', 0.14190955), ('phone', 0.08581352), ('search', 0.083066285), ('warrant', 0.03568498), ('estate', 0.032001145), ('involved', 0.026251603), ('quality', 0.02599625), ('real', 0.025753625), ('ensure', 0.019929796)]	Complaints to City	42	45	50	43	41	46	41
1	[('associate', 0.14829391), ('client', 0.11122169), ('photograph', 0.10852045), ('office', 0.08708323), ('follow', 0.061714146), ('emergency', 0.042823374), ('transfer', 0.033178225), ('color', 0.024401158), ('restaurant', 0.02243146), ('time', 0.018097397)]	Police Incident Report	38	41	22	45	42	51	42
2	[('name', 0.15037733), ('victim', 0.084349126), ('crime', 0.082644366), ('officer', 0.07667519), ('domestic', 0.041578013), ('violence', 0.03346935), ('injury', 0.030954132), ('capitol', 0.030555379), ('location', 0.027925178), ('type', 0.021750417)]	Police Incident Report	15	19	20	20	21	29	21
3	[('permit', 0.2758547), ('issue', 0.11701758), ('building', 0.07663683), ('build', 0.06414055), ('building_permit', 0.059510738), ('ref', 0.056181964), ('construction', 0.03771535),	Parcel Records, Permits, Plans	12	12	6	4	5	7	5

	('without', 0.024624562), ('administration', 0.020187119), ('since', 0.02015884)]								
4	[('list', 0.24116392), ('service', 0.123515554), ('well', 0.072589695), ('within', 0.069015294), ('person', 0.06510219), ('month', 0.06098719), ('law', 0.03411746), ('enforcement', 0.032960806), ('residence', 0.028089434), ('last', 0.026984973)]	911/Law Enforcement Service	19	20	5	22	30	24	30
5	[('show', 0.10214498), ('use', 0.08015429), ('policy', 0.058283392), ('maintenance', 0.039954595), ('year', 0.03847074), ('procedure', 0.037714776), ('parkway', 0.036550127), ('today', 0.02697065), ('operation', 0.025670737), ('old', 0.024925187)]	Public Works & Utilities	46	55	54	35	51	38	51
6	[('avenue', 0.35911864), ('drive', 0.17917691), ('account', 0.054597758), ('return', 0.032051444), ('son', 0.024467614), ('cell', 0.023386981), ('wonder', 0.019287072), ('geotechnical', 0.016777175), ('generate', 0.015776062), ('pole', 0.010756759)]	Parcel Records, Permits, Plans	35	35	25	26	33	26	33
7	[('home', 0.07116078), ('park', 0.063897714), ('house', 0.056650896), ('lot', 0.049521584), ('place', 0.048671838), ('back', 0.04694188), ('tree', 0.041362315), ('area', 0.039702497), ('residential', 0.029248808), ('land', 0.02735575)]	Property Liens	40	43	60	24	57	25	57
8	[('order', 0.06940024), ('item', 0.05680285), ('invoice', 0.04215782), ('detail', 0.03767509), ('purchase', 0.03386113), ('number',	Purchasing Records, Contracts	31	34	47	10	13	12	13

	0.029808844), ('sale', 0.029691221), ('several', 0.022504056), ('http', 0.018583613), ('commission', 0.018436963)]								
9	[('local', 0.10987223), ('international', 0.038030066), ('live', 0.03651968), ('builts', 0.035503764), ('range', 0.033208173), ('move', 0.027478758), ('contractor', 0.025758244), ('print', 0.023395741), ('approx', 0.021365745), ('shoot', 0.020594789)]	Criminal Record Check	48	51	53	44	47	50	47
10	[('arrest', 0.22508794), ('notice', 0.10653793), ('citation', 0.093008235), ('come', 0.054579236), ('say', 0.054288175), ('arrest_report', 0.030788284), ('accept', 0.023258643), ('revocation', 0.01854626), ('recently', 0.017429836), ('officer', 0.016972534)]	Criminal Record Check	21	27	13	25	17	31	17
11	[('contract', 0.12024548), ('bid', 0.08086701), ('submit', 0.0620963), ('ordinance', 0.048248757), ('subcontractor', 0.04418181), ('service', 0.042105265), ('project', 0.041909378), ('rfp', 0.036090374), ('approve', 0.03256977), ('award', 0.031801503)]	Purchasing Records, Contracts	18	22	4	8	11	8	11
12	[('camera', 0.08036688), ('entity', 0.074918844), ('release', 0.058930222), ('intend', 0.021181582), ('help', 0.021179346), ('pull', 0.019310843), ('attachment', 0.018467927), ('expedite', 0.01831748), ('appreciated', 0.0180274), ('content', 0.017454458)]	Criminal Record Check	43	44	57	56	48	60	48

13	[('call', 0.32521334), ('note', 0.088374816), ('nineoneone', 0.0701484), ('cad', 0.066656), ('stop', 0.04774236), ('frame', 0.034719706), ('dispatch', 0.03386697), ('turn', 0.019149732), ('time', 0.017925613), ('recording', 0.016269712)]	911/Law Enforcement Service Calls	17	15	14	21	18	23	18
14	[('communication', 0.0688568), ('correspondence', 0.06803626), ('staff', 0.04231121), ('meeting', 0.04045327), ('concern', 0.033581033), ('member', 0.033549283), ('note', 0.02915484), ('official', 0.028396739), ('limited', 0.023606328), ('minute', 0.020674579)]	City Gov. Meeting Notes	34	38	10	15	39	17	39
15	[('tax', 0.10910112), ('amount', 0.062115442), ('check', 0.047044717), ('payee', 0.028844092), ('employment', 0.023935616), ('performance', 0.02222605), ('initial', 0.021986058), ('great', 0.021605315), ('outstanding', 0.020619854), ('possible', 0.02048308)]	Checks and Deposits	49	52	18	42	29	47	29
16	[('hit', 0.091981694), ('mention', 0.062275108), ('next', 0.06079646), ('run', 0.054484926), ('applicant', 0.052003473), ('study', 0.049023617), ('requester', 0.039278906), ('support', 0.03417416), ('apply', 0.029796638), ('downtown', 0.025259892)]	Auto Collision Report	52	57	35	46	45	48	45
17	[('map', 0.08679456), ('possible', 0.057452887), ('identify', 0.04500377), ('free', 0.03915043), ('appreciate', 0.038111392),	Parcel Records, Permits, Plans	47	50	59	40	50	43	50

	('formal', 0.0354281), ('feel', 0.034807913), ('block', 0.028777761), ('much', 0.026496025), ('hard', 0.026162075)]								
18	[('data', 0.1092429), ('agreement', 0.0844293), ('limited', 0.08090052), ('body', 0.07428584), ('video', 0.062064104), ('cam', 0.048334192), ('room', 0.043254204), ('surveillance', 0.032963812), ('collect', 0.032702666), ('dash', 0.01876143)]	Crime Photo/ Video	33	36	1	32	36	35	36
19	[('theft', 0.22061245), ('along', 0.06716178), ('auto_theft', 0.063398145), ('parking', 0.062286988), ('auto', 0.04646996), ('pl', 0.043082036), ('southcenter', 0.035162725), ('lot', 0.024039797), ('main', 0.020844543), ('university', 0.018294595)]	Police Incident Report	28	32	2	38	22	49	22
20	[('application', 0.085657544), ('license', 0.068344794), ('payroll', 0.06762842), ('employee', 0.053203892), ('current', 0.052077144), ('business', 0.0499591), ('refer', 0.03578757), ('number', 0.031211086), ('manager', 0.026353242), ('medical', 0.025645034)]	Employee Benefits & Payroll	36	37	40	16	23	14	23
21	[('present', 0.22439376), ('history', 0.09230829), ('report_incident', 0.059092462), ('fund', 0.051201187), ('period', 0.050926022), ('time', 0.038735375), ('animal', 0.023132412), ('supplemental', 0.02311513), ('signal', 0.020913407), ('closure', 0.018397162)]	Criminal Record Check	44	42	38	48	40	33	40

22	[('vehicle', 0.16597652), ('car', 0.12360624), ('street', 0.08259031), ('intersection', 0.06479602), ('driver', 0.059679396), ('involve', 0.04571497), ('accident', 0.044755075), ('motor', 0.031106036), ('insurance', 0.02516581), ('around', 0.015663475)]	Auto Collision Report	24	25	43	18	20	20	20
23	[('involve', 0.16531257), ('investigation', 0.12932137), ('child', 0.042590715), ('usts', 0.03794015), ('division', 0.031155419), ('unresolved', 0.027624652), ('service', 0.02494772), ('training', 0.02422004), ('asts', 0.022617245), ('similar', 0.01915818)]	Human Services Cases	30	30	55	31	28	28	28
24	[('fire', 0.14456978), ('site_assessment', 0.09028297), ('environmental_site', 0.08902072), ('phase_environmental', 0.06846374), ('enforcement', 0.06332058), ('info', 0.04320781), ('light', 0.03646397), ('scene', 0.035504606), ('side', 0.033469994), ('dept', 0.02751706)]	Environ. Assess./ Hazardous Materials	37	33	16	28	31	21	31
25	[('collision', 0.16592814), ('occur', 0.14634575), ('collision_report', 0.088421896), ('cannabis', 0.0380842), ('event', 0.031146232), ('memoranda', 0.023629144), ('addition', 0.023382638), ('enterprise', 0.021961), ('good', 0.020487789), ('attempt', 0.018331006)]	Auto Collision Report	20	21	32	39	14	42	14
26	[('email', 0.05471243), ('send', 0.044262353), ('letter', 0.0327627), ('fee', 0.025692467), ('disclosure', 0.025541063), ('cost',	City Emails & Social Media	9	11	3	3	15	3	15

	0.02162389), ('mail', 0.021287313), ('following', 0.020823436), ('electronic', 0.01912893), ('time', 0.01801677)]								
27	[('demand', 0.09400265), ('past', 0.08884601), ('community', 0.076444395), ('activity', 0.06819106), ('regulation', 0.06323379), ('year', 0.046756167), ('chapter', 0.03797209), ('transcript', 0.03222767), ('every', 0.028584126), ('track', 0.026003918)]	Uncategorized	51	49	58	57	52	58	52
28	[('case', 0.37224895), ('number', 0.20965174), ('case_number', 0.046436027), ('steal', 0.043556415), ('report_case', 0.026024856), ('update', 0.019941686), ('analysis', 0.019081214), ('disposition', 0.010940141), ('view', 0.010393641), ('stolen', 0.009620244)]	Police Incident Report	8	7	9	12	9	11	9
29	[('property', 0.24129005), ('owner', 0.09174423), ('address', 0.05381202), ('lien', 0.05202367), ('research', 0.039118163), ('due', 0.029557053), ('payoff', 0.025631377), ('amount', 0.025016405), ('housing', 0.022862263), ('sidewalk', 0.022836013)]	Property Liens	22	23	8	14	27	13	27
30	[('violation', 0.12519231), ('code', 0.08656456), ('property', 0.06848214), ('open', 0.06713956), ('code_violation', 0.057894886), ('zone', 0.04222084), ('building', 0.03874377), ('fire', 0.036294296), ('unit', 0.030973408), ('apn', 0.030083403)]	Building Code Violations	7	8	26	2	2	2	2

31	[('video', 0.25615332), ('traffic', 0.16091378), ('audio', 0.11414809), ('recording', 0.058318987), ('window', 0.013530725), ('passenger', 0.01229218), ('inside', 0.01213723), ('processing', 0.011798941), ('photographs', 0.011124628), ('bac', 0.011021805)]	Crime Photo/ Video	11	9	7	19	24	18	24
32	[('llc', 0.07916555), ('sign', 0.066731155), ('district', 0.055866495), ('sheet', 0.046132334), ('hearing', 0.043102253), ('prevent', 0.038710307), ('term', 0.035818845), ('connection', 0.029579524), ('operate', 0.023355097), ('rental', 0.023327863)]	Uncategorized	54	58	48	51	60	45	60
33	[('accident', 0.39958552), ('accident_report', 0.18422489), ('auto_accident', 0.114594795), ('auto', 0.09796612), ('pedestrian', 0.019592816), ('architectural', 0.014912473), ('usaa', 0.010008935), ('everything', 0.009694831), ('fault', 0.006688628), ('instruction', 0.006613127)]	Auto Collision Report	3	3	28	13	7	19	7
34	[('location', 0.21679133), ('claim', 0.09818342), ('state', 0.0697553), ('burglary', 0.06911095), ('insured', 0.056019105), ('loss', 0.027791245), ('insure', 0.025631163), ('measure', 0.024490217), ('farm', 0.022929706), ('state_farm', 0.022389304)]	Police Incident Report	32	17	12	49	32	37	32
35	[('email', 0.26834995), ('id', 0.1306745), ('complete', 0.08180797), ('social', 0.06045802), ('post', 0.036991253), ('credit',	City Emails & Social Media	29	31	36	50	35	39	35

	0.036454022), ('approved', 0.030344106), ('medium', 0.025660833), ('exceed', 0.023591131), ('sam', 0.013001699)]								
36	[('tell', 0.05282206), ('store', 0.048417732), ('leave', 0.046139732), ('action', 0.035686594), ('door', 0.027427835), ('unknown', 0.026335958), ('speak', 0.023881935), ('officer', 0.023780545), ('hwy', 0.021363195), ('enter', 0.021137886)]	Police Incident Report	53	56	33	58	55	59	55
37	[('year', 0.15321907), ('party', 0.100014806), ('last', 0.07572979), ('apartment', 0.036389094), ('even', 0.030202717), ('happen', 0.028766967), ('sure', 0.028110817), ('possession', 0.027636135), ('investigate', 0.025434613), ('additional', 0.023297438)]	Criminal Record Check	45	39	29	41	38	32	38
38	[('address', 0.37964672), ('contact', 0.07585882), ('charge', 0.06375808), ('name', 0.051895127), ('jurisdiction', 0.043860443), ('response', 0.041658256), ('commercial', 0.039081343), ('following', 0.031024534), ('assistance', 0.029206535), ('advance', 0.02876223)]	Property Liens	41	29	44	33	34	22	34
39	[('work', 0.053015515), ('pay', 0.038354263), ('line', 0.030709907), ('review', 0.030431146), ('employee', 0.029498309), ('benefit', 0.02925975), ('exist', 0.025858663), ('recent', 0.024774328), ('total', 0.024744458), ('determine', 0.023626313)]	Employee Benefits & Payroll	39	40	37	17	43	16	43

40	[('photo', 0.22963522), ('offense', 0.06913104), ('duo', 0.047197696), ('video', 0.02510717), ('packet', 0.02099043), ('detain', 0.020919355), ('checkpoint', 0.01420033), ('mcallister', 0.010638743), ('engage', 0.009663894), ('utc', 0.008954173)]	Crime Photo/ Video	2	2	24	36	25	44	25
41	[('insurance', 0.10682117), ('reference', 0.09845432), ('type', 0.087784335), ('number', 0.07008757), ('location', 0.06628888), ('transaction', 0.06523592), ('insure', 0.06116695), ('occurrence', 0.058346972), ('occurrence_location', 0.05399189), ('type_auto', 0.037329476)]	Auto Collision Report	1	1	27	27	10	34	10
42	[('agency', 0.11959616), ('respond', 0.11368961), ('police department', 0.096744515), ('drainage', 0.048156034), ('ticket', 0.039263982), ('id', 0.025365304), ('videos', 0.025021417), ('condominium', 0.01925798), ('plumb', 0.019178534), ('pc', 0.018561114)]	Uncategorized	60	46	45	59	54	55	54
43	[('project', 0.076902196), ('plan', 0.029102111), ('management', 0.02659861), ('development', 0.025086623), ('section', 0.024244718), ('construction', 0.023521416), ('work', 0.020521961), ('improvement', 0.01656264), ('electrical', 0.015781369), ('compliance', 0.015751444)]	Parcel Records, Permits, Plans	16	16	52	5	16	4	16
44	[('plan', 0.11660539), ('site', 0.099882215), ('certificate', 0.09434634), ('occupancy', 0.08399464), ('certificate_occupancy',	Parcel Records, Permits, Plans	14	14	39	7	8	5	8

	0.07622669), ('site_plan', 0.045080233), ('final', 0.040189173), ('permit', 0.034261044), ('interested', 0.03197086), ('variance', 0.025503768)]								
45	[('dob', 0.22922716), ('check', 0.07751342), ('suspect', 0.061426185), ('individual', 0.053063616), ('avondale', 0.03744808), ('background', 0.036360003), ('protection', 0.028581668), ('renovation', 0.026687257), ('background_check', 0.023604758), ('√¢,Ç¨¬¢', 0.014700745)]	Criminal Record Check	13	13	41	29	12	36	12
46	[('locate', 0.12590905), ('building', 0.11562841), ('property', 0.074399255), ('inspection', 0.07156111), ('property_locate', 0.060741402), ('plan', 0.059727844), ('parcel', 0.052651398), ('drawing', 0.0375317), ('permit', 0.029983826), ('spill', 0.028118191)]	Parcel Records, Permits, Plans	10	10	17	6	4	6	4
47	[('page', 0.089249946), ('program', 0.08123574), ('control', 0.062193304), ('survey', 0.050580304), ('do', 0.043303214), ('summary', 0.036729675), ('pdf', 0.035272434), ('truck', 0.030877778), ('garage', 0.02812399), ('station', 0.027182342)]	Uncategorized	56	48	30	53	59	53	59
48	[('incident', 0.52505744), ('incident_report', 0.15395674), ('assault', 0.055209246), ('around', 0.04214052), ('resident', 0.029872352), ('instrument', 0.01338698), ('hospital', 0.013050693), ('inventory',	Police Incident Report	6	6	23	9	3	9	3

	0.011345894), ('fraud', 0.00896936), ('dollar', 0.008546473)]								
49	[('section', 0.051301472), ('require', 0.033593092), ('exempt', 0.030290857), ('portion', 0.028061418), ('exemption', 0.024064451), ('right', 0.024005381), ('denial', 0.022793548), ('check', 0.02138788), ('interest', 0.019739026), ('deny', 0.019162482)]	Criminal Record Check	26	28	21	37	44	41	44
50	[('give', 0.10310085), ('floor', 0.048503865), ('cause', 0.045723896), ('personnel', 0.03707043), ('source', 0.03620974), ('pick', 0.03261719), ('floor_plan', 0.023982473), ('assign', 0.023205146), ('responsible', 0.017972155), ('identification', 0.015963526)]	Uncategorized	58	59	49	54	53	57	53
51	[('matter', 0.08015114), ('tenant', 0.042593118), ('pre', 0.03209588), ('represent', 0.030175263), ('non', 0.028152522), ('specification', 0.027090076), ('via', 0.023603652), ('appreciate', 0.023358887), ('quarter', 0.020137502), ('following', 0.018199366)]	Public Works & Utilities	50	47	34	52	49	56	49
52	[('first', 0.043509472), ('security', 0.03722707), ('proposal', 0.033257924), ('set', 0.031995006), ('purpose', 0.029597947), ('related', 0.028097117), ('federal', 0.027885847), ('easy', 0.022838237), ('card', 0.021906871), ('act', 0.020320265)]	Uncategorized	55	54	56	47	56	40	56
53	[('department', 0.46678814), ('police_department', 0.14740212), ('police',	Police Admin Reclass	25	26	31	30	19	30	19

	0.13515571), ('interview', 0.025639657), ('officer', 0.016585566), ('investigator', 0.014546079), ('prevention', 0.012628612), ('answer', 0.010397196), ('cpl', 0.010000569), ('discovery', 0.003616073)]								
54	[('message', 0.049739394), ('text', 0.04523615), ('boulevard', 0.040963557), ('another', 0.037343223), ('perform', 0.03510619), ('fill', 0.03369129), ('lease', 0.033601668), ('own', 0.033182345), ('foot', 0.028789343), ('space', 0.027618203)]	Uncategorized	59	60	51	55	58	54	58
55	[('environmental', 0.048749115), ('hazardous', 0.0456608), ('storage', 0.043166153), ('material', 0.042970687), ('hazardous_material', 0.034293775), ('tank', 0.034046005), ('property', 0.033831228), ('site', 0.03341891), ('assessment', 0.028234059), ('storage_tank', 0.025999954)]	Environ. Assess./ Hazardous Materials	5	5	19	1	1	1	1
56	[('police', 0.30953634), ('police_report', 0.20703475), ('copy_police', 0.105848916), ('request_police', 0.036158293), ('involve', 0.031712394), ('try', 0.023015141), ('officer', 0.019914081), ('approximately', 0.015022011), ('send', 0.012208854), ('plate', 0.010990596)]	Police Incident Report	4	4	15	11	6	10	6
57	[('water', 0.1396025), ('utility', 0.08118722), ('sewer', 0.07739645), ('anything', 0.06501578), ('plat', 0.054275632), ('fine', 0.05123444), ('municipal', 0.040840622),	Public Works & Utilities	27	24	11	23	37	15	37

	('payment', 0.03251683), ('electric', 0.029127842), ('joemillgmailcom', 0.026157256)]								
58	[('statement', 0.12197899), ('certified', 0.07444108), ('witness', 0.059776813), ('court', 0.042062994), ('witness_statement', 0.035788205), ('form', 0.030713223), ('officer', 0.027909148), ('cooperation', 0.023561912), ('criminal', 0.023481019), ('write', 0.02166455)]	Witness Statements	23	18	42	34	26	27	26
59	[('seek', 0.06251169), ('damage', 0.05873098 2), ('break', 0.05518673), ('result', 0.05407579), ('hold', 0.04934643), ('facility', 0.04679601), ('evidence', 0.045945678), ('doc', 0.032376543), ('night', 0.022582108), ('cover', 0.022135464)]	Police Incident Report	57	53	46	60	46	52	46

Final Top 10 Topics

The final topic rankings were calculated using the dampened popularity metric with a threshold of .2 for including a PRR's composition score in the total for a topic.

Rank	Topic Key Words
1	[('environmental', 0.048749115), ('hazardous', 0.0456608), ('storage', 0.043166153), ('material', 0.042970687), ('hazardous_material', 0.034293775), ('tank', 0.034046005), ('property', 0.033831228), ('site', 0.03341891), ('assessment', 0.028234059), ('storage_tank', 0.025999954)]

2	[('violation', 0.12519231), ('code', 0.08656456), ('property', 0.06848214), ('open', 0.06713956), ('code_violation', 0.057894886), ('zone', 0.04222084), ('building', 0.03874377), ('fire', 0.036294296), ('unit', 0.030973408), ('apn', 0.030083403)]
3	[('incident', 0.52505744), ('incident_report', 0.15395674), ('assault', 0.055209246), ('around', 0.04214052), ('resident', 0.029872352), ('instrument', 0.01338698), ('hospital', 0.013050693), ('inventory', 0.011345894), ('fraud', 0.00896936), ('dollar', 0.008546473)]
4	[('locate', 0.12590905), ('building', 0.11562841), ('property', 0.074399255), ('inspection', 0.07156111), ('property_locate', 0.060741402), ('plan', 0.059727844), ('parcel', 0.052651398), ('drawing', 0.0375317), ('permit', 0.029983826), ('spill', 0.028118191)]
5	[('permit', 0.2758547), ('issue', 0.11701758), ('building', 0.07663683), ('build', 0.06414055), ('building_permit', 0.059510738), ('ref', 0.056181964), ('construction', 0.03771535), ('without', 0.024624562), ('administration', 0.020187119), ('since', 0.02015884)]
6	[('police', 0.30953634), ('police_report', 0.20703475), ('copy_police', 0.105848916), ('request_police', 0.036158293), ('involve', 0.031712394), ('try', 0.023015141), ('officer', 0.019914081), ('approximately', 0.015022011), ('send', 0.012208854), ('plate', 0.010990596)]
7	[('accident', 0.39958552), ('accident_report', 0.18422489), ('auto_accident', 0.114594795), ('auto', 0.09796612), ('pedestrian', 0.019592816), ('architectural', 0.014912473), ('usaa', 0.010008935), ('everything', 0.009694831), ('fault', 0.006688628), ('instruction', 0.006613127)]
8	[('plan', 0.11660539), ('site', 0.099882215), ('certificate', 0.09434634), ('occupancy', 0.08399464), ('certificate_occupancy', 0.07622669), ('site_plan', 0.045080233), ('final', 0.040189173), ('permit', 0.034261044), ('interested', 0.03197086), ('variance', 0.025503768)]
9	[('case', 0.37224895), ('number', 0.20965174), ('case_number', 0.046436027), ('steal', 0.043556415), ('report_case', 0.026024856), ('update', 0.019941686), ('analysis', 0.019081214), ('disposition', 0.010940141), ('view', 0.010393641), ('stolen', 0.009620244)]
10	[('insurance', 0.10682117), ('reference', 0.09845432), ('type', 0.087784335), ('number', 0.07008757), ('location', 0.06628888), ('transaction', 0.06523592), ('insure', 0.06116695), ('occurrence', 0.058346972), ('occurrence_location', 0.05399189), ('type_auto', 0.037329476)]

The 60 topics generated by the LDA algorithm were grouped into 19 data type categories. We calculated the popularity of the data types using the same set of methods used for calculating topic popularity above. The results are as follows:

	Data Type	W	P .2	P .5	DP W .2	DP W .5	DP P .2	DP P .5
0	911/Law Enforcement Service Calls	8	7	8	14	9	13	9
1	Auto Collision Report	2	2	1	4	3	4	3
2	Building Code Violations	9	10	7	13	7	14	7
3	Checks and Deposits	19	19	17	18	15	19	15
4	City Emails, Social Media	7	8	9	10	10	9	10
5	City Government Meeting Notes	17	17	18	15	17	15	17
6	Complaints to City	18	18	19	19	18	18	18
7	Crime Photo and Video	4	4	3	9	8	10	8
8	Criminal Record Check	5	5	4	3	5	3	5
9	Employee Benefits, Payroll	14	15	14	11	11	12	11
10	Environmental Assessment, Hazardous Materials	6	6	6	5	4	7	4
11	Human Services Cases	16	16	16	16	14	17	14
12	Parcel Records, Permits, Plans	3	3	5	1	2	2	2
13	Police Incident Report	1	1	2	2	1	1	1

14	Property Liens	10	9	12	7	12	6	12
15	Public Works, Utilities	12	13	11	12	16	11	16
16	Purchasing Records, Contracts	11	12	10	6	6	8	6
17	Uncategorized	13	11	15	8	19	5	19
18	Witness Statements	15	14	13	17	13	16	13

Top Categories by Method

	W	P .2	P .5	DP W .2	DP W .5	DP P .2	DP P .5
Crime	1	1	1	1	1	1	1
Environment	4	5	3	5	4	6	4
Government	3	3	4	4	5	4	5
Human Services	8	8	8	8	6	8	6
Property	2	2	2	2	2	2	2
Public Works	6	7	6	7	7	7	7
Spending	5	4	5	3	3	3	3
Uncategorized	7	6	7	6	8	5	8