Supporting Information for:

PROXYING FOR UNOBSERVABLE VARIABLES WITH INTERNET DOCUMENT-FREQUENCY

Albert Saiz[a], Uri Simonsohn[b]

[a] (corresponding author) The Wharton School, University of Pennsylvania, 1466 Steinberg-Dietrich Hall, 3620 Locust Walk, Philadelphia, PA 19104-6302, saiz@wharton.upenn.edu
[b] The Wharton School, University of Pennsylvania, 548 Hunstman Hall, 3730 Locust Walk, Philadelphia, PA 19104, uws@wharton.upenn.edu

Supporting Information for "Proxying for unobservable variables with internet document-frequency"

We supplement our paper with the nine additional sections listed below:

A. Further details and analyses on the demonstration with the five socioeconomic variables.
B. An examination of whether the relationship between document-frequency and occurrence-frequency is monotonic.
C. An exploration of the number of documents needed for document-frequency to be a stronger proxy.
D. Summary statistics with the number of documents and occurrences for variables used in the paper.
E. OLS regressions as in Table 2 that exclude actual or proxy variables (omitted variables)
F. Guns and murder: proxy that is suspect of bias toward dependent variable
G. Correlations of document frequencies with alternative World corruption indexes
H. Alternative data indexes and search engines
I. Detailed discussion of the results for state-level corruption regressions, and further replications

A - Further details and analyses on the demonstration with the five socioeconomic variables.

We conducted our document-frequency estimations of the five socioeconomic variables using the internet search engine Exalead®[1], and the newspaper database Newsbank®.[2] We focus on contemporaneous web searches and on newspapers published in the five years between 9/1/2001 and 31/8/2006, because the Newsbank's coverage is very limited before the initial date.

We queried both Newsbank and Exalead utilizing specially designed Perl scripts, adding considerable time delays between queries to avoid imposing unreasonable burdens on the servers. We obtained occurrence-frequency data for the five variables we employed in the initial demonstration from aggregate census counts, the FBI's Uniform Crime Reports, and HUD State of the Cities Database. For state-level poverty we use the percentage of population with income below one half of the state median. For poverty at the city-level we do not have microdata for all the cities so we use instead the official poverty rate as reported by the Census. [3]

To estimate document-frequency we conducted proximity searches with the keywords "African American OR African Americans," "Hispanic OR Hispanics," "Immigrant OR Immigrants," "poverty," and "murder." We used all cities with a population of 100,000 or more in the 2000 Census and all 50 states as locations. We excluded cities that have the same name as another city of more than 100,000 inhabitants,

[1] At the time of our data-collection, only Exalead provided the option of conducting proximity searches, i.e. of restricting the search to documents containing the keywords of interest within 16 words of each other.
[2] We considered two other newspaper databases: *Lexis-Nexis* and *Factiva*. We chose *Newsbank* because it has the largest set of local newspapers and because, unlike *Lexis-Nexis*, it does not place a limit on the number of documents found on a single query.
[3] Note that these poverty rates are computed utilizing a nation-wide nominal income threshold, overestimating poverty in cheaper cities and underestimating it for expensive ones. This measurement error induces a conservative bias in our estimated correlations.

such as Arlington and Springfield. As mentioned, we calculated document-frequency as the ratio of documents obtained via a proximity search and the total number of documents with the name of the location.

The distributions of both occurrence- and document-frequencies tend to have a right skew, and we conduct all analyses on the log of these variables.[4] Figure S1 shows the occurrence- and document-frequency distributions of the share of African Americans across cities and their log-standardized version. The graphs also display the normal distribution with mean and variance corresponding to the data.

Figure S1

To assess the validity of the log-transformation we estimated the parameter λ in Box-Cox regressions of the general form:

$$(2) \qquad \frac{(\hat{Y}_{p,l,k})^{\lambda} - 1}{\lambda} = \alpha_{p,k} + \beta_{p,k} \frac{(Y_{p,l})^{\lambda} - 1}{\lambda} + \varepsilon_{p,l,k}$$

Where $\hat{Y}_{k,p,l}$ is the relative document-frequency of occurrence *p* with regards to location *l* as proxied by keyword *k* and $Y_{p,l}$ is the corresponding occurrence-frequency. The estimate of λ indicates the optimal Box-Cox transformation to both variables. $\lambda = 0$ indicates a log transformation.

We fitted the Box-Cox model for our 5 keywords, both for the internet and newspapers, both at the state- and city-level. The average of the resulting 20 estimates of λ was $M = -0.107$, $SE = 0.037$. While this is statistically significantly different from zero it is quite close to it. Indeed, 12 of the 20 point estimates are not statistically

[4] We add 1 to the numerator so that the few cases with 0 documents can be included in the analyses.

different from 0. Furthermore, the log transformations are very highly correlated with those resulting from using the λs from the Box-Cox transformation.

To provide an intuitive sense of the relationship between document and occurrence-frequency for these variables, Figure S2 depicts quintile averages for each at the city-level. The vertical axes contain the occurrence-frequency of the variable being proxied and the plotted lines contain the average of occurrence-frequency by quintile of document-frequency in the left column and by quintile of occurrence-frequency in the right column. For example, the two plots in the first row show that in cities with the highest quintile of document-frequency about African Americans, 31 percent of the population is African American, compared to 48 percent for cities in the highest quintile of occurrence-frequency of African Americans. Overall, the document-frequency figures show increasing profiles, albeit they are flatter than those of the occurrence-frequency figures.

*** Figure S2 ***

In the text we refer to the cross-correlations between the document-frequency of African Americans and the occurrence-frequency of the four other variables we consider. Table S1 presents the full set of cross-correlations alluded to there.

*** Table S1***

B. An examination of whether relationship between document-frequency and occurrence-frequency is monotonic.

The paper demonstrates that document frequencies can be correlated with occurrence frequencies *on average*. Here we examine whether the relationship between the two is monotonic. This would not be the case if, for instance, at very high or low levels of occurrence *changes* in empirical frequencies were negatively related to *changes* in publication frequencies at the margin.

To examine this issue we pooled observations from all variables at the city level and estimated a spline regression. That is, we identified cutoff points for quintiles of occurrence-frequency (of all standardized variables pooled) and then each observation's occurrence-frequency was compared to these 'knots'. In particular, let the new five spline variables be represented by S_i with i=1 to 5, and the inter-quintile cutoff point separating quintile i from quintile $i+1$ be represented by k_i. The value of S_i is determined by the following conditions:

If $y < k_i$ then $S_i = 0$

If $k_i \leq y \leq k_{i+1}$ then $S_i = y - k_i$

If $y > k_{i+1}$ then $S_i = k_{i+1} - k_i$

Note that $y = S_1+S_2+S_3+S_4+S_5$.

A regression with document-frequency as the dependent variable and S_1 through S_5 as predictors estimates the marginal impact of occurrence-frequency on document-frequency separately for variation in occurrence-frequency happening in each of its five quintiles.

Considering that we pooled observations across all phenomena we include slope dummy interactions for them (e.g., a "murder" dummy interacted by its occurrence-

frequency).[5] Table S2 shows the results for both internet and newspaper based document-frequencies. All point estimates are positive, and with a few exceptions significant, indicating that within each quintile of occurrence-frequency, a marginal increase in occurrence-frequency is associated with an increase in document-frequency in the margin.

Table S2

C. An exploration of the number of documents needed for document-frequency to be a stronger proxy.

As mentioned in our discussion of data-check #4, if absolute keyword frequencies are small, variation across locations will be overridden by sampling error and hence the resulting proxy will be unreliable. In this section we gauge the relationship between the average number of documents found for each location and keyword and the fitness of document-frequency as a proxy for occurrence-frequency.

The ideal way to do so would be to query the full databases of documents we use (Newsbank and Exalead) and to create random subsamples of varying sizes from the resulting sets of documents. This approach, unfortunately, is prohibitively costly, as it requires *downloading* and analyzing the millions of documents that are obtained with the queries (just with "New York" there are over 60 million web documents).

As an alternative, we conducted queries on the full universe of documents but restricting searches so that only documents published during shorter periods of time would be considered. We focus on newspaper data at the city-level using Newsbank. In particular, rather than conducting a single query per location-variable pair for all

[5] Main effect dummies are not included because the variables were standardized separately. African American share is the excluded interaction.

documents published between 2001 and 2006, we conducted 60 such queries per location-variable pair (e.g. "crime" NEAR "Los Angeles"), restricting the results to have been published during each of the 60 months.[6] The resulting document-frequencies are hence based, on average, on samples 1/60 the size of the original sample. By adding up partial sums for different (randomly selected) months we can then create larger samples.

We assess the impact of increasing average absolute numbers of location-keyword documents by monitoring the evolution of the correlation between actual occurrence-frequency and document-frequencies computed over samples of increasingly larger sizes.[7]

Figure S3 shows the results from this exercise conducted on two different random subsets (without replacement) of 30 months each. The x-axis contains the average number of documents across cities (average number of documents containing the name of the city in proximity to the keyword) in the cumulative sample, as more and more months are added in random order. The Y-axis shows the correlation of the proxy arising from that sample with the corresponding occurrence-frequency. Random sample 1 is plotted with dark points, whereas sample 2 is pictured using transparent diamond signs.

Figure S3

The results depicted in Figure S3 are highly comparable for the two random subsamples we employed (which have no overlap). They suggest that the two samples tend to converge to the baseline correlation independently, providing further evidence of

[6] We conduct these searches only on Newsbank because internet searches with date restrictions, although possible, are not very reliable. Most notably, they obviously do not retrieve documents that were uploaded in the past but which are no longer available.

[7] The fact that we are sampling at the month-level rather than independently at the document-level reduces the efficiency of our samples. Truly random subsamples should converge faster, leading to an even smaller number of documents required to achieve a robust proxy.

the usefulness of document frequencies as proxies. In quantitative terms, of course, queries that retrieve a large average number of hits generate stronger correlations with occurrence-frequencies. If a word appears in only a few instances in the newspaper database, chances are that differences in relative frequencies across cities are mostly driven by random noise. However, in these data, we see remarkably fast improvements with moderate growth in word frequencies. To see this we calculated the square root of the square error of the correlations in Figure S3. The square error is defined here as the square of the difference between the correlations obtained with each cumulative sample (Figure S3) and the final correlations we obtained in Table 1. This is just a descriptive measure of "how far" the correlations in each sequential iteration of the cumulative sample are from the final correlations between occurrence and document frequencies using the full sample. Note that we have 5 variables and 2 independent sequences for each, for a total of 10 sequences. We display the square error information for these ten sequential samples in Figure S4 (top). We see that the root of the square error actually declines quite rapidly.

Figure S4

Taking the average across 30 intervals in the number of observations – Figure S4 (bottom) – the mean square error in this sample is between 0.05 and 0.07 in the range of 100-200 average observations per city. Since the average correlation in the overall sample is 0.35 for newspapers, this means that after 100 average keyword hits we already obtain mean correlations that are between 80-85% of the final average one. Nevertheless, it is always true that larger textual samples improve the strength of the relationship between document and occurrence frequencies.

D. Summary statistics with the number of documents and occurrences for variables used in all of the analyses in the paper

In our discussion of the results of the data-checks with the data for the five socioeconomic variables we mention the number of documents found for different queries. In Table S3 we present the average number (and standard deviation) of documents found for different queries, and in Table S4 the absolute frequency of occurrence for different phenomena we proxied.

Table S3

Table S4

E. *OLS Regressions as in Table 2 that Exclude Actual or Proxy Variables*

In table S5 we reproduce the specifications in Table 2 of the paper. Columns 1 and 2 reproduce verbatim the same-numbered columns in Table 2 of the paper. Columns 3 through 6 sequentially exclude each individual variable for which hypothetically we would like to proxy for.

Table S5

Our contention is that including the document-frequency proxies could help (at least modestly) in attenuating omitted variable (OV) bias in the coefficient of the population variable, which by assumption is the coefficient of interest. This can be seen by comparing the coefficients in Table 2 (in which we control for the document-frequency proxies) to the coefficients in the Supplemental Table S5 (in which we omit each socio-demographic control in turn). To be fair, most of the coefficients are qualitatively similar, so that one would reach similar conclusions about the impact of

population on crime even with some missing controls. However, in the case of omitting the African-American share (Table S5 versus Table 2, column 3) using the document-frequency proxies helps to reduce the OV bias on the population variable considerably. A priori, in mean squared error terms, we would have been better off controlling for the noisy proxies than omitting them in this example.

F. *Guns and Murder: Proxy that is Suspect of Bias toward Dependent Variable*

We provide an example of how careful researchers could proceed when endogeneity concerns are present. In order to qualitatively illustrate good practice with regards to using the proxies as controls, we committed ourselves to the verbal example we initially provided in the paper. This we do in supplemental Table S6.

Table S6

The problem that the proxy may be spuriously correlated with the variable of interest should generally be considered by serious researchers any time that a proxy variable is used. In this context, document-frequencies provide us with extra opportunities to analyze data patterns with which to confirm or dissipate suspicions about the quality of the proxy. For example, in supplemental Table S6 we now show OLS regressions like those in Table 2. In Table S6 columns 1 and 2 we try to "account" for murder rates by city using the log of population, African American, Hispanic, and immigrant shares, and the poverty rate, together with the document-frequency of the word "guns" (internet in column 1, news in column 2).

Our concern here is that the media and internet talk more often about guns when these are associated with a crime. That would bias the true relationship between actual

gun ownership and murder rates upward, because in cities with more murders document-frequencies will tend to show more occurrences of the word gun, even if no causal relationship is present. Notice that this is also a very important concern in the use of other proxies that are unrelated to the internet. One of the advantages of document-frequencies is that one can obtain alternative proxies as placebos or refinements of the initial queries. In the text, we suggest two alternative approaches, which we now illustrate. First, note that the two initial regressions in Supplemental Table S6 suggest a positive link between gun document-frequency and murder rates at the city level, even after controlling for population and socioeconomic characteristics. This association is actually significant for the newspaper document frequency. However it is likely that the press more often reports about guns when they are associated with crimes. Our first approach to deal with this potential problem consists in performing another search for documents with the word gun, but that exclude words related to violent crime; concretely we used the syntax: *cityname* NEAR (guns OR gun) AND NOT (crime OR murder OR homicide OR homicides OR manslaughter OR kill OR kills OR killed OR slays OR slay OR slain OR assassinated OR injured OR injury OR injuries OR robbery OR robbed OR violence OR theft OR assault OR rape OR raped). As we suspected, a full 35% of "gun" hits in the Internet actually corresponded to pages making mention to violence. This problem is more severe in the case of newspapers, where a full 59% of pages mentioning guns also make reference to violence. Note that if, say, only 1 or 2% of the "gun" documents contained allusions to crimes, this would seem like a more minor concern.

One of the solutions we propose is to use excluding those gun documents that make reference to violent crime. This we do in columns 3 and 4. Interestingly, the

associations between such "cleansed" gun document frequencies and murder are now insignificant. In columns 5 through 10 we look at a dependent variable that we know should be related to gun ownership (see Azrael et al, 2004): the standardized share of suicides by metropolitan area in which a firearm is used between years 2000 and 2004 (firearm suicides divided by total suicides; we obtained the data from the 2000-2004 Mortality files of the US Center for Disease Control and Prevention – CDC). As could be expected, comparing columns 7 and 8 (cleansed) with 5 and 6 (contaminated), the cleansed version of the data seems to perform better: we obtain positive and stronger associations with gun-related suicides (albeit the newspaper data seems to be less reliably associated with this variable). For comparison, in columns 9 and 10, we present the results of using an alternative document-frequency proxy, as we suggested in the previous version of the paper: the relative frequency of the phrase "gun show," which does not have a potential direct relationship with crime. This variable is also associated, as it should, with the share of gun-executed suicides.

In sum, a simple analysis of the data patterns in the proxy would make us suspicious of using the document-frequency of the keyword *guns*. In this case, we'd have to recommend alternative proxies or implementing a MIMIC approach with multiple proxies where the document frequencies are both causes and consequences of guns, and where we model explicitly the multi-directional relationship between gun ownership, gun document-frequency, and guns.

G. *Correlations of document frequencies with alternative World corruption indexes*

Besides TI's CPI, other indexes of corruption at the country level exist. At a referee's request, we obtained all the national indexes from the World Bank Worldwide Governance Indicators. Here, we focus on indexes collected by the World Bank for which there are more than 100 observations (countries). We present the results in supplementary Table TS7.

Table S7

The correlation with the World Bank composite index (a weighted average of the other indexes allowing for data imputation as calculated by the source) is similar to that with Transparency International's measure (0.6). The 16 correlations between internet and newspaper document frequencies and the other (presumably noisier) data on corruption are always positive and on average around 0.5.

H. Alternative data indexes and search engines

In the development phase of this project we experimented with different search engines. We found Exalead® to be the best engine for our purposes because of its extended Boolean operators and the ability to be scraped. No other web engine allows using proximity operators. Most other engines do not allow automated scripts. However, we started the project using Ask.com® and Yahoo® search, which allowed us, with daily numerical limitations, a number of automated searches. Since we started with world corruption as an example, we retain data for our queries with Yahoo® and Ask.com®, as well as newspaper queries using Factiva®. The shortcomings of these alternative search engines is that they do not perform proximity searches, so that we only quantified the

number of pages that contained both the name of the country and the word corruption, not necessarily in textual proximity. Factiva® focused, at the time of our queries, on larger newspapers and syndicated TV news transcripts, with less coverage of local news (which are better for the city document-frequencies).

Table S8

In Table S8 of the supplemental materials section we show that the correlations between Transparency International's corruption index and the document-frequencies obtained with the alternative data repositories and search algorithms is very similar. However we want to note that, especially with Yahoo and Ask, the inability to perform proximity searches was really a potential issue at the local level: many pages contained both the name of a city (say Philadelphia Enquirer) and then the keyword in reference to another city (e.g. "Arizona approves new *immigrant* law"). In the paper, we focus on Exalead® in order to use Boolean proximity search and therefore minimize false positives, and also in order to take advantage of its more liberal data scrapping restrictions.

I.. More detailed discussion of the results for state-level corruption regressions

One of the two existing measures of corruption at the state-level is that of Glaeser and Saks (18). They construct their measure based on the number of government officials convicted for corrupt practices through the (federal) Department of Justice (DOJ). In particular they divide the average number of DOJ corruption convictions over the 1976-2002 period by the state's average population during that same period.[8]

[8] Corporate Crime Reporter, http://www.corporatecrimereporter.com/corruptreport.pdf, constructs essentially the same index.

As they acknowledge in their paper, there is a problem with deflating convictions by population, as doing so assumes that the number of government officials that could be corrupt has a linear relationship with population. States, of course, differ in the proportion of their citizens working for the government and hence at risk of engaging in the kind of behavior which could lead to a federal conviction.

With this consideration in mind, and particularly because size of government is one of the predictors used by GS, we modified their denominator, such that we divided DOJ convictions by the average number of *government employees* between 1976-2002.[9] In the paper we contrasted the results of document-frequency to this – in our view improved – benchmark. Here we expand the comparison to include Glaeser and Saks' original index (GS).

Altogether we have five measures of corruption at the state-level: (i) the original *GS* index, (ii) *GS* computed deflating by number of public employees rather than population for 1976-2002 (iii) Boylan and Long (2003)'s survey, (iv) internet based document-frequency index, and (v) newspaper based document-frequency index. In order to compare these variables measured in different units we log-standardize all indexes.

The correlations among all indexes are presented on Table S9. A graphic comparison between measures (i) from GS and (iv) [internet] can be appreciated in figure S5.

***Figure S5 ***

[9] Glaser and Saks (2006) point out that their preferred deflator would have been the number of public officials by state, for which data are not available. Number of public employees, however, is available. We suspect it is more highly correlated with number of officials than state population is.

The average correlation between the internet measure and the three occurrence-frequency based measures is .49 (column 1). Interestingly, internet document-frequency is more highly correlated with the DOJ and survey-based indexes than they are with each other (although the difference is not significant at conventional levels).

***Table S9 ***

When convictions are divided by public employees rather than population, the correlations with other corruption measures increase (e.g. from .43 to .59 with internet document-frequency and from .31 to .41 with Boylan & Lang (2003)'s survey. This is consistent with our claim that number of public employees is a more appropriate denominator for corruption convictions.

In Table S10 we expand on Table 3 from the paper. Most notably we present in column 1 the results employing the corruption measure based on a population deflator. In general, the results obtained with the number of public employees' deflator are of greater magnitude and more similar to those obtained with the other measures, further suggesting it the more apt deflator. Nevertheless, if the document-frequency based regressions are compared to column 1 the point estimates remain strikingly similar, with the main exception, not surprisingly, of government size, which is positive and significant only in column 1.

***Table S10 ***

Throughout their paper, *GS* show numerous others regressions studying the relationship between corruption and a variety of additional variables, controlling for all variables included in Table S10 except income inequality. To dispel any concerns about model selection, in Table S11 we report the results from replicating the subset of these

additional regressions which *GS* find to have a significant relationship with corruption (at the 5% level). Some of the estimates using the log version of GS's measure are no longer significant, but point estimates for the occurrence-frequency and document-frequency based measures of corruption are remarkably similar.

***Table S11 ***

Figure S1: Log-standardizing the Data Sources – African Americans in US Cities

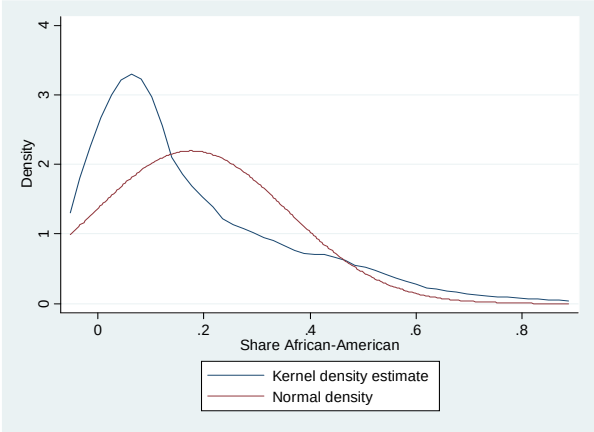
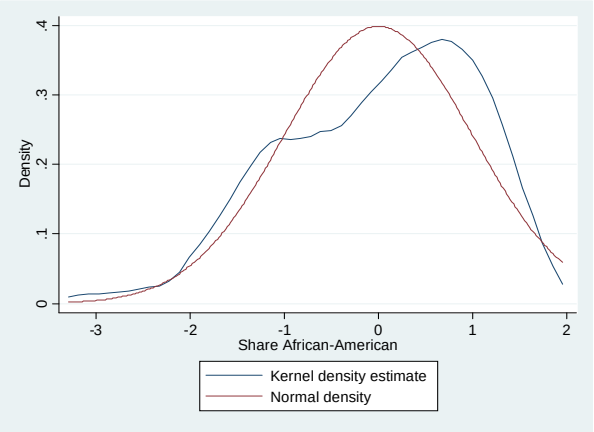
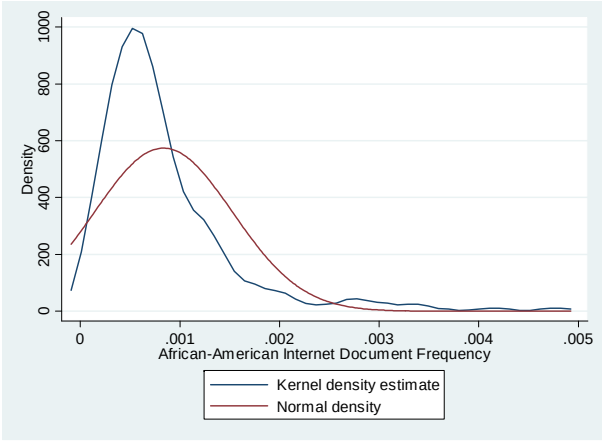
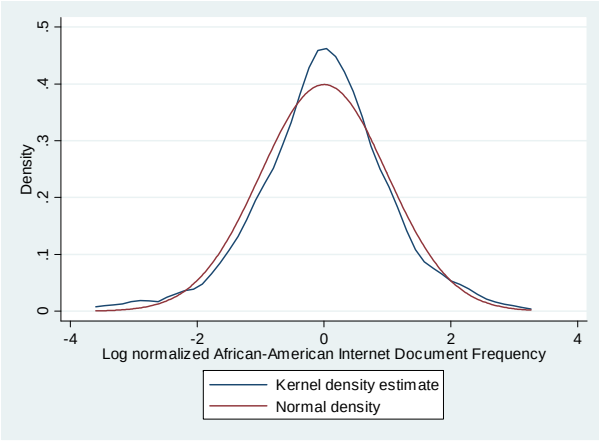
	Raw Data	Log Standardized Data
Empirical Occurrence Frequency		
Internet Document Frequency		

Figure S2: Average Data by Frequency Quintiles (cities)

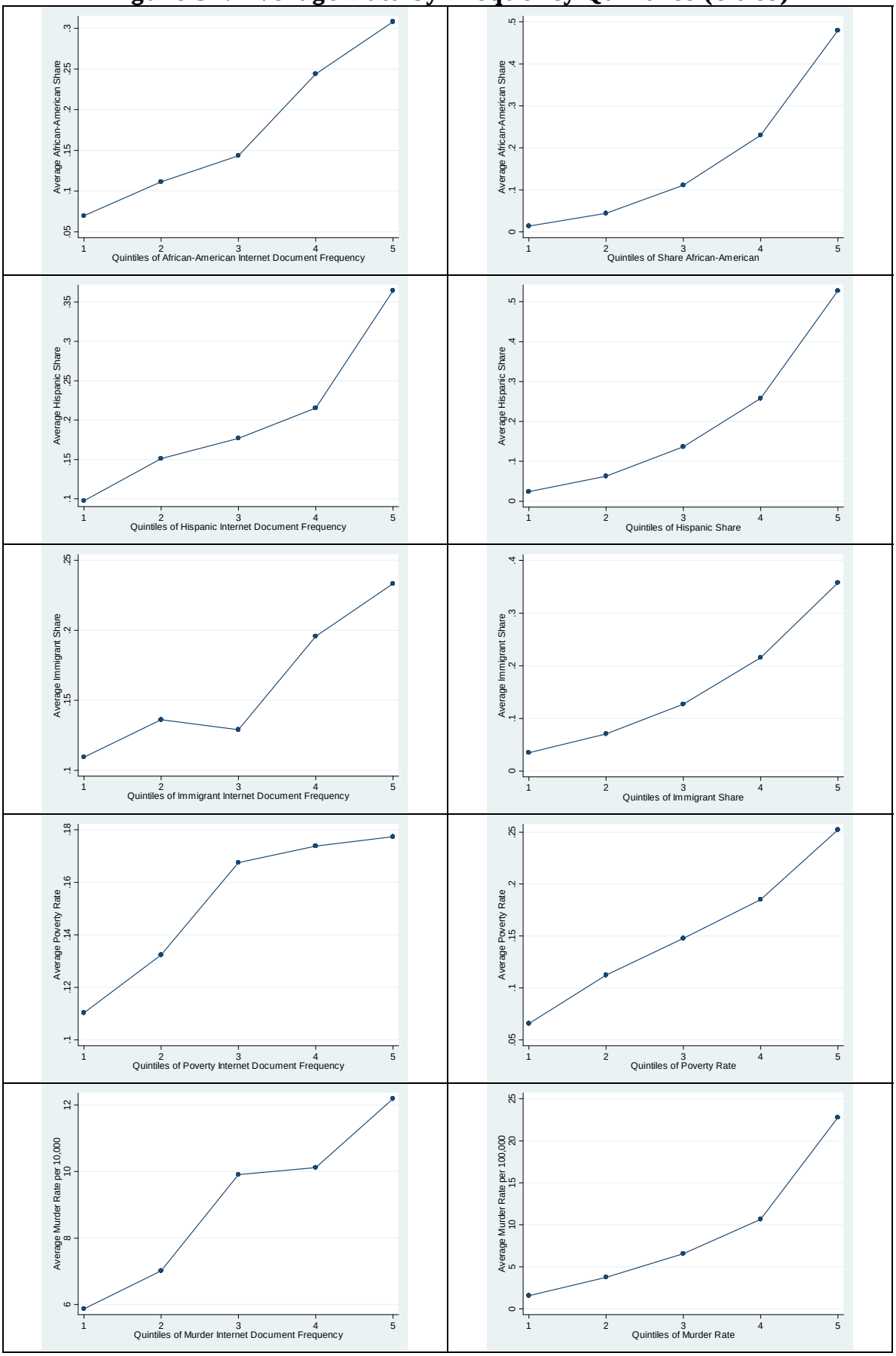


Figure S3
Correlations by Sample Size across Alternative Samples

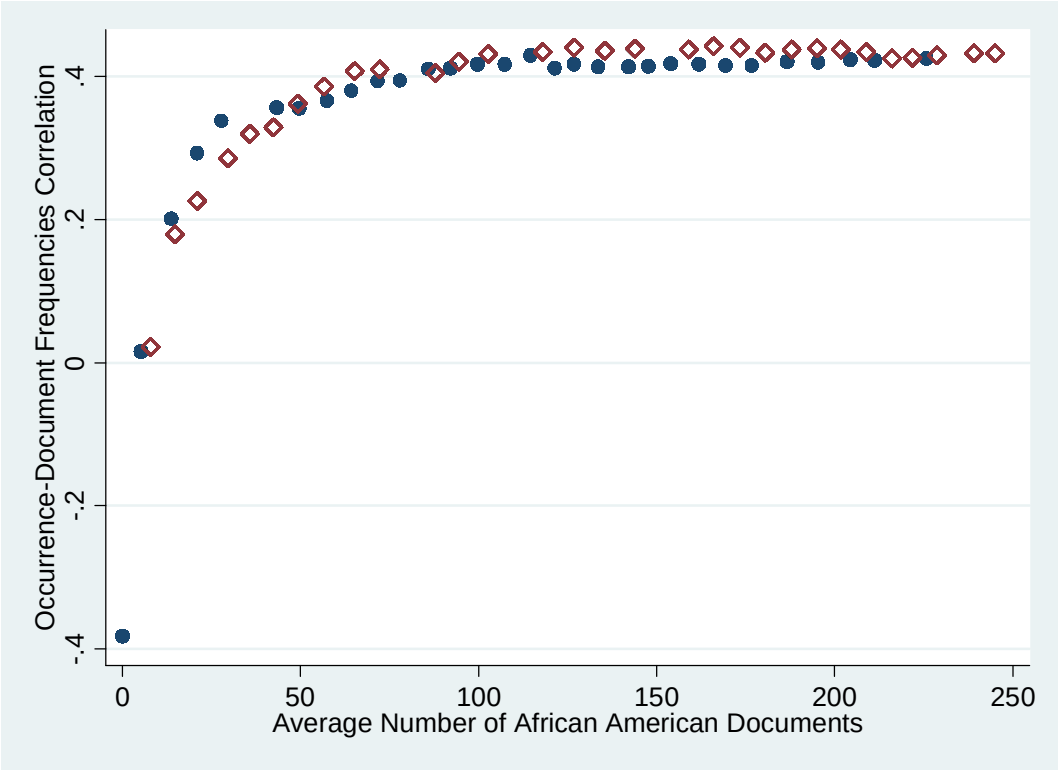
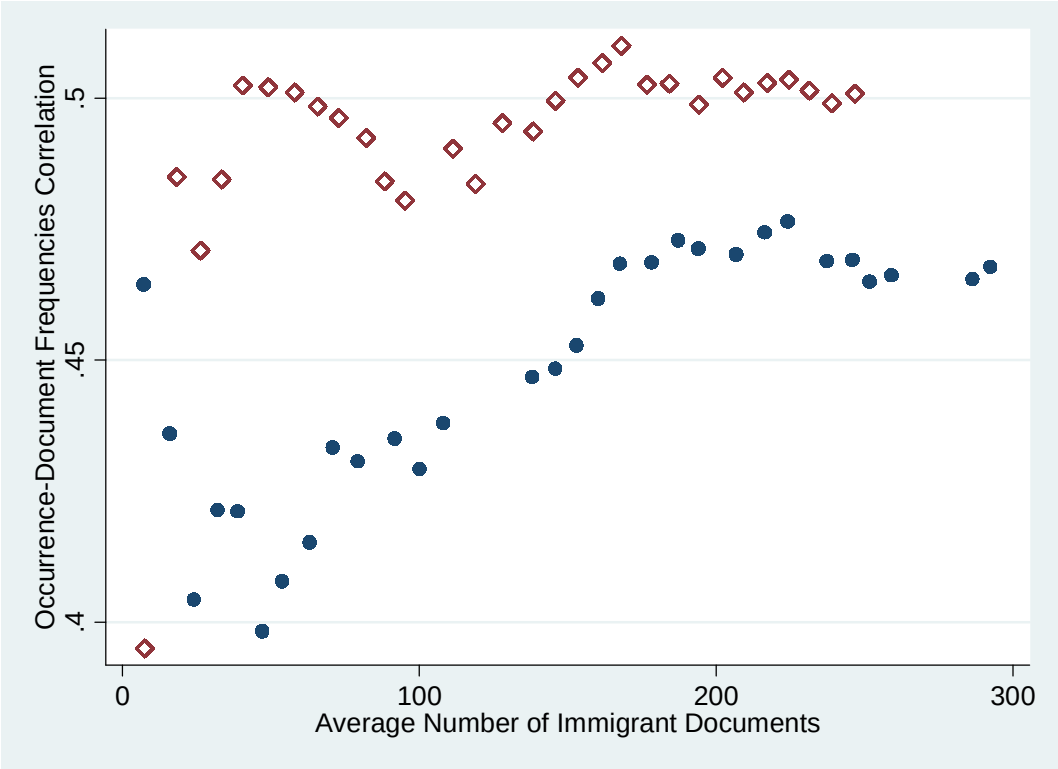


Figure S3 (Continued)



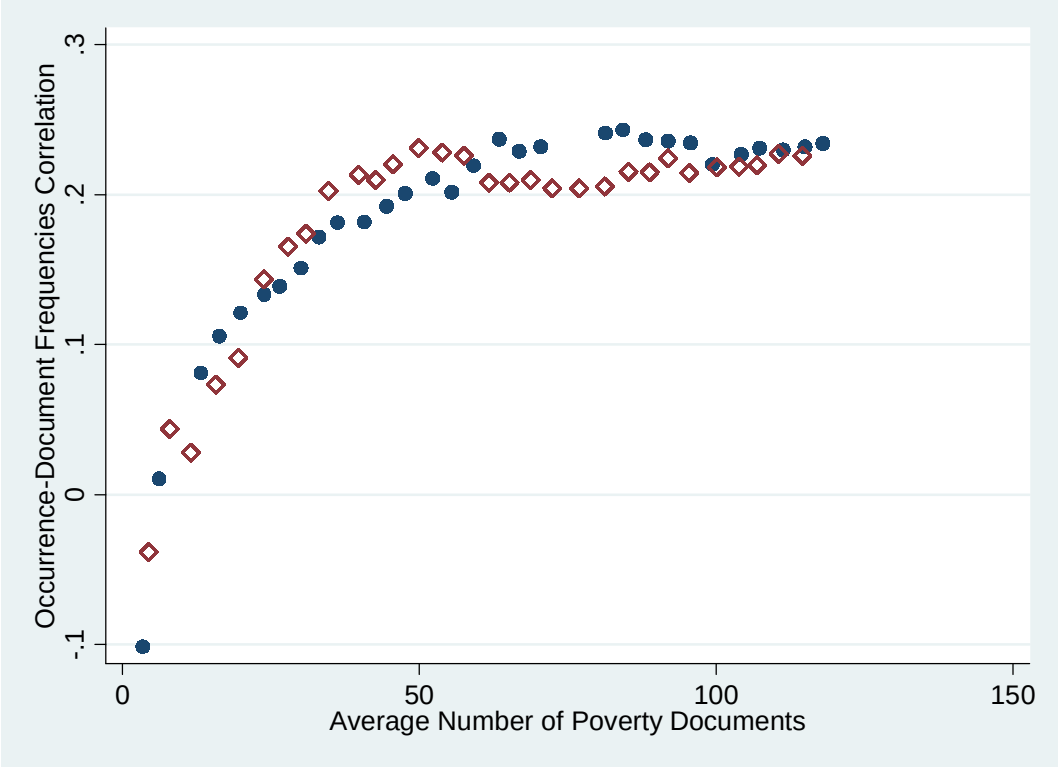


Figure S3 (Continued)

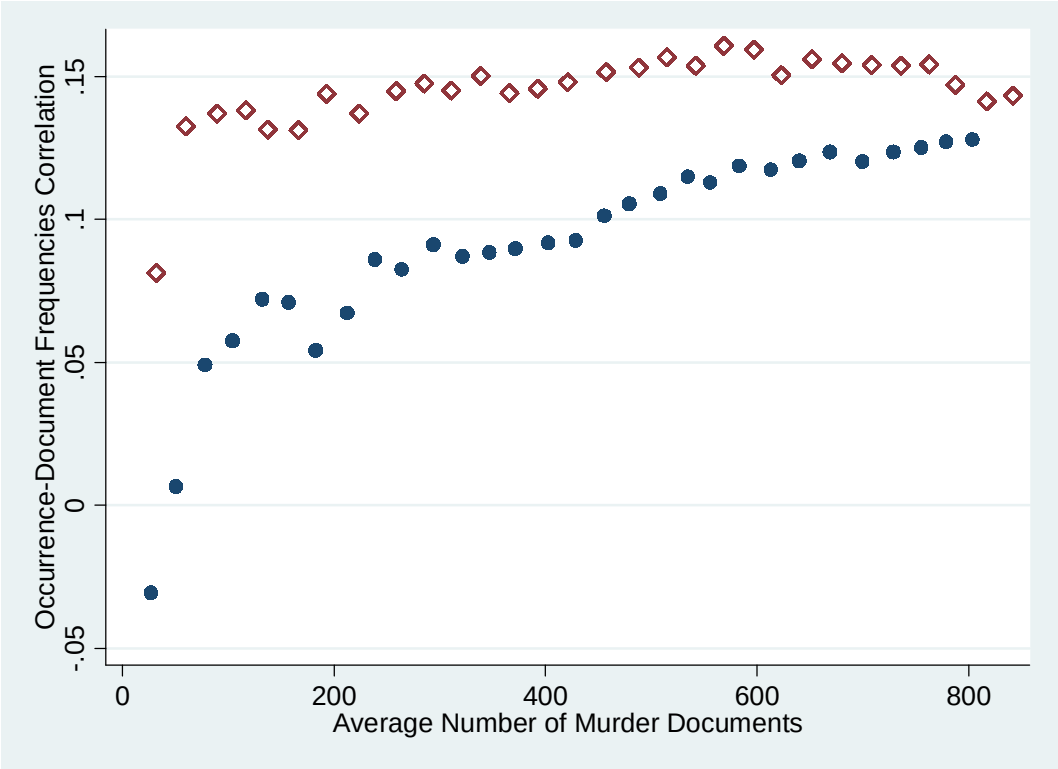
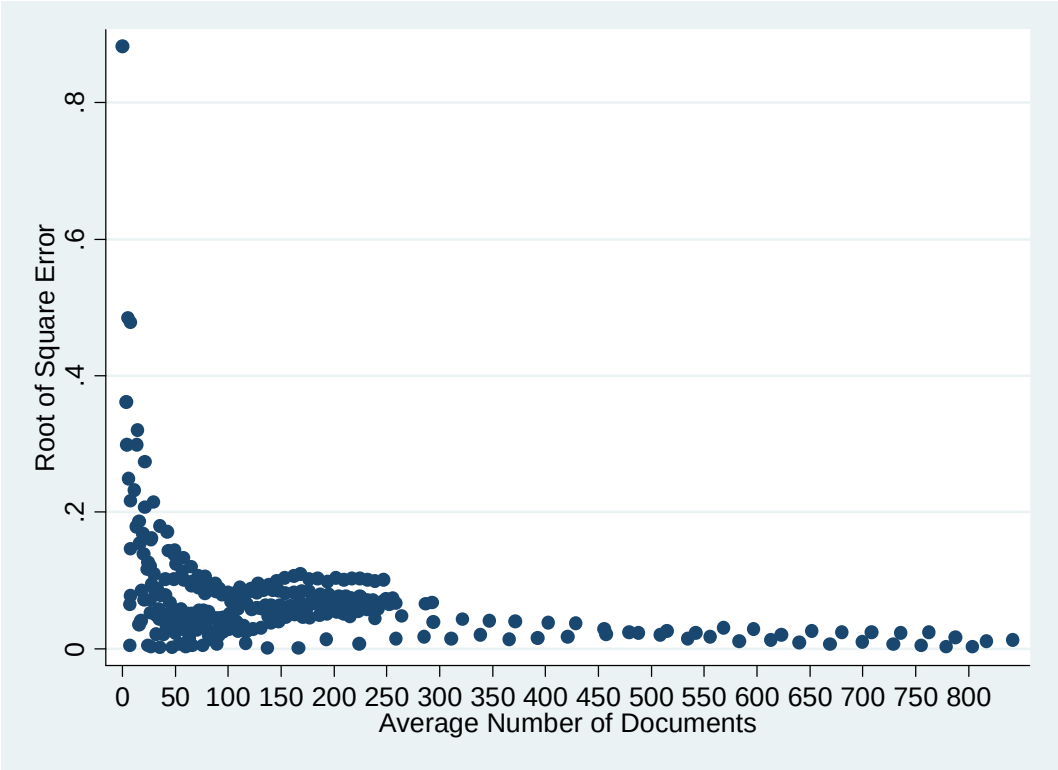


Figure S4

All Series



Means across Series

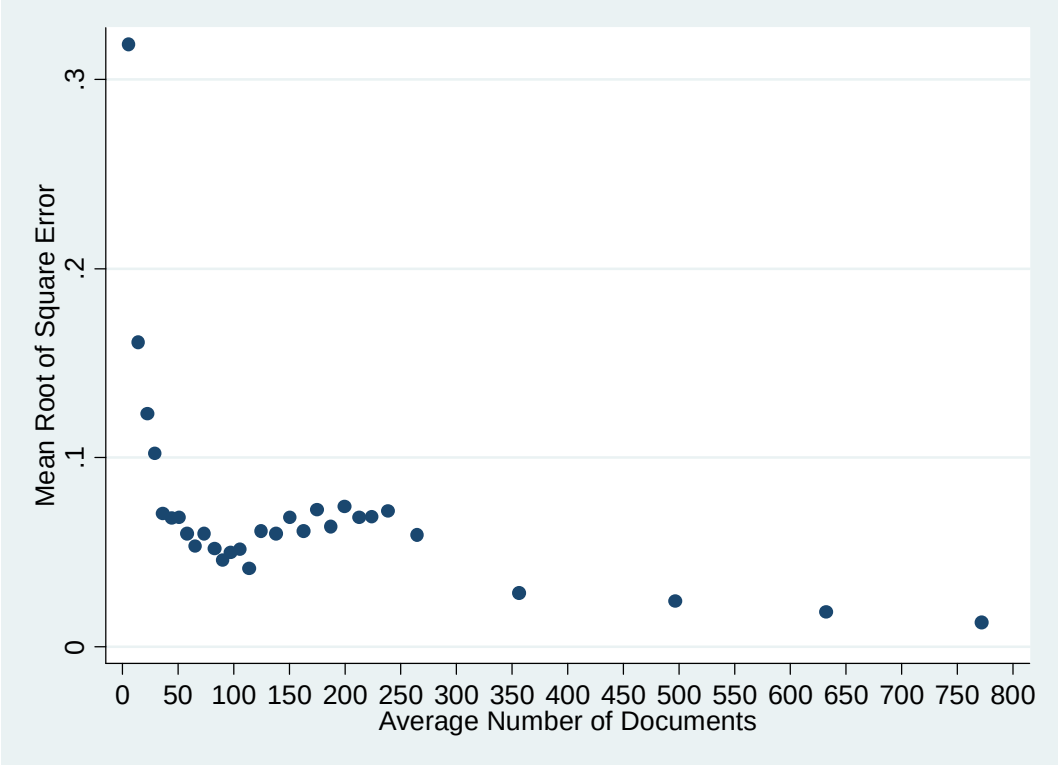


Figure S5: Corruption in the USA

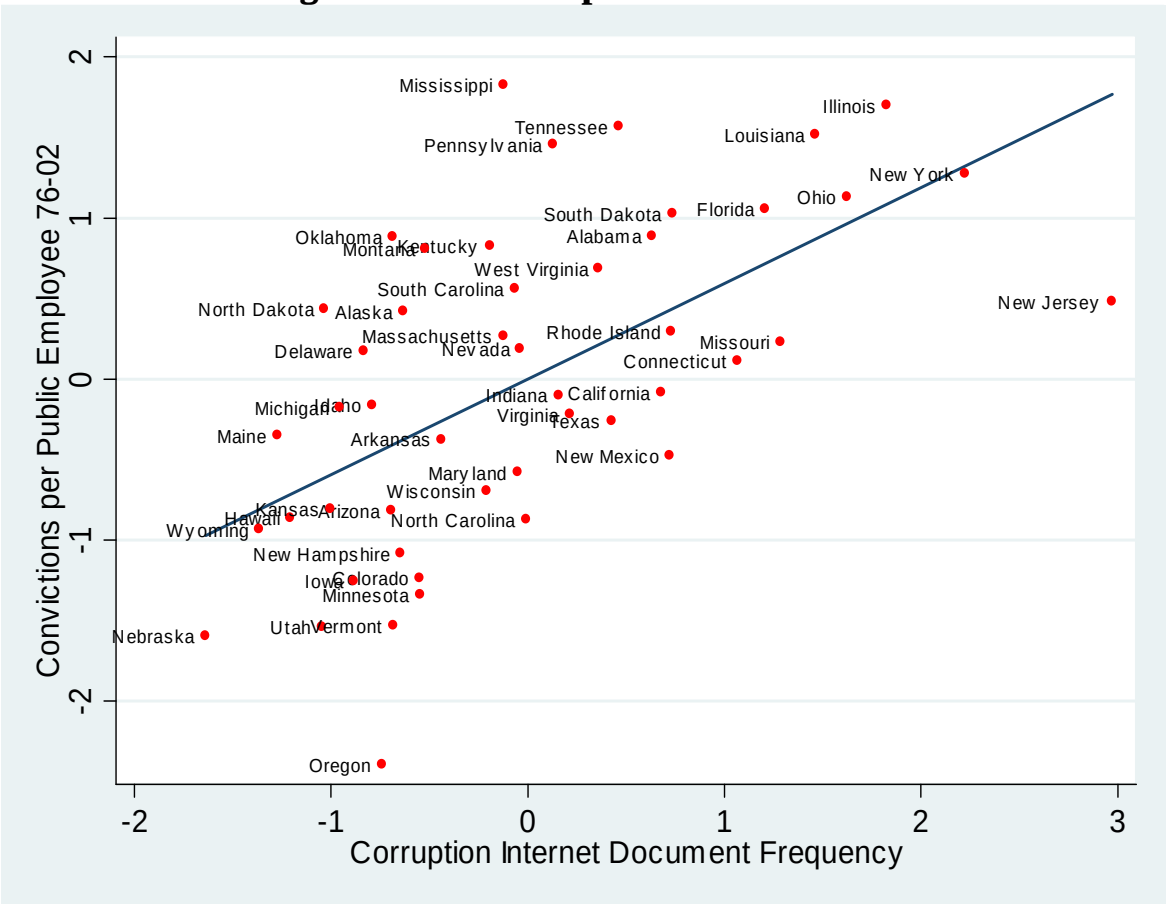


TABLE S1

Cross Correlations of Frequency of African-Americans with Other Frequencies

(1)	(2)	(3)	(4)
	Frequency of African Americans		
	Occurrence-Frequency	*Document-Frequency*	
		Internet	Newspapers
Occurrence-Frequency (States)			
African-Americans	1.00	0.70	0.82
Hispanics	0.16†	-0.08†	-0.05†
Immigrants	0.21†	-0.02†	0.06†
Poverty rate	0.21	0.47	0.36
Murder rate	0.80	0.62	0.65
Occurrence-Frequency (Cities population>250,000)			
African-Americans	1.00	0.67	0.61
Hispanics	-0.54	-0.40	-0.37
Immigrants	-0.48	-0.35	-0.31
Poverty rate	0.48	0.34	0.38
Murder rate	0.79	0.59	0.51

Notes: Each row in table reports the correlations between the occurrence-frequency of the variable listed in column (1), with frequency of African Americans. All correlations are statistically significant at the 5% level unless otherwise indicated.

† Not significant at 5%

TABLE S2

Spline Regressions Assessing if Relationship Between Occurrence-Frequency and Document-Frequency is Monotonic

	Dependent Variable	
	Internet document-frequency	Newspaper document-frequency
Predictors		
Spline 1	0.361***	0.428***
	(0.121)	*(0.120)*
Spline 2	0.287	0.573***
	(0.181)	*(0.180)*
Spline 3	0.668***	0.584***
	(0.220)	*(0.219)*
Spline 4	0.116	0.135
	(0.246)	*(0.245)*
Spline 5	1.027***	0.958***
	(0.203)	*(0.202)*
Socioeconomic Variable Dummies (K = 5)	yes	yes
Observations (City [226] * Socieconomic Variable [5])	1,130	1,130
R^2	0.13	0.12

Notes: Entries in table are point estimates from OLS regressions that pool all city-level observations for frequencies of African Americans, Hispanics, Immigrants, and Poverty and Murder rate. Robust standard errors are in parentheses below parameter estimates. The dependent variable is the log-standardized document-frequency. The five predictors are splines for the corresponding occurrence-frequency. These splines measure the distance between a given observation's occurrence-frequency and each of the 5 cutoff points between quintiles of occurrence-frequency,bounded by 0 from below and by the next quintile from above (see text for details). The reported point estimates assess the impact of changes in occurrence-frequency, within each of its five quintiles, on document-frequency. All five point estimates for the splines being positive indicates that higher occurrence-frequency is associated with higher document-frequency within each of the 5 quintiles of occurrence-frequency.

TABLE S3

Number of Documents: Averages and Standard Deviations

Panel A: The Internet

Documents with City Name and Keyword:	States			Large Cities			Small Cities		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
African-American	50	35,957	48,777	62	20,721	30,555	165	3,827	6,108
Hispanic	50	16,864	20,351	62	9,010	12,214	165	1,383	1,904
Immigrant	50	10,715	15,707	62	6,913	14,020	165	1,123	2,099
Poverty	50	5,265	5,355	62	3,027	6,710	165	877	1,983
Murder	50	13,043	13,764	62	10,495	21,454	165	2,558	4,695
Corruption	50	2,801	4,471	62	1,763	4,079	165	410	1,109
Total	50	32,100,000	24,600,000	62	18,000,000	17,500,000	165	7,315,665	18,700,000

Panel B: Local Newspapers (DataBank)

Documents with City Name and Keyword:	States			Large Cities			Small Cities		
	N	Mean	Std. Dev.	N	Mean	Std. Dev.	N	Mean	Std. Dev.
African-American	50	1,079	1,046	62	1,013	1,142	165	278	472
Hispanic	50	1,472	2,046	62	1,079	1,198	165	282	420
Immigrant	50	1,710	2,580	62	1,271	1,743	165	264	427
Poverty	50	691	551	62	476	540	165	145	287
Murder	50	3,085	2,927	62	3,054	3,241	165	1,119	1,485
Corruption	50	403	568	62	334	529	165	94	236
Total	50	817,391	705,792	62	705,126	599,778	165	226,077	251,421

TABLE S4

Occurrence Frequencies: Averages and Standard Deviations

	States				Large Cities				Small Cities			
Population Percent	N	Mean	Std. Dev.	σ/μ Ratio	N	Mean	Std. Dev.	σ/μ Ratio	N	Mean	Std. Dev.	σ/μ Ratio
African-American	50	10.33	9.70	0.94	62	22.36	18.87	0.84	165	15.61	17.33	1.11
Hispanic	50	8.81	9.44	1.07	62	20.47	19.16	0.94	165	19.89	20.36	1.02
Immigrant	50	7.71	5.85	0.76	62	16.10	12.50	0.78	165	16.01	12.51	0.78
Poverty	50	20.76	1.88	0.09	62	17.41	5.64	0.32	165	14.39	6.76	0.47
Murder Rate*	50	4.66	2.46	0.53	62	13.07	9.82	0.75	165	7.45	8.04	1.08
Corruption Rate**	50	3.13	1.48	0.47	62	NA	NA	NA	165	NA	NA	NA

* Murders per 10,000

** Convictions per 100,000 employees

TABLE S5

Demographic Variables in an OLS Regression Setup

Panel A

Dependent Variable: Standardized Actual Murder Rate (US CITIES)

	All controls Missing (1)	All Controls (2)	Black Missing (3)	Hispanic Missing (4)	Immigrant Missing (5)	Poverty Missing (6)
Log of Population	0.541***	0.223***	0.350***	0.215***	0.190***	0.293***
	(0.085)	(0.051)	(0.062)	(0.053)	(0.052)	(0.059)
Standardized Share African American		0.502***		0.461***	0.519***	0.692***
		(0.046)		(0.046)	(0.047)	(0.048)
Standardized Share Hispanic		0.234***	0.083		0.059	0.422***
		(0.059)	(0.072)		(0.040)	(0.066)
Standardized Share foreign born		-0.230***	-0.290***	-0.055		-0.353***
		(0.059)	(0.073)	(0.040)		(0.067)
Standardized Poverty rate		0.389***	0.600***	0.446***	0.427***	
		(0.042)	(0.047)	(0.041)	(0.042)	
Constant	-6.628***	-2.730***	-4.288***	-2.642***	-2.334***	-3.592***
	(1.040)	(0.624)	(0.759)	(0.645)	(0.636)	(0.727)
Observations	222	222	222	222	222	222
R-squared	0.16	0.74	0.6	0.72	0.72	0.64

Panel B

Dependent Variable: Standardized Actual Murder Rate (US STATES)

	All controls Missing (1)	All Controls (2)	Black Missing (4)	Hispanic Missing (5)	Immigrant Missing (6)	Poverty Missing (7)
Log of Population	0.533***	-0.016	0.380***	-0.036	-0.094	0.045
	(0.119)	(0.105)	(0.127)	(0.120)	(0.107)	(0.110)
Standardized Share African American		0.699***		0.690***	0.700***	0.771***
		(0.100)		(0.115)	(0.107)	(0.103)
Standardized Share Hispanic		0.524***	0.503**		0.232***	0.556***
		(0.137)	(0.196)		(0.084)	(0.145)
Standardized Share foreign born		-0.378**	-0.384*	0.075		-0.451***
		(0.145)	(0.207)	(0.095)		(0.152)
Standardized Poverty rate		0.228**	0.386***	0.256**	0.269***	
		(0.085)	(0.117)	(0.096)	(0.088)	
Constant	-7.959***	0.234	-5.677***	0.536	1.403	-0.665
	(1.783)	(1.577)	(1.904)	(1.798)	(1.606)	(1.644)
Observations	50	50	50	50	50	50
R-squared	0.29	0.76	0.5	0.69	0.73	0.73

Standard errors in parentheses

* significant at 10%; ** significant at 5%; *** significant at 1%

Table S6

Proxy that is Suspect of Bias Toward Dependent Variable

	OLS Regressions: Dependent Variable									
	Standardized City Murder Rate				Standardized City Gun Suicide Relative Rate					
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**
Document Frequency of Guns	0.055	0.078**			0.100*	-0.039				
	(0.041)	(0.039)			(0.056)	(0.055)				
Doc. Frequency of Guns-Excluding References to Crime			0.017	-0.006			0.143**	0.044		
			(0.040)	(0.038)			(0.055)	(0.054)		
Document Frequency of "Guns Shows"									0.161***	0.118**
									(0.056)	(0.054)
Includes Full List of Controls in Table S5(2)	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Observations	220	220	220	220	227	227	227	227	227	227
R-squared	0.69	0.7	0.69	0.69	0.38	0.37	0.39	0.37	0.39	0.38

All OLS regressions control for all the independentvVariables in Table S5 (occurrence frequencies). Standard errors in parentheses
* significant at 10%; ** significant at 5%; *** significant at 1%

TABLE S7

Correlations Between Occurrence and Document Frequencies: World Corruption

		Internet	Newspapers	Countries
WGI	Worldwide Governance Indicators (World Bank)	0.60	0.59	156
DRI	Global Insight Global Risk Service	0.50	0.50	134
BRI	Business Environment Risk Intelligence	0.56	0.57	117
BTI	Bertelsmann Transformation Index	0.35	0.14[†]	115
EIU	Economist Intelligence Unit	0.49	0.47	144
GWP	Gallup World Poll	0.49	0.59	125
GAD	Cerberus Corporate Intelligence Gray Area Dynamics	0.36	0.33	146
PRS	Political Risk Services International Country Risk Guide	0.50	0.44	131
WMO	Global Insight Business Condition and Risk Indicators	0.60	0.53	156
	Average Correlations (except WGI)	0.49	0.46	

[†] Not significant at 5%

Table S8

Alternative Search Engines and News Database

Correlations Between World-Corruption Document and Occurrence Frequencies (TI)

	Exalead	Ask.com	Yahoo	Factiva	Newsbank	TI
Exalead	1					
Ask.com	0.60	1				
Yahoo	0.69	0.76	1			
Factiva (Major News)	0.59	0.47	0.36	1		
Newsbank	0.60	0.61	0.49	0.79	1	
Transparency International (TI)	**0.60**	**0.68**	**0.62**	**0.66**	**0.63**	1

Notes: The table uses the sample of observations with complete observations for all variables (N=152)
All correlations are significant at the p>0.01 level.

TABLE S9

Correlations of State-Level Corruption Measures

	Document Frequency		Corruption Convictions[a]		Survey[d]
	Internet	*Newspapers*	*per inhabitant*[b]	*per public employee*[c]	
Document Frequency					
Internet	1				
Newspapers	0.75	1			
Corruption Convictions[a]					
per inhabitant[b]	0.43	0.45	1		
per public employee[c]	0.59	0.60	0.90	1	
Survey[d]	0.44	0.51	0.31	0.41	1

[a] Convictions correspond to Federal Department of Justice convictions on corruption charges of state officials (as used in Glaeser & Saks, 2006)

[b] Division of total number of convictions by population, original Glaeser & Saks (2006) indicator

[c] Division of total number of convictions by number of public employees (authors' calculations).

[d] Survey of State House Reporters, Boylan & Lang (2003)

TABLE S10

Replication of Regressions Establishing Correlates of State Level Corruption in Glaeser and Saks (2006): Table 4 (1)

	(1)	(2)	(3)	(4)	(5)
Dependent Variable: Corruption as measured by	Convictions[a] per Inhabitant (76-02)	Convictions per Public[c] Employee (76-02)	Document Frequency *Internet*	Survey [d]	Document Frequency *Newspapers*
Income Inequality	0.786***	0.811***	0.927***	0.344	0.795***
	(0.168)	*(0.172)*	*(0.220)*	*(0.361)*	*(0.226)*
Ln(Income)	0.652***	0.759***	0.788***	0.599	1.050***
	(0.174)	*(0.192)*	*(0.231)*	*(0.403)*	*(0.235)*
Share of population in state with 4+ Years of College	-0.655***	-0.835***	-0.468***	-0.642**	-0.521***
	(0.152)	*(0.156)*	*(0.156)*	*(0.243)*	*(0.168)*
Share of all employees employed by the state government	0.386**	0.015	-0.401***	-0.052	-0.359**
	(0.173)	*(0.172)*	*(0.127)*	*(0.233)*	*(0.147)*
Ln(Population)	-0.009	-0.02	0.088	-0.199	-0.137
	(0.166)	*(0.178)*	*(0.121)*	*(0.175)*	*(0.117)*
Share of population living in urban environment	0.153	0.255	0.334***	0.660***	0.263*
	(0.188)	*(0.184)*	*(0.118)*	*(0.145)*	*(0.154)*
Census Region Dummies					
South	0.109	0.008	-0.523*	0.661	0.029
	(0.479)	*(0.478)*	*(0.309)*	*(0.427)*	*(0.302)*
Northeast	0.55	0.472	0.015	0.039	0.552
	(0.479)	*(0.466)*	*(0.335)*	*(0.449)*	*(0.343)*
Midwest	-0.003	-0.234	-0.55	-0.616	-0.544
	(0.521)	*(0.534)*	*(0.335)*	*(0.388)*	*(0.373)*
Observations	48	48	48	45	48
R^2	0.54	0.52	0.56	0.5	0.49

Notes: Entries in table are point estimates from log-standardized regressions. Robust standard errors are below in parentheses. Columns contain regressions employing different dependent variables. Document-frequencies are the ratios of documents found with the keyword "corruption" and the name of the state over the number of all documents found with the name of that state. Regressions exclude Washington State and Georgia (see text)

* significant at 10%; ** significant at 5%; *** significant at 1%

[a] Convictions correspond to Federal Department of Justice convictions on corruption charges of state officials (as used in Glaeser & Saks, 2006)

[b] per *inhabitant* corresponds to dividing the number of convictions by the population of the state.

[c] per *public employee* corresponds to dividing the number of convictions by number of public employees the state.

[d] From (Boylan and Lang, 2003)

TABLE S11

Corruption Regressions with Additional Predictors Significant at the 5% Level in (Glaeser and Saks, 2006)

Dependent Variable: Corruption as measured by	Convictions per Inhabitant 76-02	Convictions per Public Employee 76-02	Internet Document Frequency	Newspaper Document Frequency
Racial Dissimilarity	0.402**	0.343	0.288	0.346*
	(0.169)	*(0.209)*	*(0.206)*	*(0.204)*
Share Black	0.381***	0.371**	0.317*	0.367
	(0.132)	*(0.155)*	*(0.183)*	*(0.221)*
Local Share of Gov. Employment	1.112	2.012	1.793	1.306
	(1.368)	*(1.483)*	*(2.102)*	*(1.831)*
Integrity ranking, 2002	-0.025***	-0.026***	-0.015*	-0.019*
	(0.007)	*(0.008)*	*(0.008)*	*(0.011)*

Notes:

Entries in table are point estimates from log-standardized regressions. Robust standard errors are below in parentheses. Variables in this table are those found by Glaeser and Saks (2006) to be significant at the 5% level in tables other than their Table 4 (which we replicate in our Table S6). These regressions also control for 1970 income, education, population, share government employment, urban share, and regional dummies. Regressions exclude Washington State and Georgia (see text).

* significant at 10%; ** significant at 5%; *** significant at 1%