PROXYING FOR UNOBSERVABLE VARIABLES WITH INTERNET DOCUMENT-FREQUENCY

Albert Saiz
Massachusetts Institute of Technology

Uri Simonsohn
University of Pennsylvania

Abstract
The internet contains billions of documents. We show that document frequencies in large decentralized textual databases can capture the cross-sectional variation in the occurrence frequencies of social phenomena. We characterize the econometric conditions under which such proxying is likely. We also propose using recently-introduced internet search volume indexes as proxies for fundamental locational traits, and discuss their advantages and limitations. We then successfully proxy for a number of economic and demographic variables in US cities and states. We further obtain document-frequency measures of corruption by country and US state and replicate the econometric results of previous research studying its covariates. Finally, we provide the first measure of corruption in American cities. Poverty, population size, service-sector orientation, and ethnic fragmentation are shown to predict higher levels of corruption in urban America. (JEL: H00, J11, C81, B40, D73)

1. Introduction

A number of research projects in economics are hampered by data availability or plagued by omitted-variables bias. Economists and other social scientists have become active in the search for new sources of information with which to describe data patterns and test hypotheses that hitherto remained unexplored. In this paper, we provide economists and other researchers with a new source of *quantitative* data that uses the information contained in decentralized text databases, such as the internet and digital news repositories, which contain billions of documents. Concretely, we study if there is useful information in the frequency with which different topics are written about.

Assuming that, ceteris paribus, the more often a phenomenon occurs the more likely somebody is to write about it, aggregate measures of what large numbers of people write about in a decentralized fashion, *document-frequency*, should be correlated with the relative frequency with which the discussed phenomena have

The editor in charge of this paper was edited by Stefano DellaVigna.

Acknowledgments: This is an improved version on an older circulating working paper (Saiz and Simonsohn, 2007). We thank participants at departmental presentations at Wharton, Berkeley, and IZA-Bonn, and at NARSC and SJDM conferences, the editor, and three referees for useful comments. Remaining errors are ours. Saiz acknowledges support from the Research Sponsors Program of the Zell/Lurie Real Estate Center at Wharton. Shalini Bhutani, David Kwon, Caleb Li, Joe Evangelist, and Blake Wilmarth provided excellent research assistance. Saiz is also a Research Fellow at IZA.
E-mail: saiz@mit.edu (Saiz); uws@wharton.upenn.edu (Simonsohn)

Journal of the European Economic Association February 2013 11(1):137–165
 DOI: 10.1111/j.1542-4774.2012.01110.x

occurred, *occurrence-frequency*. Here we show that such hypothesized correlation can be exploited to proxy for the occurrence-frequency of economic and social phenomena that are difficult to observe.

We do not expect, of course, document-frequency to be correlated with occurrence-frequency in all circumstances. We do believe, however, that it is possible to judge ex-ante whether such an association is likely. In fact, we propose a simple econometric framework that leads to eight data checks that researchers can use to assess if a given document-frequency is likely to be a valid proxy for a particular occurrence-frequency. We also propose the use of *placebo* keywords to asses if spurious correlation is an issue.

For reasons outlined in our econometric framework, we focus on demonstrations that involve proxying for variation in the relative occurrence of a given social phenomenon across different locations.

For each occurrence-frequency variable to be proxied we construct a document-frequency proxy based on documents on the Internet (as indexed by the search engine *Exalead*) and another based on articles printed in US newspapers (as stored in the newspaper database *Newsbank*). Somewhat surprisingly, the two proxies are typically very similar.[1]

We operationalize document-frequency as the ratio of the number of documents containing a keyword relevant to the economic or social phenomenon of interest in textual proximity to the name of the target location, over the total number of documents containing the name of such location. For instance, in January of 2009 Exalead had indexed around 2,500 web documents with the word "corruption" in textual proximity to the word "Sweden", out of the nearly 16 million with "Sweden" in them. For "Russia" in contrast, a word identifying a more corrupt country contemporaneously, these figures were about 28,000 and 22 million respectively.

We begin the assessment of document-frequency measures by proxying for five variables whose true value is known with precision: percentage of population that is Hispanic, African-American and immigrant, and the crime and poverty rates across US states and cities. We obtain strong statistically and economically significant positive correlations with occurrence-frequencies: an average correlation of 0.56 for state-level variables and of 0.38 for city-level ones.

We then show the usefulness of the method as a way to complement other incomplete or unreliable data sources by imputation (proxying for religious denomination shares in the United States) or to generate indexes of latent variables based on factor analysis (proxying for the share of gay population in US cities).

Finally, we proxy for corruption, a variable that is difficult to measure, is of interest to economists and other social scientists, and has considerable variation at different levels of aggregation. At the country level, our document-frequency proxy is

1. We employ Exalead because it is one of the few engines offering (reliable) "proximity" searches, which restrict documents to contain two keywords within a certain distance from each other (in Exalead's case 16 words). Correlations obtained using alternative search engines without the proximity are somewhat weaker, but always significant. The interested reader can consult the Online Appendix for details.

highly correlated with the leading indicator of corruption published by Transparency International as well as with other similar proxies. To assess the extent to which such correlation is sufficiently high for our measures, we attempt to replicate published econometric research that employed Transparency International's measure as the dependent variable. To this end we re-estimate OLS regressions published in a recent literature review on the economics of corruption (Svenson 2005). Our results prove qualitatively identical and quantitatively very similar to those obtained in the literature.

We then repeat this exercise for corruption in US states. Our measure of corruption is positively and significantly correlated with the existing measures of state-level corruption (Boylan and Long 2003; Glaeser and Saks 2006). In addition, employing our measure as a dependent variable we successfully replicate the existing published results establishing the econometric determinants of state-level corruption.

Finally, we further demonstrate the ability of document frequencies to allow economists and other social scientists to study otherwise unobservable phenomena by creating the first index of corruption for US cities. This allows us to study for the first time the econometric determinants of urban corruption in America. We find that larger, poorer, service-oriented, minority-dense cities outside of the South tend to be more victimized by corrupt officials.

The paper contributes to research efforts that focus on garnering, combining, and aggregating different sources of information (Clemen 1989; Surowiecki 2004) as illustrated by the prediction markets literature in economics (Wolfers and Ziztewitz 2004).

Our research is especially related to a growing literature in economics and other disciplines that employs text-mining techniques in order to obtain word frequencies in large databases of documents and uses them to make inferences. This research has concentrated primarily on unveiling information already contained *within* the texts, or to make inferences specifically about the authors of the documents being queried, be it about their biases (Gentzkow and Shapiro 2010), beliefs (Antweiler and Frank 2004; Bangerter and Heath 2004), preferences (Liu 2006; Godes and Mayzlin 2004), or sentiments (Tetlock 2007). Online downloading and search behavior have also been used to predict scientific citations (Perneger 2004) and trends in software patents (Rech 2007). Our paper pushes the scope of text-mining approaches by demonstrating that there can be *real-world* informational content in the corpus of texts, but irrespective of the intentional content of any given document and without the need of strong behavioral assumptions about their individual writers.

In parallel to our early research (Saiz and Simonsohn 2007),[2] a new literature has recently emerged that uses online search behavior to predict associated offline behavior. While document-frequencies are based on the *supply* of information on large textual databases, this new literature focuses on the *demand* for information contained in consumer patterns of behavior in search engines. Ginsberg et al. (2009) and

2. A number of papers have used or cited the methodology laid out in our early working paper, including Berger and Heath 2008; Farrall 2009; DellaVigna and La Ferrara 2010; Feldman, Goldenberg and Netzer 2010.

Pelat et al. (2009) use the number of times that users search for terms associated with specific illnesses in Google® to predict subsequent volumes of doctor visits. Choi and Varian (2009a, b), and Da, Engelberg, and Gao (2011) use search behavior to predict the demand for products for which consumers tend to do online research prior to buying (such as real estate, cars, and stocks).

The existing research along these lines has so far been limited to using *demand for information* (DellaVigna and Gentzkow 2010)—as measured by search volumes—to analyze or forecast high-frequency movements of flow variables in situations where offline *behavior* (e.g. renting a car) will typically follow online *behavior* (e.g. searching for information about car rentals). In addition to the *supply* measures that we introduced, we further propose expanding the scope of such *demand*-side internet-search intensities to measure variation in fundamental stock variables. To do so we generate analog proxies constructed with search volume indexes (SVI) about a topic in the target geographical area. We find that *demand* measures can also work in situations where searches are associated with individuals looking for the *local* consumption of goods or services that cater specifically to the community whose frequency we are measuring. For example, we find that in metropolitan areas with more Catholics, people tend to search more often for sentences including the keyword Catholic, looking for information about local schools and religious services. However, we show that SVI measures may not be good proxies absent a consumption motivation of the group that is cognate (or unequivocally associated) with the keyword of interest.

In relation to the previous literature we therefore make four notable contributions. First we demonstrate that analyses of document-frequencies need not be limited to making inferences about the authors of the written text, about the specific events described in the text, or to forecast offline behavior that tends to follow online behavior, but more generally to proxy for the relative prevalence of *any* fundamental variable that can be expressed in frequencies. Conceptually we suggest that large databases of documents (such as the internet) can be thought of as records of large numbers of independent witnesses of phenomena that may have happened to themselves or to others. This may increase the value to economists and social scientists of the easy access they now enjoy to large databases of documents without the need of making strong assumptions about the relationship between online and offline behavior. Indeed, the method is applicable to the historical research of large textual databases that preceded the internet. Second, we advance the conditions under which such proxying is likely to be valid, aiding future researchers in their decision on whether to use document-frequencies as proxies. Third, we validate empirically the use of document-frequencies with several demonstrations. Fourth, we show that, in some circumstances, researchers can use the demand-side of the market for information to generate proxies based on internet search volumes, and provide guidance as to where to use them in combination with document-frequencies.

The paper proceeds as follows. In Section 2, we develop a simple econometric framework to help researchers identify the situations in which it is justified to use document frequencies as proxy variables, as well as to warn about some of the pitfalls involved. Section 3 demonstrates that document-frequencies are strongly correlated

with major economic and social variables in the US context. Next, we further show how the econometrician can use document-frequency proxies to complement existing data sources or to generate indexes of latent variables (Section 4). In Section 5 we replicate the econometric results of previous papers studying world and US state corruption. Our methodology also allows us to investigate city-level corruption for the first time. In Section 6 we compare document-frequency proxies to proxies obtained using internet search counts, thus linking our research to the parallel literature on search-volume frequencies. Section 7 concludes.

2. Econometric Framework

In this section we describe the econometric conditions that must hold to obtain useful proxies, with a data-check summary after each such condition. We shall denote transformations (in a sense specified in what follows) of the occurrence-frequency of phenomenon p in location l by $Y_{p,1}$. The corresponding transformations of the document-frequency obtained by querying a document database for the relative frequency of a keyword k appearing in textual proximity to the target location will be denoted by $\hat{Y}_{p,l,k}$ (we utilize subscripts only when needed). We focus on a linear first-order approximation to characterize the relationship between Y and $\hat{Y}$:

$$\hat{Y}_{p,l,k} = \alpha_{p,k} + \beta_{p,k} Y_{p,l} + \varepsilon_{p,l,k}, \tag{1}$$

where $\alpha_{p,k}$ is a phenomenon-keyword fixed effect, $\beta_{p,k}$ corresponds to the impact of the occurrence of phenomenon p on the frequency of documents containing keyword(s) k, and $\varepsilon_{p,l,k}$ is the residual.

Equation (1) is useful for organizing our discussion of various data checks that can be performed to assess the ex-ante likelihood of a high correlation between $\hat{Y}$ and Y. A very intuitive problem can arise if α (the keyword fixed effect) varies across queries. This is why we focus on cross-sectional variation on a given phenomenon (e.g. murder rates across cities) rather than across phenomena (e.g. murder versus car-theft rate within a city). For this same reason, we conduct all our searches in English, independently of the target country.

DATA CHECK 1. Do the different queries maintain the phenomenon and keyword constant?

Our basic premise, that all else constant, the occurrence of a phenomenon increases the likelihood that a document about it will be written, is equivalent to assuming that $\beta_{p,k} > 0$. Two data checks can be used to assess the validity of this premise. The first is straightforward: the variable of interest must be expressed in terms of a relative frequency. The second is that the keyword chosen be more likely to be employed following the occurrence than the non-occurrence of the phenomenon of interest.

The keyword "education" exemplifies violations of both requirements. First, "education" characterizes a term which does not have a frequency interpretation.

Second, both an increase and a decrease in the quality of education in a given location may lead to more documents with the keyword "education".

The second requirement can be assessed empirically (as we do in the next section) by sampling the content of the documents resulting from a given query and assessing whether keyword k is often utilized to demark the non-occurrence of Y.

DATA CHECK 2. Is the variable being proxied a frequency?

DATA CHECK 3. After sampling the contents of documents found: is the keyword k employed predominately to discuss the occurrence rather than non-occurrence of phenomenon p?

The variable $\varepsilon_{p,l,k}$ in equation (1) captures factors that influence $\hat{Y}_{p,l,k}$ other than $Y_{p,l}$. Here, we discuss three such factors: sampling error, measurement error, and violation of the "redundancy-condition" for proxy variables.

Note that it is important to use a textual proximity algorithm in order to minimize false positives in the number of documents assigned to a location. The fact that the target location and keyword appear in textual proximity in the document makes it more likely that the keyword is used in reference to the location of interest. Furthermore, conditional on match quality, sampling error is also reduced as sample size increases and consequently sampling error will play a smaller role for topics and locations where the number of documents is large. We bootstrapped our data to assess what is "large enough" (henceforth, the interested reader can consult the Online Appendix for details of calculations that we report in the text but which we do not display in the paper). Adding two alternate non-overlapping monthly random sequences of newspaper articles, we note that the mean root square deviation of the correlations in our examples with respect to their final value was quickly reduced to only around 20% after an average number of 100 documents containing the keyword by location.

DATA CHECK 4. Is the average number of documents found large enough for variation in document-frequency to be driven by factors other than sampling error?

Another cause for a large noise-to-signal ratio is lack of variation in the underlying occurrence-frequency. To exemplify this problem we proxied for cancer rates across US states and countries employing the document-frequency of the word "cancer" in proximity to the name of the target locations. We expected cancer rates to vary much more across countries than US states and hence for document-frequency to be a better proxy for the former. Data from the Center for Disease Control and GLOBOCAN confirmed both expectations. The coefficient of variation for cancer rates across states is 0.15 compared to 0.7 across countries, and in fact the correlation between occurrence and document-frequency was much higher for variation across countries 0.34 ($p < .01$) than across states -0.06 (not significant).

Noise-to-measurement problems are especially likely to arise in the context of trying to capture temporal variance in relatively short panels of data. Therefore,

researchers should be skeptical of using relatively noisy proxy variables (document-frequencies or any other) to capture high-frequency changes in flow variables.

Most fundamental attributes of a location are typically stock variables and evolve slowly, however. For example, the correlation between measures of country corruption, as summarized by the World Bank from different sources, in 2000 and 2009 was 0.94. *Changes* in perceived corruption (or other slowly changing fundamental economic, social, or demographic characteristics of a location) will certainly be too small relative to their levels for them to be reliably captured using proxy variables in short panels. Nevertheless, as digitized text databases increase in temporal span, using document proxies to capture low-frequency *changes* in otherwise unmeasured social phenomena will no doubt be fruitful.

DATA CHECK 5. Is the expected variance in the occurrence-frequency of interest high enough to overcome the noise associated with document-frequency proxying?

One likely source of measurement error is keywords with multiple meanings leading to false positives; that is, to documents that do contain k but which are not actually about p. This can be easily fixed by replacing a keyword for a synonym with fewer other meanings (for instance using "African Americans" rather than "Blacks").

DATA CHECK 6. After inspecting the content of the documents found: does the chosen keyword have as its primary or only meaning the occurrence of the phenomenon of interest?

The final aspect of ε that we discuss deals with its possible correlation with covariates of Y. This could be a problem if $\hat{Y}_{p,l,k}$ is estimated to learn about the relationship between $Y_{p,l}$ and other variables, X_l. A prerequisite for such use of proxy-variables is that $\text{Cov}(X, \hat{Y} \mid Y) = 0$ (Wooldrige 2001): conditioning on occurrence-frequency, document-frequency should be uncorrelated with the covariates of interest.

We consider two possible violations of this condition. The first occurs if X directly impacts $\hat{Y}$, independently of Y. As an example consider a researcher interested in the possible link between gun ownership and crime, but who lacks data on gun ownership. In this case, Y_l = *gun ownership in city l*, which could be proxied by $\hat{Y}_{k,l}$ with k = "guns". If the tendency to write about guns increases not only as more guns are owned, but also as more guns are used (e.g. in violent crime), then the correlation between the two will be a biased estimate of the relationship between gun *availability* and crime and specifically biased towards the relationship between gun *use* and crime.

Fortunately, this potential problem can be empirically assessed and addressed. One way to diagnose it is to conduct queries that combine keywords for the occurrence of interest and its covariates (e.g. k = "*gun* AND *murder*"); the greater the share of documents that include the keyword for the covariate, the greater the potential bias. If a problem is identified, k can be modified to reduce or eliminate it by, for example, employing keywords less likely to be used only in association with X (e.g. "*gun show*") or by explicitly requesting the search engine to exclude keywords associated with X (e.g. excluding "*murder"*). Comparisons of the results obtained when such

corrections are and are not implemented should provide guidance of the extent to which $\text{Cov}(\hat{Y}, X \mid Y) \neq 0$ is driving the results, as we show in more detail in the Online Appendix.

DATA CHECK 7. After inspecting the content of the documents found: does the chosen keyword also result in documents related to the covariates of the occurrence of interest?

The second scenario under which the redundancy condition may be violated is the presence of an omitted variable Z which affects both X and $\hat{Y}$ independently of Y. For example, suppose that there is more debate about socioeconomic issues in more cosmopolitan cities. Estimates of the correlation between a given covariate X, for instance average years of schooling, and the document-frequency of a given socioeconomic issue like "poverty" ($\hat{Y}$) will be biased towards the relationship between schooling and cosmopolitanism (Z)—that is, $\text{Cov}(\hat{Y}, X)$ will be biased towards $\text{Cov}(Z, X)$.

To fix this problem additional keyword searches can be conducted to proxy either for the underlying omitted variable (e.g. "cosmopolitan") or for the suspected latent variable influenced by the omitted one (e.g. "socioeconomic") and assess the impact of controlling for this additional document frequency on the parameter estimates of interest.

DATA CHECK 8. Are there plausible omitted variables that may be correlated both with the document-frequency and its covariates? If so, control for the omitted variable with an additional placebo document-frequency variable.

3. Proxying for Major Economic and Social Variables

3.1. Data Checks

To conduct this initial demonstration we selected variables capturing very salient socioeconomic dimensions in the United States and readily available at the state and city level (we encourage researchers to independently assess document-frequencies as proxies for other variables that comply with the data checks). In particular, we constructed proxies for the share of the population that is African American, Hispanic, and foreign born, and for both murder and poverty rates (see Online Appendix for more details on data sources and procedures). We define document frequencies by calculating the following ratio: the number of pages containing the relevant keyword within 16 words of the target location divided by the total number of pages containing the name of the target location. We used all cities with a population of 100,000 or more in the 2000 Census and all 50 states as locations (we present separate results for the more important—and reduced—group of cities with 250,000 inhabitants or more). We excluded cities that have the same name as another city of more than 100,000 inhabitants, such as Arlington and Springfield.

We begin by exemplifying the use of the econometric data checks with these variables, and then present the correlations between our proxies and the true values.

The variables of course pass Data Checks 1 (keywords are kept constant across locations), and 2 (occurrence of the phenomena can be expressed in relative frequency terms). For Data Checks 3 (keyword is more commonly used for occurrence rather than non-occurrence of phenomenon) and 6 (keyword's primary meaning is that of the phenomenon of interest) we conducted searches for each of the keywords in proximity to the word "city" and examined the contents of the first 50 documents found. For "African American" and "Immigrants" all 50 documents were true positives (e.g. *Cleveland's African American Museum* and the *Coalition for Humane Immigrant Rights* in Los Angeles). For "Hispanics" and "poverty" 49 out of 50 were true positives. For "murder", in contrast, only 14 out of 50 pages made direct allusion to actual murder cases or murder rates; many documents referred to murder mystery clubs, TV shows, or songs.

In the pool of 250 documents sampled, no document used a keyword in a sentence that discussed the absence or reduced occurrence of the phenomena (e.g. "no immigrants" or "lack of poverty"). Data Check 3 is hence passed by all keywords, while "murder" has some difficulty with Data Check 6.

Data Check 4 requires raw document-frequency to be high enough that variation in relative frequencies across locations can have a reasonable signal-to-noise ratio. In our data, the cross-city average number of internet documents found for a given keyword ranged between 476 (for "*poverty*" at the city level) and 35,957 (for "*African Americans*" at the state level). To investigate if these numbers are large enough in this context we generated two alternative monthly samples using newspaper stories and aggregated them in a random sequence. We found that *the mean root square deviation of the correlations in our examples with respect to their final value was quickly reduced to only around 20% after an average number of 100 documents containing the keyword per location*. Nevertheless, it is always true that larger textual samples improved the strength of the relationship between document and occurrence frequencies (see supplemental materials in the Online Appendix).

Data-check 5, requiring occurrence-frequency to experience substantial variation, will typically consist of a qualitative a-priori assessment, as researchers will not have access to the variable they desire to proxy for. For the variables in this demonstration, however, we can directly assess the variation in occurrence-frequency. The coefficients of variation are relatively high across the board, hovering around 75%–110%. Poverty is the notable exception, with a coefficient of variation of just 9% at the state level. We should expect, therefore, that poverty's document-frequency will be less strongly correlated with its occurrence-frequency.

Data Checks 7 and 8 do not apply here since we are not estimating regressions. We take the natural logarithms of document and occurrence-frequencies in order to approximate the relationship in equation (1), although we remark that Spearman rank correlations (insensitive to monotonic transformations) were always very similar to Pearson correlations. Furthermore, log-log specifications are commonly used in empirical research in economics and fit the data very well (see online supporting

information for details). For comparability between two variables measured in different units, throughout the text, we use standardized versions of Y and $\hat{Y}$.

3.2. Results: Document-Frequencies

Table 1 shows the correlations between document-frequency and occurrence-frequency for these five variables. All correlations are positive, with 28 out of the 30 being significant at the 5% level.

Considering that all five variables are related to socioeconomic status it is possible that rather than five independent demonstrations, the previous correlations capture the *same* correlation between document-frequency and occurrence-frequency of low socioeconomic status, five times. A more troubling concern is that this single correlation could be spurious.

For example, one may worry that large numbers of documents are written about African Americans in relation to Philadelphia not because of Philadelphia's large African American community, but because of Philadelphia's large Democratic Party voter base, which might lead to greater discussion of socioeconomic issues in general, *including* those pertinent to the African American community.

To address this concern we computed cross-correlations of document-frequency and occurrence-frequency of African Americans with the occurrence-frequency of the remaining four variables. This is the variable with the highest correlation between document and occurrence frequencies; under the null of a common latent spurious factor the problem here should be the largest. Contrary to the notion that a single latent variable drives all correlations in Table 1, several of the cross-correlations are negative. Furthermore, they are similar to the cross-correlations within occurrence-frequencies. For example, the correlation of the document-frequency of African Americans and the occurrence-frequency of Hispanics between cities is $r = -0.40$, while the correlation of the *occurrence-frequency* of African American with the occurrence-frequency of Hispanics is $r = -0.54$.

An alternative way to address the concern of a single latent variable driving our five demonstrations, as suggested in the econometric framework, consists of controlling for the suspected omitted variable with an additional placebo document-frequency proxy. Because we are concerned with an overall tendency to discuss socioeconomic issues, we estimated the relative document-frequency of the placebo keyword "socioeconomic" for all states and cities and used it as a control in regressions of document-frequency on occurrence-frequencies. We find that partial correlations between document-frequency and occurrence-frequency that control for the relative frequency of "socioeconomic" are very similar to the raw correlations. In particular, the average partial correlation between document-frequency and occurrence-frequency is $r = 0.41$, across the five variables for states, and $r = 0.44$ for large cities. The respective raw correlations were $r = 0.52$ and $r = 0.37$.

The previous results establish that document frequencies are strongly correlated with occurrence frequencies *on average*. To examine whether such relationships

TABLE 1. Correlations between occurrence, document, and search frequencies.

	Internet documents			Local newspapers			Internet search volume indexes (SVI)		
		Cities			Cities			Cities	
	US states	pop. > 100k	pop. > 250k	US states	pop. > 100k	pop. > 250k	US states	pop. > 100k	pop. > 250k
African-Americans[a]	0.70	0.43	0.67	0.82	0.50	0.61	0.74	0.24	0.37
Hispanics[a]	0.50	0.43	0.43	0.74	0.48	0.56	0.74	−0.14*	−0.10*
Immigrants[a]	0.51	0.37	0.44	0.69	0.40	0.46	0.74	0.49	0.36
Poverty rate[b]	0.41	0.34	0.31	0.37	0.26	0.20*	−0.15*	0.06*	0.00*
Murder rate[c]	0.48	0.29	0.26	0.36	0.13	0.02*	0.47	0.23	0.46
Average	**.519**	**.375**	**.422**	**.596**	**.354**	**.371**	0.51	0.17	0.22
N	50	227	62	50	227	62			

Notes: Entries in table are correlations between the occurrence and document-frequency or search volume indexes for each variable described in the first column. Occurrence frequencies are defined as the log-standardized population shares of the variable of interest in the target location (e.g. percentage of Hispanics in Georgia). Internet document frequencies are log-standardized ratios of the number of documents citing the phenomenon in textual proximity to the target location, divided by the total number of pages citing the target location. Internet search volume indexes are provided by Google and capture the relative frequency of searches for the keyword of interest originating in the target locations. Internet document frequencies are obtained with the search engine Exalead® while Newspaper frequencies are obtained with Newsbank®. Correlations are significant at the 5% level unless stated otherwise.

*Correlation not significant at 5%.

[a]As percentage of the overall population reported 2000 US Census.

[b]Poverty rate is the percentage of households below half the median of state income measures.

[c]Murder rate is the average murder rate per 10,000 in 2000–2005 according to the FBI's Uniform Crime Reports.

are monotonic we further estimated spline regressions and found that for all five quintiles of occurrence-frequency, the relationship with document-frequency was positive.[3]

3.3. Search Volume Indexes as Proxies for Local Socioeconomic Attributes

Subsequently to our initial research (Saiz and Simonsohn, 2007) a new literature has emerged that uses internet search volume indexes (SVI) to forecast economic and social phenomena. The number of times that a specific text has been searched constitutes a direct measurement of online behavior. This allows researchers to capture the prevalence of offline *behavior* that will typically follow such online *behavior* (Ginsberg et al. 2009; Pelat et al. 2009, Choi and Varian 2009a,b; Da, Engelberg, and Gao 2011).

In this section we are interested in increasing the scope of SVI by treating the search logs as yet further massive textual databases containing word frequencies in order to proxy for cross-sectional variation in fundamental economic and social variables that are unlikely to change at high frequencies. To do that, we use the Google Insights© application, which allows us to retrieve data on SVI.

Unfortunately, there are insufficient search volumes where the relevant keywords and target location appear in the same sentence in Google Insights. However, the application allows researchers to download the number of queries by the approximate location of the individual undertaking the search (country, state, and metro areas). Therefore, we use the log-standardized version of the Google Insights indexes (which are based on relative comparisons of total number of searches using the keyword of interest divided by total search activity) by country, state, or metro area to proxy for the relative prevalence of the phenomenon signified by the keyword.[4]

The results are presented in the last three columns of Table 1. In these applications, SVI measures show themselves less reliable than their document-frequency counterparts. Hereafter, we present similar columns with SVI correlations to facilitate comparisons by the reader throughout the paper's tables or in the text. In Section 6 we summarize how these correlations compare to those obtained from document-frequencies overall, and suggest the environments in which these informational demand-side proxies can be of use to researchers.

3. In an online supplemental materials section we also show that one can substitute the occurrence frequencies in these examples by document frequencies as noisy control variables in an OLS setup.

4. Note that this constitutes a departure from the approach that we advocate elsewhere in the paper: a textual proximity algorithm is more likely to minimize false positives. In contrast, by using the relative prevalence of a word by origin destination we are allowing the search "corruption in Washington DC" from a household in Philadelphia to be attributed to the latter metropolitan area (a false positive).

4. Alternative Applications

4.1. Complementing Existing Data

We next illustrate the use of document-frequency proxies *to complement*, rather than substitute for, available data with missing or mismeasured observations. For this econometric application we focus on denominational and religious shares by metropolitan area in the United States. The main source of information as to religious shares comes from the Religious Congregations and Membership Study (RCMS), completed by the Association of Statisticians of American Religious Bodies (ASARB). Their measure of the number of adherents by religious group is calculated by obtaining estimates from the ministers of local congregations (Finke and Scheitle 2005). We focus on the main congregations and religious groups in the United States: Catholic, Episcopalian, Jewish, Lutheran, Methodist, Presbyterian, Mormon, Pentecostal, and Baptist.

Consistent with our data strategy and modeling, we focus on the shares of each denomination over the total number of religious adherents, and treat the ASARB data as the gold standard (albeit those data may not be absent of measurement error and biases). Our second measurement comes from the 2001 American Religious Identification Survey (Kosmin Mayer and Keysar 2001) under the auspices of the Institute for the Study of Secularism in Society and Culture of Trinity College. The sample was based on a series of national RDD (random digit dialing) surveys conducted through a private firm (ICR) as an addendum to their national telephone omnibus sample. In total, over 50,000 respondents were interviewed over a span of approximately four months, but we also have information about their spouses, if any was present in the household. ARIS relatively oversampled households in the largest 77 metropolitan areas (MSAs/PMSAs), but sample sizes in many of the other cities are small. From both ASARB and ARIS we calculate the share of adherents by denomination divided by the total number of religious adherents in the relevant metro areas. We then used Exalead® and Newsbank® to obtain the relative document frequencies of the denomination's adherents (e.g. Lutheran or Lutherans) in textual proximity to the name of the different cities. Since denominational shares are relative to each other, and in order to reduce noise in the denominator, we divide the number of each denomination's hits by the sum of all denominational hits (results using total number of city hits are very similar).

In Table 2 , Panel A, we show correlations between the different measures for US metropolitan areas. All correlations between document-frequencies and occurrence-frequencies (from ASARB) are positive, with unweighted averages of 0.33 for the internet and 0.42 for newspapers (0.41 and 0.48, respectively, population-weighted averages). In this application, the proxies based on SVI obtain slightly higher correlations.

Nevertheless, the correlation between the phone survey (ARIS) denominational frequencies and the ASARB data is much stronger, 0.66 on average across the denominations of interested. Therefore, as expected, the quality of the information provided by a carefully-conducted representative survey is superior to the proxies.

TABLE 2. Using document frequencies to impute religious shares.

Panel A

All available cities

ASARB religion data correlations with:	Internet documents	News	Phone survey (ARIS)	SVI
Catholic	0.415	0.631	0.885	0.548
Episcopalian	0.219	−0.012*	0.424	0.024*
Jewish	0.363	0.506	0.744	0.547
Lutheran	0.483	0.508	0.716	0.741
Methodist	0.402	0.468	0.764	0.698
Presbyterian	0.364	0.428	0.424	0.519
Mormon	0.531	0.563	0.766	0.597
Pentecostal	0.102*	0.224	0.417	−0.068*
Baptist	0.625	0.691	0.836	0.854
Average	0.389	0.445	0.664	0.496
Cities (N)	223	223	223	101

Panel B

	Cities with small sample in ARIS survey		
	Correlation of ASARB with document frequency imputation	Coefficients of multivariate regression with two variables: ASARB denominational share is the dependent variable	
	(1)	**(2)** Measure from ARIS survey	**(3)** Imputation from Internet and News
Catholic	0.607	0.710	0.285
Episcopalian	0.171*	0.322	0.668*
Jewish	0.508	0.303	0.900
Lutheran	0.673	0.399	1.033
Methodist	0.509	0.421	0.618
Presbyterian	0.519	0.069*	1.089
Mormon	0.673	0.369	1.023
Pentecostal	0.352	0.251	0.778
Baptist	0.795	0.592	0.633
Average	0.534	0.382	0.781
Cities (N)	68	68	

Notes: Panel A displays correlations between log-standardized frequencies of religious adherents (e.g.Catholic share) from the Association of Statisticians of American Religious Bodies (ASARB) with respect to the proxies. The internet and newspaper proxies are based on the ratio of number of documents containing the name of the religion adherent (e.g. Lutheran) in textual proximity to the name of the city divided by the total number of documents containing the name of the city. Internet search volume indexes (SVI) are provided by Google and capture the relative frequency of searches for the keyword of interest originating in the target locations. The phone survey is an estimate of denominational frequencies from a random-dialing sample (ARIS: American Religious Identification Survey). Panel B, column 1, shows similar correlations but this time using imputed values for ARIS religious frequencies: we use an "in" city sample with more than 200 survey observations and the document-frequency proxies to impute survey frequencies for the "out" sample (cities with low number of ARIS observations). Columns 2 and 3 show the coefficient of an OLS regression with the ASARB frequencies on the left-hand side and both the "out" ARIS survey observations (column 2) and the document-frequency imputations (column 3) on the right-hand side.

* Not significant at 5%.

However, we can combine the information from the proxies to obtain a more complete picture of the phenomena under study.

We provide two examples focusing on the document-frequency proxies in Panel B of Table 2. In the first example, we censor the ARIS survey by eliminating cities with less than 200 valid religion responses. This corresponds to a hypothetical experiment where the coverage of ARIS was reduced by 67 cities. We therefore assume that the econometrician has access to the censored survey data, but does not possess the *gold-standard* data from ASARB. Could we use document frequencies to supplement the existing data? A way to do this is by imputing the survey data for the missing cities. To do so, we ran an OLS regression with the survey (ARIS) data on the left-hand side and the document-frequency proxies on the right-hand side (newspaper and internet for all denominations) for the cities for which we have more than 200 survey responses (the observations in our hypothetical researcher's information set). We then used the predictions from the regression coefficients to impute values of the "missing" ARIS observations (67 cities with less than 200 valid survey values). In Table 2, Panel B, column 1, we show correlations between the *imputed* values and denominational shares from ASARB. All the correlations are positive and significant, with an unweighted average of 0.51. Note that we are purposefully using a rather unsophisticated data imputation technique, and yet showing that document-frequencies are useful to complement the survey data.

In Table 2, Panel B we also consider another scenario where, rather than missing, we do have complete city coverage in the survey, but simply assign less credibility to the data from cities with small samples (we focus again on cities with less than 200 observations, but the conclusions are not sensitive to this threshold). Columns 2 and 3 present the coefficients of an OLS regression where the ASARB reference data is the dependent variable on the left-hand side and *both* the ARIS survey data (estimated coefficient in column 2) and the imputation using document-frequencies (estimated coefficient in column 3) are included *simultaneously* as right-hand-side explanatory variables. In the sample of cities with relatively small survey samples, both the survey data and the imputation from document-frequencies are simultaneously predictive of higher corresponding denominational frequencies. This suggests that we could combine the information of both proxies (survey and document-frequency imputations) in order to improve inference when using data from cities with small survey samples.

More generally, document-frequency proxies can be used to *complement* other data sources in a wide array of more sophisticated econometric approaches to inference, including errors-in-variables IV (Carter and Fuller 1980); using multiple proxies as simultaneous controls (Lubotsky and Wittenberg 2006); Bayesian multiple imputation (Rubin 1996); and in structural equations with latent variable approaches (Bollen 1989).

4.2. Creating Latent Variable Indexes

Variables based on document-frequencies are noisy proxies for measurable or latent variables of interest to the econometrician. We can therefore combine the information

they provide to generate latent variable indexes based on factor analysis. Here we do this for the case of the percentage of gay population by metropolitan area. Certain metropolitan areas in the United States are known for their social tolerance and tend to attract individuals of gay sexual orientation. It is difficult to measure exactly the proportion of gay individuals in a city. For an approximation, we obtained data from the Census table PCT14 that details the share of total couples where both partners are male. Of course, this measure leaves out all gay individuals who are not in a stable relationship *and* cohabitating.

As *complementary* proxies for the share of gay populations we generate document frequencies for the words "gay" and "homosexual" as well as SVI proxies from internet users searching for the keyword "gay". In Table 3 we present the correlations between these proxies and the share of same-sex cohabiting couples in the 2000 Census for the US cities for which we can obtain the data, and for all states. All correlations with document-frequencies are positive and close to 0.25 in the unweighted sample of cities. SVI proxies are quite strong in this application, an issue to which we will return in Section 6. All correlations are generally stronger and significant when we weight the data by population.

TABLE 3. Using document frequencies and factor analysis to create a gay population index.

	Correlations with frequency of homosexual couples in US Census			
	Cities (unweighted) **(1)**	Cities (pop. weighted) **(2)**	*States (unweighted)* **(3)**	States (pop. weighted) **(4)**
"Gay" frequency: Internet	0.34	0.60	0.21*	0.49
"Homosexual" frequency: Internet	0.19	0.53	0.43	0.66
"Gay" frequency: Newspapers	0.27	0.48	0.18*	0.39
"Homosexual" frequency: Newspapers	0.16	0.32	0.21*	0.44
"Gay" internet search volume index	0.60	0.72	0.72	0.78
Document-based gay index	0.33	0.61	0.36	0.64
N	106	106	50	50

Notes: Entries in table are correlations between the occurrence and document-frequencies of proxies for the prevalence of gay population by metropolitan area or state. Occurrence frequencies are defined as the log-standardized population shares of cohabiting same-sex couples in the 2000 Census. Internet document frequencies are log-standardized ratios of the number of documents citing the keywords "gay" or "homosexual" in textual proximity to the target location, divided by the total number of pages citing the target location. Internet document frequencies are obtained with the search engine Exalead® while Newspaper frequencies are obtained with Newsbank®. Internet search volume indexes are provided by Google and capture the relative frequency of searches for the keyword of interest originating in the target locations. The Gay Index is the first principal component arising from a factor analysis of all document-frequencies. Correlations are significant at the 5% level unless stated otherwise.

*Not significant at 5%.

In the next-to-last row of Table 3, we present the correlations between the Census data and a gay index constructed using the first factor in a principal-factor analysis of the four *document-frequency* proxies. This provides us with a gay index that could be used to complement the same-sex couple data from the Census.

We next check if the information contained in the newly created gay index conforms to what we know about gay metropolitan areas. Black et al. (2002) study the urban economic determinants of the share of gay couples in the Census; they find that the only robust variable that predicts higher same-sex cohabitation by metropolitan area in their dataset is its median house value, which they interpret as a proxy for urban amenities. These authors focus on large cities—metropolitan areas with population of at least 700,000—arguing that "estimates for smaller cities are unreliable due to the low incidence of gay and lesbian couples in the data." While we cannot exactly replicate their sample, we execute an OLS specification where the log-standardized census same-couple frequency appears on the left-hand side and the explanatory variable is the log-standardized median house value in each metro area.[5] We use population weights in order to focus on the largest metropolitan areas and find, similar Black et al. (2002), that median housing prices are strongly predictive of same-sex cohabitation. The results suggest that an increase of one standard deviation in home values is associated with a half standard deviation increase (0.511) in Census gay-couple frequencies. The cross-sectional model fits the data quite well, with an R-squared of 0.45. The estimated coefficient is virtually the same—0.505—that we obtain using the document-frequency gay index variable as the dependent variable. Therefore, we reach an identical conclusion to Black et al. (2002) using the document-frequency-based gay index. Of course, one could combine the information of our newly-created gay index with the SVI proxies, Census data, and other sources to generate *even more sophisticated indexes*.

5. Proxying for Corruption: Econometric Applications

The results from the previous sections demonstrated that document-frequencies can be significantly correlated with the occurrence-frequency of major economic and social phenomena. In this section we examine whether such correlation can be exploited by the econometrician to construct proxies with which to explore the determinants of unobservable dependent variables. We decided to focus on corruption for two main reasons. First, doing so reduces possible concerns of data snooping to a minimum, as we require the exact same technique to replicate prior findings in settings with independent sources of variation (replicating studies by different sets of researchers who employ different measures of corruption and different sets

5. Remember that all data are log-standardized in order to facilitate comparison of magnitudes. The use of proxies does not generally allow us to determine the exact value of the parameters as measured in the scale of the original unobserved data; however, if we had an imperfect occurrence-frequency measure for at least a subset of the data, we could use it to convert our standard-deviation proxy measure into a percentage-point measure.

of predictors). Second, it characterizes the ideal application of document-frequency based proxying, as corruption is a phenomenon that is otherwise difficult to measure (Olken, 2009); Transparency International's Corruption Perceptions Index (CPI), for instance, averages information from 16 different surveys, some including over 4,000 respondents. Quantifying document-frequency, in contrast, is virtually free and can in principle be conducted at any level of aggregation.

5.1 Replication of Econometric Specifications with Country-Level Corruption

We conducted searches for "corruption" in proximity to the name of 154 countries, deflating the resulting number of documents by the total of documents containing a country's name. As hypothesized, the resulting measure proved positively correlated with the CPI ($r = 0.62$).

A possible concern is that the documents the search engines are finding are actually discussing Transparency International's CPI, which could mean that only when a measure already exists can document-frequency be employed. To address this possibility we conducted a new set of queries excluding documents containing the word "*transparency*", obtaining a similarly high correlation with the CPI ($r = 0.60$).

Similarly, we also found strong correlations between document-frequency proxies and other widely used measures of international corruption. In contrast, the SVI proxy performed relatively poorly in this application, obtaining a correlation of 0.34 with Transparency International's corruption measure.

Several papers have investigated econometrically the determinants of corruption across countries. Svensson (2005) provides a recent review of this literature, where he presents results from regressions using alternative measures of corruption as dependent variables and variables hypothesized to influence corruption by multiple economic theories as predictors. Table 4 in this paper combines the regression specifications reported in Tables 2, 3, and 4 in Svensson's paper (we utilize the same data sources for the predictors).

Because each predictor has a different set of missing observations, in the upper panel of Table 4 we report results from separate univariate regressions (each *cell* is a separate regression), and in the lower panel those from a single multivariate regression (each *column* is a separate regression).

In the four univariate regressions we obtain the same qualitative results with our document-frequency proxies and with the CPI. Furthermore, point estimates (recall that these are log-standardized regressions) and significance levels are quite close. The lower panel, with the multivariate regressions, also shows results similar for the CPI and for our document-frequency proxies, with the exception of the number of days to open a business which is not significant with the internet proxy but is with the CPI. The results using internet document frequencies suggest that one standard deviation increase in log income decreases corruption by about 0.9 standard

TABLE 4. Replication of regressions establishing correlates of country level corruption in Svensson 2005 (Tables 2, 3, and 4).

Dependent variable is corruption as measured by:	(1) Transparency International's Corruption Perception Index	(2) Internet Based Document Frequency (Exalead)	(3) Newspaper Based Document Frequency (NewsBank)	N
Only education in 1970 as a predictor				
Log (Average education [in years] 1970, adults 25+)	−0.679*** (*0.100*)	−0.527*** (*0.106*)	−0.446*** (*0.093*)	*96*
Only GPD per capita in 1970 as a predictor				
Log (real GDP in 1970)	−0.761*** (*0.060*)	−0.618*** (*0.074*)	−0.483*** (*0.089*)	*105*
Only Imports/GDP as a predictor				
Log (average[imports/GDP] 1980–2004)	−0.081 (*0.083*)	−0.0903 (*0.072*)	−0.196** (*0.079*)	*145*
Only days required to open a new business as a predictor				
Log (days to open new business)	0.601*** (*0.055*)	0.265*** (*0.070*)	0.341*** (*0.110*)	*84*
All four predictors				
Log (average education [in years] 1970, adults 25+)	0.072 (*0.096*)	0.180 (0.167)	0.167 (0.137)	
Log (real GDP in 1970)	−0.747*** (*0.139*)	−0.889*** (0.167)	−0.611*** (0.159)	*54*
Log (average[imports/GDP] 2000–2004)	−0.181** (*0.075*)	−0.205** (0.078)	−0.316** (0.120)	
Log (days to open new business)	0.271*** (*0.075*)	−0.050 (0.091)	0.215 (0.130)	

Notes: Entries in the table are point estimates from log-standardized regressions where corruption proxies at the country level are the dependent variables. Robust standard errors below parenthesis. Horizontal lines separate regressions employing different subsets of predictors. Sample sizes vary due to missing observations. Columns separate regressions employing different dependent variables. Document-frequencies are the ratios of documents found with the keyword "corruption" in textual proximity to the name of the country over the number of all documents found with the name of the country. Specifications replicate those published in Svensson (2005). See text for data sources.
*Significant at 10%; **significant at 5%; ***significant at 1%.

deviations; a similar increase in trade openness decreases corruption by 0.2 standard deviations.

5.2. Replication of Econometric Specifications with State-Level Corruption

We are aware of two existing measures of state-level corruption in the United States. The first, by Boylan and Long (2003), is based on a survey of 834 state House

reporters (with 293 responses from 45 different states). The second, by Glaeser and Saks (2006), is based on the number of government officials convicted for corrupt practices through the (federal) Department of Justice (DOJ).[6] Our document-frequency measure of corruption proved significantly correlated with both, $r = 0.44$ and $r = 0.59$ respectively (the two existing measures are correlated $r = 0$ .41 between them).

Table 5 replicates a regression specification from Glaeser and Saks (2006), for the three alternative measures of corruption just discussed.[7] The three significant predictors for the conviction-based measure are also significant predictors for our document-frequency measures and are of similar magnitudes. Most of the other point estimates are similar as well, except for "share of employees employed by the state" which is not correlated with the corruption conviction rate but is (negatively so) with the document-frequency measure. Using the internet document-frequency proxy, a one standard deviation change in inequality and income is associated, respectively, with 0.9 and 0.78 standard deviations increases in corruption. Conversely, a one-standard-deviation increase in inequality is associated with a decrease in corruption of about half a standard-deviation. The results obtained with the survey of house state reporters are not very dissimilar qualitatively, but many of the parameter estimates are not significantly different from zero. It is noteworthy that document-frequencies also replicate the other significant results in Glaeser and Saks (2006).

5.3. The Economic and Social Determinants of City-Level Corruption

We now turn to using internet document-frequencies to produce the first econometric assessment of corruption in US cities. We start by presenting a ranking of our corruption proxy for cities larger than 250,000 inhabitants in Table 6. The top ten cities are consistent with our priors on corruption; they include Chicago, New Orleans, Philadelphia and San Diego. Conversely, among the bottom ten we find cities seldom used as examples of corrupt local governments.

We next estimate regressions employing city-level corruption document-frequency as the dependent variable in Table 7. Given that there are no existing measures of city-level corruption, we use as a benchmark a state-level regression. Unfortunately, some of the state-level covariates are unavailable for cities (e.g. income inequality) and some lack variation (e.g. percentage of the population living in cities), so we estimated a new state-level regression so as to keep identical specifications.

While the results of the benchmark specification for state-level and city-level variation in corruption are different, both city- and state-level regressions weakly indicate that locations with lower income, larger populations, and more minorities and

6. Glaeser and Saks deflate the number of convictions by the state population. We deflate instead by the number of public employees in the state. See the Online Appendix for a detailed discussion with additional results.

7. In line with Data Check 4 both Washington state and Georgia were not included in the analyses because they lead to a high rate of false positives (with most articles found referring to Washington DC and the country of Georgia instead of the respective states).

TABLE 5. Replication of regressions establishing correlates of state level corruption in Glaeser and Saks 2006 (Table 4(1)).

Dependent Variable: Corruption as measured by	**(1)** Federal corruption Convictions per public employee (1976–2002)[a,b]	**(2)** Survey of journalists[c]	**(3)** Corruption Document frequency: *Internet*	**(4)** Corruption Document frequency: Newspapers
Income inequality	0.811***	0.344	0.983***	0.795***
	(*0.172*)	(*0.361*)	(0.210)	(*0.226*)
Ln(Income)	0.759***	0.599	0.827***	1.050***
	(*0.192*)	(*0.403*)	(0.256)	(*0.235*)
Share of population in state with 4+ Years of College	−0.835***	−0.642**	−0.366**	−0.521***
	(*0.156*)	(*0.243*)	(0.169)	(*0.168*)
Share of all employees employed by the state government	0.015	−0.052	−0.334***	−0.359**
	(*0.172*)	(*0.233*)	(0.120)	(*0.147*)
Ln(Population)	−0.02	−0.199	0.169	−0.137
	(*0.178*)	(*0.175*)	(0.123)	(*0.117*)
Share of population living in urban environment	0.255	0.660***	0.204*	0.263*
	(*0.184*)	(*0.145*)	(0.110)	(*0.154*)
Census region dummies				
South	0.008	0.661	−0.43	0.029
	(*0.478*)	(*0.427*)	(0.273)	(*0.302*)
Northeast	0.472	0.039	0.5	0.552
	(*0.466*)	(*0.449*)	(0.336)	(*0.343*)
Midwest	−0.234	−0.616	−0.238	−0.544
	(*0.534*)	(*0.388*)	(0.369)	(*0.373*)
Observations	48	45	48	48
R^2	0.52	0.5	0.56	0.49

Notes: Entries in the table are point estimates from log-standardized regressions where different corruption proxies at the US state level are the dependent variables. We mimic the econometric specifications in Glaeser and Saks (2006). Robust standard errors below parentheses. Document-frequencies are the ratios of documents found with the keyword "corruption" in textual proximity to the name of the state divided 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 and Sak 2006).
[b]per *public employee* corresponds to dividing the number of convictions by number of public employees in the state.
[c]From Boylan and Lang 2003.

immigrants tend to be more victimized by corruption. However, the coefficients using the DOJ data are very imprecisely estimated in the state regression.

On the contrary, the results using the city document frequencies are quite well defined and comparable across newspapers and the internet. A one-standard-deviation increase in log income is associated with a decrease in about 0.2 standard deviations in corruption: poverty begets corruption in urban America. This is, of course, fully coincident with the results from previous literature at other aggregation levels. Larger cities tend to be perceived as more corrupt (with a standardized impact of about 0.20). A one-standard-deviation increase in African American and immigrant shares

TABLE 6. Document-frequency corruption ranking at the city level (pop. > 250,000).

Rank	City	Rank	City
1	Las Vegas	31	Sacramento
2	New Orleans	32	Fort Worth
3	New York	33	Honolulu
4	Los Angeles	34	Long Beach
5	Chicago	35	Corpus Christi
6	San Diego	36	Austin
7	St. Louis	37	Milwaukee
8	Miami	38	Houston
9	San Jose	39	St. Paul
10	Philadelphia	40	Santa Ana
11	Oklahoma City	41	Minneapolis
12	Newark	42	Buffalo
13	Detroit	43	Portland
14	Cleveland	44	Wichita
15	Boston	45	Raleigh
16	San Francisco	46	Pittsburgh
17	Phoenix	47	Tampa
18	Riverside	48	Cincinnati
19	Atlanta	49	Tucson
20	El Paso	50	Anchorage
21	Memphis	51	Mesa
22	Baltimore	52	Charlotte
23	Virginia Beach	53	Nashville-Davidson
24	San Antonio	54	Albuquerque
25	Dallas	55	Tulsa
26	Denver	56	Indianapolis
27	Seattle	57	Colorado Springs
28	Oakland	58	Louisville
29	Lexington-Fayette	59	Omaha
30	Fresno	60	Jacksonville
		61	Anaheim

Notes: The table displays the ranking of cities with population of more that 250,000 inhabitants in the 2000 Census ranked by corruption internet document frequency. Document frequencies are the ratios of documents found with the keyword "corruption" in textual proximity to the name of the city divided over the number of all documents found with the name of that city.

is associated each, broadly speaking, with about a 0.15 standard-deviation increase in corruption. Minorities are more likely to be victimized by corrupt officials, probably exploiting the politics of ethnic fragmentation. These latter results are highly consistent with those found by La Porta et al. (1999) across countries, and by Olken (2006) in Indonesian villages.

One of the benefits of obtaining city-level data is the possibility of studying covariates that vary at a finer level of aggregation than at the state level. As an example we examine if industrial cities tend to experience more corruption. To this end we add as a predictor of city-level corruption the share of employment in the manufacturing sector, which proves—somewhat surprisingly—negatively associated with corruption (see fourth column). We hypothesize that this is due to reverse causation: metro areas

TABLE 7. OLS Identifying correlates of city-level corruption.

Independent variable: Corruption as measured by	(1) State-level occurrence Frequency (Convictions per public employee)	(2) City-level Internet document frequency	(3) City-level Newspaper document frequency	(4) City-level Internet document frequency	(5) City-level Internet document frequency	(6) City-level Internet document frequency
Log of income	−0.226	−0.167**	−0.231***	−0.156**	−0.156**	−0.158**
	(*0.171*)	(0.077)	(0.078)	(0.079)	(0.070)	(0.070)
Share workers in public administration	0.192	0.021	0.069	−0.016	−0.042	−0.04
	(*0.180*)	(0.053)	(0.050)	(0.055)	(0.053)	(0.054)
Log of population	0.083	0.232***	0.134**	0.205***	0.183***	0.185***
	(*0.229*)	(0.050)	(0.054)	(0.053)	(0.050)	(0.049)
Share African-American	0.381**	0.135*	0.089	0.154**	0.11	0.095
	(*0.171*)	(0.073)	(0.074)	(0.075)	(0.068)	(0.074)
Share foreign born	2.31	0.163*	0.187**	0.197**	0.174**	0.164**
	(*4.209*)	(0.083)	(0.076)	(0.085)	(0.074)	(0.076)
South	0.53	−0.401**	−0.113	−0.441**	−0.332**	−0.366**
	(*0.546*)	(0.170)	(0.165)	(0.173)	(0.164)	(0.175)
Northeast	1.113**	0.345	0.527*	0.368	0.295	0.245
	(*0.510*)	(0.231)	(0.287)	(0.228)	(0.211)	(0.222)
Midwest	0.684	−0.043	−0.169	0.07	0.093	0.072
	(*0.597*)	(0.206)	(0.167)	(0.217)	(0.215)	(0.212)
Share workers in manufacturing				−0.142*	−0.149*	−0.146*
				(0.082)	(0.079)	(0.079)
"Socioeconomic" document frequency					0.307***	0.312***
					(0.059)	(0.061)
State-level corruption						0.054
						(0.080)
Observations	50	224	224	224	224	224
R-squared	0.3	0.2	0.19	0.21	0.3	0.3

Notes: Entries in table are point estimates from log-standardized OLS regressions with corruption as the dependent variable. Robust standard errors below point estimates. Corruption in column 1 corresponds to the ratio of DOJ corruption convictions to the number of public employees in the respective state. The dependent variable in all other columns except (1) is the document-frequency of corruption at the city level. Column 5 adds the document-frequency of the word "socioeconomic" to the specification from column 1 to control for idiosyncratic differences in the tendency to discuss socioeconomic issues across cities. Column 6 adds the document-frequency of corruption at the State level to account for state level variation in corruption.

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

with corrupt politicians may see footloose industries move out or new industrial firms avoid moving in.

We next examine two possible econometric concerns regarding our city-level analyses of corruption. The first is the possibility that variation in our corruption proxy is driven by variation in the tendency to write about social issues in relation to different cities. For example, we find that larger cities are more corrupt on average; the concern here is that this correlation may result from people being more inclined to writing about economic and social issues in general with regards to larger cities. As suggested in our discussion of such problem in Data Check 8, we assess the potential importance of this concern by estimating the document-frequency of a placebo variable that may proxy for the omitted variable in question. In particular, in the fourth column we add as a control the document-frequency of the word "socioeconomic". Although it proves a significant predictor, the point estimates for the other variables remain largely unchanged, suggesting omitted variables of the kind we considered are not a problem in the original specification.

The second concern we address is the possibility that our city-level regression results capitalize on state-level variation in corruption. To address this concern in column 6 we control for our state-level document-frequency measure of corruption (i.e. the dependent variable in the first column). We find that (i) surprisingly, state-level document-frequency of corruption is not a significant predictor of city-level corruption (controlling for city observables), and (ii) more importantly, its introduction in the model does not greatly influence the point estimates of the other independent variables. This strongly suggests we are capturing variation in corruption above and beyond state-level corruption.

6. Document Frequencies and Internet Search Volume Indexes

In this section, we summarize and contrast the relative strengths of document-frequency and SVI proxies. There are two main noteworthy differences between the two measures. First, SVI measures are based on intentional internet search behavior. In contrast, gathering information from decentralized textual archives does not require any online behavior to exist, and can be performed on older textual pre-internet sources. It does not rely on individuals purposefully focusing on a specific search term: the frequency of words in large decentralized textual databases contains information beyond and above the *intentional* content of the individual texts themselves. Second, the document-frequency proxies are based on textual proximity. This maximizes the probability that the keywords make reference to the target locations. However, SVI are based on the location of the individuals performing the search, who may be actually looking for information about distant geographic areas. In this context the SVI proxies by origin area can only be expected to work well in situations where search volumes are: (i) associated with the prevalence of the phenomenon (similar to document frequencies); and (ii) *likely to have a strictly local scope*.

In Table 8 we summarize and compare the previous results in the paper with respect to SVI and *internet* document-frequency (IDF) correlations with actual data.

TABLE 8. Search volume indexes (SVI) versus document-frequencies: unweighted correlation summary.

PANEL A Google insights (SVI)				
		Cities		
	US states	pop.>100k	pop.>250k	World
African Americans[a]	0.74	0.24	0.37	
Hispanics[a]	0.74	−0.14*	−0.10*	
Immigrants[a]	0.74	0.49	0.36	
Poverty rate[b]	−0.15*	0.06*	0.00*	
Murder rate[c]	0.47	**0.23**	0.46	
Average	***0.51***	**0.17**	**0.22**	
Average religion		**0.50**		
Gay Census		**0.60**		
World Corruption (TI)				**0.34**
US State Corruption (Glaeser–Saks)	0.09*			

PANEL B Internet document frequencies (IDF) (Same sample as Panel A)				
		Cities		
	US states	pop > 100k	pop > 250k	World
African Americans[a]	0.70	0.43	0.67	
Hispanics[a]	0.50	0.43	0.43	
Immigrants[a]	0.51	0.37	0.44	
Poverty rate[b]	0.41	0.34	0.31	
Murder rate[c]	0.48	0.29	0.26	
Average	***0.52***	***0.38***	***0.42***	
Average religion		**0.39**		
Gay Census		**0.34**		
World Corruption (TI)				**0.70**
US State Corruption (Glaeser-Saks)	**0.59**			

Notes: Entries in Panel A displays correlations between log-standardized ocurrence frequencies and Search Volume Indexes (SVI). SVI are obtained from Google Insights and are defined as the number of Google searches containing the keyword of interest relative to all searches originating in the target metropolitan area. Panel B displays correlations between log-standardized occurrence and document-frequencies for each variable described in the first column, as laid out in the text and previous tables, this time replicating the sample available in Panel A. Internet document frequencies are obtained with the search engine Exalead®. Correlations are significant at the 5% level unless stated otherwise.

*Correlation not significant at 5%.

[a]As percentage of the overall population reported 2000 US Census.

[b]Poverty rate is the percentage of households below half the median of state income measures.

[c]Murder rate is the average murder rate per 10,000 in 2000–2005 according to the FBI's Uniform Crime Reports.

With respect to the main socioeconomic variables, the SVI proxy held its own on average at the state level, but was a bit less consistent, not always positive, and not always significant. However, at the metropolitan-area level the socio-economic SVI proxies performed relatively poorly, obtaining average correlations that were only 50% of the equivalent IDF. Moreover four of the ten correlations were not significantly different from zero, and two were actually negative. In contrast IDF correlations seem more consistent and are always positive and significant throughout all the applications.

This is, perhaps, not a surprise given that we expect more noisiness in searches that are not necessarily made in reference to the location where the search is performed.

The results with respect to country and state-level corruption using SVI were particularly disappointing. At the country level the correlations obtained were one half of those using IDF. At the state level, the correlation was basically close to zero: a lot of searches about corruption are geared towards finding about corruption elsewhere (note that there are insufficient search volumes to use a word-proximity approach).

SVI, however, did much better in proxying for fundamental social characteristics in the cases of religion and homosexuality. Religion searches obtained an average correlation that was 33% above the IDF, and the gay keyword correlated 74% more strongly with the Census same-sex-cohabitation data. The explanation comes from looking at the most common search phrases containing the keywords. In the case of religious denominations, most of the more common searches make reference to the denomination's church building or associated institution (e.g. Lutheran churches; or catholic school): presumably individuals looking for addresses to the local church or school. In the case of the keyword "gay", nine out of the top ten searches make allusion to explicit materials. Therefore in both cases searches are associated with individuals looking for the *local* consumption (in the neighborhood or at home) of community-specific goods or services.

In other words, SVI measures by point-of-origin of the search can be unreliable because the searchers do not necessarily look for information about local phenomena. In contrast the document-frequency approach relying on textual proximity ensures that a large amount of documents actually make reference to the phenomenon with respect to the target location. However, if internet searchers use a keyword prior to the *local* consumption of goods or services that are geared specifically to the community to which the keyword makes allusion to, then the number of searches can actually be a bellwether for the relative prevalence of such communities.

Therefore, it is always a reasonable idea to use document-frequency proxies (internet, newspaper, and others) to capture the occurrence-frequency of otherwise unobservable fundamental variables. If the econometrician is looking for recent data (after 2004), and has strong priors about the search keyword being associated with the local consumption patterns of the community of interest, then it is very likely that SVI proxies will contain additional intelligence; a diligent researcher should then combine the information contained in the multiple proxies.

7. Conclusion

We hypothesized that, ceteris paribus, the occurrence of an economic or social phenomenon increases the chances people will publish content about it. In this paper we have demonstrated that the variation in relative measures of internet and newspaper document-frequencies in reference to a phenomenon can capture cross-sectional variation of the underlying corresponding empirical occurrence-frequencies.

We began by introducing an econometric framework that specified the circumstances under which the frequency of documents containing specific keywords in relation to a given location (e.g. a country, state, or city) might be used as a proxy for the occurrence-frequency of the discussed phenomenon.

We first explored a set of variables that capture the spatial variance in economic and socio-demographic attributes that are very salient in the American context: race, ethnicity, immigration, poverty, and crime. We found document frequencies to be significantly and strongly correlated to corresponding occurrence frequencies at the state- and city-levels. These associations were shown unlikely to be spurious.

We then calculated document-frequency proxies for religious share by metropolitan area in the United States, and provided several examples on how researchers could use those to complement existing survey data. Similarly, we performed a factor analysis on a number of different document-frequency proxies for the share of gay individuals in US metropolitan areas. We use the first factor as a "gay index", and found that it correlated very strongly with the share of same-sex couples in the Census. Moreover, using the newly-created gay document-frequency index we obtained similar econometric associations to the main correlate to homosexual-couple prevalence in the existing literature.

Focusing on the measurement of corruption at the country-, state- and city-levels we also found that document-frequency proxies were highly correlated with published measures of corruption. Regression analyses utilizing the document-frequencies of country and state corruption replicated the sign, significance, and magnitude of the covariates of corruption from published papers in economics.

Our approach also allowed us to obtain and provide the first measure of corruption in American cities and study its econometric relationships with other major variables. Poverty, population size, service-sector orientation, and ethnic fragmentation were shown to predict higher levels of corruption in urban America.

Epistemologically, our results demonstrate that word frequencies in large textual databases can actually capture some of the variance in the occurrence frequency of economic and social phenomena. Econometrically, the results in the paper also illustrate the simplicity and potential usefulness of using document-frequency proxies.

As with any other proxy for economic, social, and demographic phenomena, researchers should use document-frequency proxy variables with caution and think carefully about noise-to-signal ratios, endogeneity, and potential biases. Nevertheless, when a number of restrictive conditions are met, document-frequency proxies could be used in exploratory analysis as control variables; as proxies for dependent variables; as instruments in errors-in-variables IV estimation; as latent variables or outcomes; as complements to other data sources in data imputation or factor analysis; and as a way to obtain a degree of intelligence about phenomena that are otherwise unfeasible to measure. The low cost of generating such proxies also make replication and sensitivity analyses of studies using document-frequency data straightforward for independent researchers.

References

Antweiler, Werner and Murray Z. Frank (2004). "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *The Journal of Finance*, 59, 1259–1294.

Bangerter, Adrian and Chip Heath (2004). "The Mozart Effect: Tracking the Evolution of a Scientific Legend." *British Journal of Social Psychology*, 43, 605–623.

Berger, Jonah and Gráinne Fitzsimons (2008). "Dogs on the Street, Pumas on Your Feet: How Cues in the Environment Influence Product Evaluation and Choice." *Journal of Marketing Research*, 45, 1–14.

Berger, Jonah A. and Chip Heath (2005). "Idea Habitats: How the Prevalence of Environmental Cues Influences the Success of Ideas." *Cognitive Science*, 29, 195–221.

Black, Dan, Gary Gates, Seth Sanders, and Lowell Taylor (2002). *Journal of Urban Economics*, 51, 54–76.

Bollen, Kenneth (1989). *Structural Equations with Latent Variables.* Wiley.

Boylan, Richard T. and Cheryl X. Long (2003). "Measuring Public Corruption in the American States: A Survey of State House Reporters." *State Politics and Policy Quarterly*, 3, 420–438.

Carter, R. L. and Wayne A. Fuller (1980). "Instrumental Variable Estimation of the Simple Errors-in-Variables Model." *Journal of the American Statistical Association*, 75, 687–692.

Choi, Hyunyoung and Hal Varian (2009a). "Predicting the Present with Google Trends." Google Research working paper.

Choi, Hyunyoung and Hal Varian (2009b). "Predicting Initial Claims for Unemployment Benefits." Google Research working paper.

Clemen, Robert T (1989). "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting*, 5, 559–583.

Da, Zhi, Joseph Engelberg, and Pengjie Gao (2011). "In Search of Attention." *The Journal of Finance*, 66, 1461–1499.

DellaVigna, Stefano and Matthew Gentzkow (2010). "Persuasion: Empirical Evidence." *Annual Review of Economics*, 2, 643–669.

DellaVigna, Stefano and Eliana La Ferrara (2010). "Detecting Illegal Arms Trade." *American Economic Journal: Economic Policy*, 2, 26–57.

Farrall, Kenneth N. (2009). "Suspect Until Proven Guilty, a Problematization of State Dossier Systems via Two Case Studies: US and China." Dissertation No. AAI3405375, University of Pennsylvania.

Feldman, Ronen, Jacob Goldenberg, and Oded Netzer (2010). "Mine Your Own Business: Structure Surveillance through Text Mining." Working paper, Wharton Interactive Media Initiative (http://www.whartoninteractive.com).

Finke, Roger and Christopher P. Scheitle (2005). "Accounting for the Uncounted: Computing Correctives for the 2000 RCMS Data." *Review of Religious Research*, 47, 5–22.

Gentzkow, Matthew and Jesse M. Shapiro (2010). "What Drives Media Slant? Evidence From U.S. Daily Newspapers," *Econometrica*, 78, 35–71.

Ginsberg, Jeremy, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski and Larry Brilliant (2009). "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature*, 457, 1012–1014.

Glaeser, Edward L. and Raven E. Saks (2006). "Corruption in America." *Journal of Public Economics*, 90, 1053–1072.

Godes, David and Dina Mayzlin (2004). "Using Online Conversations to Study Word-of-Mouth Communication." *Marketing Science*, 23, 545–560.

Kosmin, Barry A., Egon Mayer and Ariela Keysar (2001). *The American Religious Identification Survey (ARIS)* 2001.

LaPorta, Raphael, Florencio Lopes-de-Silanes, Andrei Shleifer and Robert Vishny (1999). "The Quality of Government." *Journal of Law, Economics, and Organization*, **15**, 222–279.

Liu, Yong (2006). "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue." *Journal of Marketing*, 70, 74–89.

Lubotsky, Darren and Martin Wittenberg (2006). "Interpretation of Regressions with Multiple Proxies." *The Review of Economics and Statistics*, 88, 549–562.

Olken, Benjamin A. (2006). "Corruption and the costs of redistribution: Micro evidence from Indonesia." *Journal of Public Economics*, 90, 853–870.

Olken, Benjamin A. (2009). "Corruption Perceptions vs. Corruption Reality." *Journal of Public Economics*, 93, 950–964.

Pelat, Camille, Clément Turbelin, Avner Bar-Hen, Antoine Flahault and Alain-Jacques Valleron (2009). "More Diseases Tracked by Using Google Trends." *Emerging Infectious Diseases*, 15, 1327–1328.

Perneger, Thomas V. (2004). "Relation Between Online 'Hit Counts' and Subsequent Citations: Prospective Study of Research Papers in the BMJ." *British Medicine Journal*, 329, 546–547.

Rech, Jörg (2007). "Discovering Trends in Software Engineering with Google Trend." ACM *SIGSOFT Software Engineering Notes*, 32, 1–2.

Rubin, Donald B. (1996). "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association*, 91, 473–489.

Saiz, Albert and Uri Simonsohn (2007). "Downloading Wisdom from Online Crowds." IZA Discussion Paper No. 3809.

Surowiecki, James (2004). *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations*. Doubleday.

Svensson, Jakob (2005). "Eight Questions about Corruption." *Journal of Economic Perspectives*, 19, 19–42.

Tetlock, Paul C. (2007). "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance*, 62, 1139–1168.

Tumarkin, Robert and Robert F. Whitelaw (2001). "News or Noise? Internet Postings and Stock Prices." *Financial Analysts Journal*, 57, 41–51.

Wolfers, Justin and Eric Zitzewitz (2004). "Prediction Markets." *Journal of Economic Perspectives*, 18, 107–126.

Wooldridge, Jeffrey M. (2001). *Econometric Analysis of Cross Section and Panel Data.* MIT Press.