DOWNLOADING WISDOM FROM ONLINE CROWDS

Albert Saiz
The Wharton School
University of Pennsylvania
Steinberg-Dietrich Hall, Suite 1466
3620 Locust Walk
Philadelphia, PA 19104-6302
saiz@wharton.upenn.edu
Tel. 215-898-28-59
Fax: 215-573-22-20

Uri Simonsohn
The Wharton School
University of Pennsylvania
Huntsman Hall 500
Philadelphia, PA 19104
uws@wharton.upenn.edu
Tel. 215-898-1630
Fax 215-898-3664

ABSTRACT

The internet contains billions of documents, is there useful information in the number of websites about different topics? We propose, based on the premise that the occurrence of a phenomenon increases the likelihood that people write about it, that the relative frequency of documents discussing a phenomenon can be used to proxy for the corresponding occurrence-frequency. After establishing the conditions under which such proxying is likely to be successful, we construct proxies for a number of demographic variables in the US and for corruption across US states and countries, obtaining average correlations with occurrence-frequencies of 0.46 and 0.61 respectively. We also replicate results from two separate published papers establishing the correlates of corruption. Finally, we construct the first index of corruption in US cities and study its correlates.

* We thank participants at departmental presentations at Wharton, Berkeley, and IZA-Bonn, and at NARSC and SJDM conferences 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, and Joe Evangelist provided excellent research assistance.

1. Introduction

When judgments made by large numbers of people are aggregated into a single estimate, be it through sophisticated prediction markets (Justin Wolfers and Eric Zitzewitz, 2004) or by simply averaging judgments of experts or even uninformed respondents (Robert T. Clemen, 1989), the result is often remarkably accurate, a regularity popularized as *The Wisdom of Crowds* in the homonymous book.[1]

In this paper we are interested in the possible "wisdom" resulting from the aggregation of a very specific kind of judgment, namely, the determination of which topic is worth writing about. Assuming that, all else constant, 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, *document-frequency*, should be correlated with the relative frequency with which the discussed phenomena have occurred, *occurrence-frequency*. Here we examine the potential for such correlation to be exploited to proxy for the occurrence-frequency of difficult to observe phenomena.

We operationalize the estimation of document-frequencies by conducting both internet (via the search engine *Exalead*®) and newspaper (via the newspaper data-bank *Newsbank*®) searches, for documents containing the keyword describing the phenomenon of interest in proximity (within 16 words) of the name of the location of interest. We focus on the internet because it integrates documents from multiple sources and because, unlike newspapers, information need not be newsworthy to be published.[2]

[1] See Surowiecki (2004) "*The Wisdom of Crowds*", Doubleday, New York.
[2] We also favored the internet because it contains billions of documents from extremely diverse sources, produced in a decentralized fashion.. At the same time, it is difficult to assess *when* internet documents were originally written so for applications where timing of occurrence is important newspapers may be preferable.

By requiring words to appear near each other, this approach dramatically reduces false-positives - instances where a document contains all searched keywords but not in relation to each other. The resulting number of documents is deflated by the total number of documents containing the keyword for the location of interest.

For instance, in August 2006 Exalead® had indexed 806 documents with the word "corruption" in proximity to "Sweden" out of the nearly 16 million pages just with the word "Sweden". For "Russia" in contrast, a keyword identifying an unambiguously more corrupt country, these figures were 11,491 and 22 millions respectively. Relative document-frequency about corruption, therefore, correctly identifies Russia as the more corrupt country. This is not an anomalous achievement. The document-frequency based corruption index we construct for 156 countries is correlated .62 with that of Transparency International (TI), the leading international indicator of corruption. Figure 1 illustrates the strong association between these two (log-standardized) variables.

*** Figure 1 ***

Of course, we do not expect 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. We introduce a conceptual framework that specifies the conditions under which this is likely to occur. The framework includes examples and eight specific data-checks that should be performed prior to utilizing document-frequency as data.

We then demonstrate the ability of document-frequency to proxy for occurrence-frequency by constructing proxies for a set of salient economic and demographic variables whose true value is readily observable (population's racial composition and

share foreign-born, and poverty and murder rates) for both states and cities in the United States. We obtain strongly significant correlations with occurrence-frequencies: an average correlation of .56 for state-level variables and of .38 for city-level ones.

Subsequently, we test the usefulness of this technique in a more natural application: proxying for variables that are not easily observable. We focus on corruption creating document-frequency based measures for corruption at the country, US state and US city level. In addition to the previously mentioned correlation of .62, at the country level, with Transparency International's index, our state level measures correlate .59 and .44 with the only two existing indices we are aware of: (Edward L. Glaeser and Raven E. Saks, 2006) and (Richard T. Boylan and Cheryl X. Long, 2003).

More notably, using our internet document-frequency-based measures of corruption as dependent variables we are able to replicate published results on the correlates of corruption at the state level (Edward L. Glaeser and Raven E. Saks, 2006) and at the country level (Jakob Svensson, 2005). Our document-frequency based measure of city level corruption is the first such fine-grained assessment of corruption in America. We utilize it to estimate regressions assessing the covariates of city level corruption. Our results indicate that covariates of corruption at the state and country level are also covariates at the city level, even after controlling for state-level corruption.

Our research should inform a growing body of research from various disciplines that takes advantage of easier access to large databases of text to obtain quantitative information. Prior research has mostly concentrated on learning about authors' beliefs (Werner Antweiler and Murray Z. Frank, 2004, Adrian Bangerter and Chip Heath, 2004, Jonah Berger and Chip Heath, 2005, Marwan Sinaceur and Chip Heath, 2005, Robert

Tumarkin and Robert F. Whitelaw, 2001, Peter D. Wiysocki, 1998), preferences (2004, 2006), sentiments (Feng Li, 2006, Paul C. Tetlock, Forthcoming) and political biases (Matthew Gentzkow and Jesse M. Shapiro, 2006). Two exceptions are, (Edward L. Glaeser and Claudia Goldin, 2004), who looked at the number of newspaper articles to *qualitatively* assess changes in corruption through time, and (Roland G. Jr. Fryer et al., 2005) who proxied for crack-cocaine availability through time.[3]

In relation to these literatures we make three notable contributions. First we demonstrate that analyses of document-frequencies need not be limited to making inferences about the authors of the written text or about the specific events described in the text, but more generally, to proxy for the relative frequency of any variable that can be expressed in frequencies. 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, and third, we validate empirically the use of document-frequency with several demonstrations.

The rest of the paper is organized as follows. In section 2 we introduce the conceptual framework, in section 3 we demonstrate the ability of document-frequency to proxy for occurrence-frequency by constructing proxies for a set salient economic and demographic variables whose true value is readily observable (population's racial composition and share foreign-born, and poverty and murder rates) for both states and cities in the United States. In section 4 we report the results from our attempts to proxy for corruption at the country, state and city level and section 5 concludes.

[3] Fryer et al. also attempted to capture cross-sectional variation in crack availability but obtained null results. We believe this was the case because their document-frequency data violate two of the data requirements put forward in this paper (datachecks #5 and #6).

2. Conceptual Framework

In this section we lay out a framework establishing the conditions under which document-frequency is likely to be a valid proxy for occurrence-frequency. We shall refer to the occurrence-frequency of phenomenon *p* in location *l* with $Y_{p,l}$ and to the corresponding document-frequency obtained by querying a document database (e.g. the internet) with a set of keywords *k*, by $\hat{Y}_{p,l,k}$ (we utilize subscripts only when needed). The following stylized equation characterizes the assumed linear relationship between *Y* and $\hat{Y}$:

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

where $\alpha_{p,k}$ is a phenomenon-keyword specific intercept, $\beta_{p,k}$ corresponds to the impact of the occurrence of phenomenon *p*, on the number of documents written about it including keywords *k* , and $\varepsilon_{p,l,k}$ to the residual.

Equation (1) is useful for identifying various datachecks that can be performed to assess conditions which make a high correlation between $\hat{Y}$ and *Y* more likely ex-ante.

2.1 $\alpha_{p,k}$: Maintaining p and k constant.

Document-frequency is unlikely to be a reliable proxy for occurrence-frequency *across* phenomena because $\alpha_{p,k}$'s do not remain constant in such comparisons. Different phenomena elicit different levels of overall interest (variation in α from *p*) and keywords relevant to different phenomena vary in how common it is for documents about such phenomena to utilize that specific keyword (variation from *k)*. As an example, suppose occurrence-frequency of cause of death by airplane and car crashes were to be

approximated by the document-frequencies for the queries for "car crash" and "plane crash". Differences in such document-frequencies could be driven not only by differences in the occurrence-frequency of such accidents, but also by the idiosyncratic appeal to write on each of the two causes of death *and* by the percentage of all documents about automobile accidents containing the keywords "car crash" vis-à-vis the percentage of airplane accidents documents containing "plane crash." This problem is greatly reduced when comparisons are made across queries that maintain both p and k constant.

Data check #1: do the queries maintain phenomenon and keywords constant?

2.2 $\beta_{p,k}$: frequencies and our basic premise.

Our basic premise, that *ceteris paribus* the occurrence of a phenomenon increases the likelihood that a written document about it will be created, is equivalent to assuming that $\beta_{p,k}>0$. Two data checks are associated with assessing the validity of this premise. The first is that the variable of interest can be expressed in terms of a relative frequency, and the second that the keyword chosen to search for documents about it is more likely to be employed following the occurrence than the non-occurrence of the phenomenon.

The keyword "education" exemplifies a violation of both requirements. First, "education" characterizes a term which cannot have a frequency interpretation (unlike, say, "high-school dropout"). 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 need not rely on subjective or intuitive judgment. It can be assessed empirically by examining the content of the documents resulting from a given query. In particular, a researcher can sample the contents of a selection of the

documents found through a particular query and assess whether keyword *k* is often utilized to demark the non-occurrence of *Y*.

Data check #2: is the variable of interest, Y*, a frequency?*

Data check #3: is the keyword k *employed predominately to discuss the occurrence (rather than non-occurrence) of the phenomenon* p*?*

2.3 $\varepsilon_{p,l,k}$: Efficiency and bias.

$\varepsilon_{p,l,k}$ captures factors that influence $\hat{Y}_{p,l,k}$ other than $Y_{p,l}$. We will discuss here four such factors with a corresponding data check for each: sampling error, occurrence variability, measurement error, and violation of the "redundancy-condition" for proxy variables. Let $\text{var}(Y_{p,l}) = \sigma_Y^2$ and $\text{var}(\varepsilon_{p,l,k}) = \sigma_{\varepsilon k}^2$. The correlation between $Y_{p,l}$ and $\hat{Y}_{p,l,k}$ is therefore $\rho\left(Y, \hat{Y}\right) = \dfrac{1}{\sqrt{1 + \dfrac{\sigma_{\varepsilon k}^2}{\beta_{p,k}\sigma_Y^2}}}$. Clearly this correlation is low when the noise-to-signal-ratio $\left(\dfrac{\sigma_{\varepsilon k}^2}{\beta_{p,k}\sigma_Y^2}\right)$ is large.

(i) Sampling error: Sampling error is reduced as sample size increases, of course, and hence, considering that document-frequency consists of the ratio of the number of documents matching the specific query over those about the location overall, the relative magnitude of $\sigma_{\varepsilon k}^2$ will tend to be smaller for topics and areas where the number of documents is "large."

Data check #4: is the average number of documents found large enough for variation to be driven by factors other than sampling error?

(ii) Occurrence variability: for a given amount of measurement error specific to the relevant keyword and geographic level, a smaller variance in the occurrence-frequency of the phenomenon will lead to a smaller correlation between occurrence and document-frequency. For example, cancer rates vary substantially more across countries than across US states (coefficients of variation of 0.7 and 0.15 respectively) and hence we would expect document-frequency about cancer to better capture occurrence-frequency across countries than states. This is exactly what we find, obtaining correlations of 0.34 ($p<.01$) and -.06 (n.s.) respectively.

<u>Data check #5:</u> is there enough variance in the variable of interest to come through the inherent noise associated with using document-frequencies?

(iii) Measurement error: Another possible cause for large $\sigma^2_{\varepsilon k}$ is that keywords often have multiple meanings, leading to false-positives; that is, to documents that do contain *k* but which are not actually about *p*. A possible way to reduce this source of $\sigma^2_{\varepsilon k}$ is to replace the keyword for a synonym with fewer other meanings (for instance using "African Americans" rather than "Blacks").

<u>Data check #6:</u> does the chosen keyword have as its primary or only meaning the occurrence of the phenomenon of interest?[4]

[4] As mentioned in the introduction (Fryer et al., 2005) computed measures of crack-cocaine availability across cities based on newspaper stories containing the word "crack", "cocaine" and the name of the city and found no cross-sectional correlation with their 4 other proxies (average correlation .02). We conducted proximity searches utilizing the same keywords and found that for most cities there simply were too few articles to make comparisons across them meaningful. For example, for the year on which most articles appeared overall, 1989, 45% of all cities had 10 or fewer documents. Variation across cities when the number of documents is so small is likely to be over-ridden by sampling error. In addition, they did not restrict their searches to those where the different keywords appeared near each other. We hand checked

(iv) Redundancy Condition: The final aspect of ε we discuss does not deal with $\sigma^2_{\varepsilon k}$, but rather with ε's 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 well-known prerequisite for such use of proxy-variables is that cov(X, $\hat{Y}$ |Y)=0 or equivalently that cov(ε,X)=0 (see e.g. Jeffrey M. Wooldrige, 2001). This condition means that, controlling for occurrence-frequency, document-frequency should be uncorrelated with the covariates of occurrence-frequency.

We consider two possible violations of this condition. The first is if *X* directly causes $\hat{Y}$ to increase, independently of *Y*. As an example, suppose gun ownership, *Y*, was to be proxied via document-frequency about guns across different cities, $\hat{Y}$, and a regression was then to be estimated with violent crime, *X*, as a dependent variable (e.g.. X=OLS($\hat{Y}$)). If the tendency to write about guns increases not only as more guns are owned, but also as more violent crimes occur, then the correlation between the two will be a biased estimate of the relationship between gun availability and crime.

This problem can be diagnosed by conducting queries that combine keywords for the occurrence of interest and its covariates. The greater the share of documents that discuss both, the greater the problem is (e.g. by assessing the share of documents with *k*="gun" that contain the keywords *"murder"* or "*crime*").

If a problem is identified, *k* can be modified to reduce or eliminate this problem by, for example, employing keywords less likely to be used to describe violent crimes,

---

the results for one city, Oakland, and found that 80% of the proximity searches were true-positives, compared to 33% of the regular searches. Their data, therefore, violated data-checks #4 and #6.

such as "gun shows" or "gun possession" instead of simply "gun" or/and by explicitly requesting the search engine to exclude keywords associating with *X* (e.g., using Boolean search to query [(*guns NEAR Oakland) NOT murder NOT crime)*].[5] Comparisons of the results obtained when such corrections are and are not implemented should provide guidance of the extent to which Corr($\hat{Y}$,X|Y)≠0 is driving the results.

<u>Data check #7</u>: does the chosen keyword also query for documents related to the covariates of the occurrence 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, if a researcher proxied for a socioeconomic variable (e.g., poverty across cities) without controlling for how cosmopolitan different cities are, then there would be concern for bias since cosmopolitanism may be associated with the tendency to write about socioeconomic issues in general (and poverty in particular). In this case, the estimated relationship between the document-frequency of poverty and X would be biased towards the relationship between cosmopolitanism and X (i.e., cov($\hat{Y}$,X) would be biased towards cov(Z, X)).

Proxying via document-frequency provides a direct fix to this problem: additional 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.

[5] NEAR corresponds to a "proximity" search; NOT excludes pages including the specified words. Some illustrative results: the query for "*gun*" on January 19th, 2007 lead to 29.1 million hits on Exalead, of which 3.3 million, or 11%, also contain the words murder, murders or murdered. In contrast, of the 155,744 documents for "*gun show*" only 4,720, or 3%, contained such words.

"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 of the variable of interest and its covariates? If so, consider proxying for the omitted variable with a new document-frequency query.

3. Demonstrations with observable occurrence-frequencies

We begin the empirical analyses with a few demonstrations of how document-frequency can be used to proxy for occurrence-frequency. Because most papers utilizing text databases have concentrated on newspapers we conducted our document-frequency estimations both on the internet, using the aforementioned search engine Exalead, and on a newspaper database. We employed *Newsbank*, a database containing more than 600 local newspapers and which also offers proximity searches. [6]

To conduct this initial demonstration we chose a set of variables that are readily available at both the state and city level and which capture salient socioeconomic dimensions in the US context. In particular we constructed proxies for share of the population that is African-American, Hispanic, and foreign born, and for both murder and poverty rates.

We obtained the first 3 variables from aggregate census counts, and crime data was obtained from the FBI's Uniform Crime Reports, via the HUD State of the Cities Database. We used the 2000 census microdata (IPUMS) to calculate poverty rates at the

[6] 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. We queried both Newsbank and Exalead utilizing specially designed PERL scripts. Importantly, we added considerable time delays between queries to avoid imposing unreasonable burdens on the queried servers.

state level as 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. [7]

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, such as Arlington and Springfield.

As mentioned earlier, we calculate 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 so we conduct all analyses on the log of these variables. For this reason we add 1 to the numerator in the odd cases when there are no documents found through the proximity search. To illustrate the need of the log transformation, Figure 2 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 that has the mean and variance corresponding to the data.

Figure 2

Furthermore, for comparability of results using variables measured in different units we standardize all document and occurrence frequency data, effectively using log-standardized frequencies throughout the rest of the paper.

[7] 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.

3.1 Data checks

Before conducting the analyses we examine whether the variables of interest pass the data-checks put forward in our framework. They of course pass checks #1 (keywords are kept constant across locations), and # 2 (occurrence of the phenomena can be expressed in relative frequency terms).

For data-check #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 50 pages made direct allusion to actual murder cases or murder rates many documents referred to murder mystery clubs, TV shows, or pop songs. In the pool of 250 documents sampled, no document made allusion to any of the keywords to signify absence or reduced occurrence of the phenomena (e.g. "no immigrants" or "lack of poverty," "less Hispanics," or similar). Data-check #3 hence passes all keywords, while data-check #6 does too with the exception of "*murder.*"

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. This data-check was reasonably met by the data, as the average number of documents found for a given keyword ranged between 410 (for "*corruption*" at the city level) and 35,957

(for "*African Americans*" at the state level). The average number of documents is substantially higher the larger the geographic level of aggregation. We therefore expect document-frequency to be a more effective proxy for occurrence-frequency for states than for large cities, and better for the latter compared to smaller cities. See Table A1 in the appendix for a full list of the number of documents found for each keyword at different levels of analysis.[8]

Data-check #5, requiring occurrence-frequency to experience substantial variation, will typically consist of a qualitative a-priori assessment. For the variables in this demonstration, however, we can directly assess the variation in occurrence-frequency since we are proxying for observable variables. In Table A2 in the Appendix we provide summary statistics for all the variables being proxied. 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, for example. We should expect, therefore, that poverty's document-frequency will be less strongly correlated with its occurrence-frequency. Finally, data-checks #7 and #8 do not apply here since we are not estimating regressions.

In sum, we expected positive correlations between occurrence and document-frequency for the social phenomena under consideration. The data-checks, however, suggest that we should encounter weaker correlations for murder (since the keyword *"murder"* recovers many unrelated documents) and for poverty, particularly at the state level (with very low occurrence variation). We also expect correlations between data and

[8] For newspapers the range is between 97 for "corruption" at the city level and 3,085 for "murder" at the state level.

document-frequency to be stronger for the larger geographic units. All of these expectations are confirmed in the data.

3.2 Results

To provide an intuitive sense of the relationship between document and occurrence-frequency for these variables, Figure 3 depicts quintile averages for each of them at the city level. The vertical axis contains the occurrence-frequency of the variable being proxied, and the figure reports 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 3

Table 1 shows the correlations between document-frequency and occurrence-frequency for the variables depicted in Figure 3 at both state and city level. All correlations are positive, with 28 out of the 30 being significant at the 5 percent level and 26 at the 1 percent level. Internet-based document-frequency is correlated on average

.439 with occurrence-frequency, almost identical to the correlation between newspaper-based document-frequency and occurrence-frequency, .440.[9]

Table 1

We interpret the positive correlations between document and occurrence-frequency as supportive of our contention that, for data that pass the multiple data-checks, greater occurrence-frequency of a specific phenomenon is associated with increased document-frequency of that same phenomenon.

Considering that the five variables we proxied for are all related to socio-economic status, however, it is possible that rather than five independent demonstrations, the above 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. This could occur if people living in cities with greater frequency of low socioeconomic status were interested in writing about socioeconomic issues for reasons other than a high local occurrence-frequency per-se. For example, one may worry that large numbers of documents are written about African Americans in Philadelphia not because of Philadelphia's large African American community, but because of Philadelphia's large Democratic Party voter base, say, which will tend to discuss *all* socioeconomic issues, *including* those pertinent to the African American community.

We address these concerns in Table 2, where we report the cross-correlations of document-frequency and occurrence-frequency of African-Americans, with the occurrence-frequency of all five demographic variables used in the above

[9] Table 1 reports Pearson correlations. Unreported Spearman correlations (based on rank and therefore not sensitive to outliers or log-standardization) were very similar. The averages across all variables are .47 for Newspapers and .45 for the Internet.

demonstration.[10] Contrary to the null hypothesis that there is a single latent variable driving all correlations in Table 1, several of the cross-correlations between African American document-frequency and the occurrence-frequency of other variables are negative, and –importantly- similar to the cross-correlations in occurrence-frequency. For example, the cross-correlation between the occurrence-frequency of Hispanics and the document-frequency of African-Americans is -.40 across cities, compared to an actual correlation between both occurrence-frequencies of -.54.

Table 2

An alternative way to address this concern, suggested in the discussion of the conceptual framework, consists of proxying for the suspected omitted variable via document-frequency. Importantly, this approach can easily be applied in situations where, unlike the present example, actual occurrence-frequencies are not observable.

If a single latent variable accounts for the multiple correlations we obtain, then partialing out the variance contained in a proxy of such a variable should substantially mute the (spurious) correlations. Because we are concerned with an overall tendency to discuss socioeconomic issues, we estimated the relative document-frequency of the keyword "socioeconomic." If the correlations from Table 1 arise because of a spurious association between the occurrence-frequency of those variables with the tendency to discuss socioeconomic issues, this variable should help us capture this trend and weaken the obtained correlations. Contrary to this prediction, we find that controlling for relative frequency of "socioeconomic" leaves the correlations between document-frequency and

[10] We focus on the African-American share because this is the variable for which document-frequency is more strongly correlated with occurrence-frequency and therefore where we have more power. Considering that we are seeking to show lack of correlation across variables this is the most conservative test we can take. We focus on states and major cities, for analogous reasons.

occurrence-frequency from Table 1 largely unchanged: .41 on average for states, .38 for cities, and .44 for large cities, compared to .52, .38 and .42 respectively.

4. Document-frequency based measures of corruption.

The results from the previous section demonstrate that document-frequency can be significantly correlated with occurrence-frequency. We now demonstrate how such correlation can be exploited to construct proxies for unobservable variables which can then be used to learn about the covariates of the variable of interest.

We chose to focus on corruption for various reasons. First, we wanted to reduce possible concerns of data snooping to a minimum. Because published papers have studied correlates of corruption both at the state and country level, by concentrating on corruption we require the exact same technique to replicate prior findings in settings with independent sources of variation. Second, the study of corruption characterizes the ideal application for the quantification of document-frequency: approximating the occurrence-frequency of a phenomenon that is otherwise very expensive to measure. Transparency International's Corruption Perceptions Index (CPI), the most commonly used international measure, averages information from 16 different surveys on experts and businessmen, some of them containing responses from more than 4,000 individuals. The high costs associated with data collection on corruption not only lead to large expenses, but also to censored, incomplete, or even nonexistent data sets. The International Crime Victim Survey from the year 2000, for example, which includes questions about bribes, was administered in only 48 countries. Quantifying document-frequency, in contrast, is virtually free and can in principle be conducted at any level of aggregation.

We present results using both internet and newspaper document-frequency, but center our discussion on the internet measures: these always work as well, if not better, than newspaper-based variables, are more widely available, and reflect documents from a much more diversified set of social agents.

4.1 Country level variation

We start by analyzing corruption at the country level. We conducted searches for "corruption" in proximity to the name of 154 countries, deflating the resulting number of documents by the number obtained searching only the countries' names. The resulting correlation between occurrence and document frequencies is strong, positive, and significant: 0.62.

An important question is the extent to which the documents we are finding are actually discussing Transparency International's CPI. On the one hand that would be good news for the validity of the technique, as it would demonstrate its ability to capture relevant information. On the other it would be bad news if document-frequency works solely because it relies on existing occurrence-frequency estimates readily available online.

To assess the extent to which our document-frequency measure relies directly on Transparency International's index, we conducted a new search adding a Boolean condition that excluded all documents containing the word "*transparency*," presumably leaving out an important share of documents that discuss corruption in relation to the CPI.[11] If document-frequency was mostly picking up variation created by the CPI, then

[11] The query was: (*(corruption NEAR <country>) NOT transparency)*).

the new index should be much less closely correlated with the CPI. The new correlation, however, is virtually identical: .60.

Replicating published results

Several papers have investigated the correlates of corruption across countries. In his review of the literature, (Jakob Svensson, 2005) estimated several regressions using various alternative measures of corruption as dependent variables, and as independent variables those hypothesized to predict corruption by various theories, making his paper an ideal benchmark.

Tables 2, 3 and 4 in his paper contain three specifications combining different correlates of corruption. We report our results for regressions using all those independent variables in our Table 3. The predictors are per-capita income in 1970, education level in 1970 (average number of years of schooling for people over 25), average (imports/GDP) between 2000-2004, and the number of days it takes to open a business in that country (we utilize the same sources cited by Svensson in his paper). The relationship of some of these variables with corruption is hard to interpret as causal, but we included them to restrict the degrees of freedom.

Because each variable has a different set of missing observations, we report both univariate regressions and a single multivariate one, with a much smaller sample size. Table 3 reports the regression results using as the dependent variable the log-standardized versions of TI's CPI index (first column), and our document-frequency based one (second column).

Table 3

In the four univariate regressions, we obtain qualitatively identical results with our document-frequency based measure and with the CPI one (both in terms of sign and statistical significance). Furthermore, point estimates (recall that these are log standardized regressions) are quite close. The only exception is 'number of days it takes to open a business,' where the document-frequency point estimate is less than half that obtained with Transparency International's CPI.

The lower panel in Table 3 shows the results combining all four predictors into a single regression. Comparing both columns the general pattern is the same: document-frequency obtains results very similar to those obtained with the CPI, with the exception of the number of days to open a business. The results from Table 3 indicate that one can learn almost *the same* about the correlates of international corruption by either conducting expensive surveys of thousands of individuals or by running a few hundred searches on the internet, which takes a matter of hours.

4.2 State level variation

We next turn our attention to corruption across states in the United States. Unlike the case of corruption across countries, no widespread index of corruption exists for different states. We are aware of two assessments of state level corruption; we used both as benchmarks for our document-frequency based index of state corruption.

The first consists of a survey conducted by {Boylan, 2003 #614}. They provided a questionnaire to 834 state house reporters, obtaining 293 responses (from 45 different states). They constructed their corruption index with the average of some of the questions in their questionnaire.

The second assessment of corruption across states is that of {Glaeser, 2006 #613}, referred to as GS for the remainder of the paper. They constructed a state-level corruption index based on the number of government officials convicted for corrupt practices through the (federal) Department of Justice (DOJ). In particular they divided the average number of DOJ corruption convictions over the 1976-2002 period by the state's average population during that same period.[12]

As GS acknowledge, 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. However, states 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 use, in addition to the index published in their paper, one which divides the same numerator of DOJ convictions by the average number of government employees in the state during the period 1976-2002.[13]

Altogether we have 5 measures of corruption at the state level: (i) the original GS index, (ii) GS computed 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, as was done in the previous sections, we log-standardize all indexes.

---

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

[13] 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.

Figure 4 depicts the relationship between measures (ii) and (iv). Corruption measured by average convictions per employee during the 1976-2002 period appears in the vertical axis and document-frequency of corruption on the internet on the x-axis. The graph shows an obvious association between both measures of corruption.

The correlations among all indexes are presented on Table 4. 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).

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

We now move to replicating previous corruption research at the state level. We estimate regressions with our various corruption measures as dependent variables and the same predictors used in GS table 4, column 1, as independent variables: income inequality (Gini in 1970), median income (in 1970), education (share of population with college degree in 1970), share of employment provided by the government, (log of) population size, share of population living in an urban area, and regional dummies. This specification nests all previous ones in GS.[14]

[14] Most of the data for the predictors used in the regressions for Table 3 were kindly provided by Raven Saks.

The results are reported in our Table 5. Column 1 uses the original GS measure, column 2 the alternative version deflated by number of public employees, column 3 an internet based document-frequency index, column 4 the survey and column 5 the newspaper based document-frequency.[15]

Comparing columns 1 and 2, we see that deflating convictions by number of public employees rather than population increases the size of most coefficients, maintaining significance mostly unchanged (consistent with the notion that deflating by number of public employees is more appropriate as it introduces less measurement error). The notable exception is the estimated impact of share of government employees, which drops to less than 5 percent of its original size and is no longer significant.

Most importantly, column 3 shows that, using our internet document-frequency measure of corruption as an independent variable, we obtain results that are largely consistent with those from columns 2 and 1. Greater income inequality, greater income levels and lower education are all associated with an increase in the internet corruption index. The point estimates are of similar magnitudes across the three columns, although education is slightly less important in the document-frequency regression. The biggest difference across columns occurs with the impact of share of employment provided by

[15] In Table 5 we exclude from the analyses the state of Georgia, because (contrary to our data-check 6) most documents allude to the Caucasian country and not to the US state of interest: for example, 34 out of the 50 first pages containing the keyword "corruption" and "Georgia" allude to the ex-Soviet Union country. We also exclude Washington State, since a majority of web pages alluding to Washington are actually in relation to the District of Columbia. 28 out of the 50 first pages containing the keyword "corruption" and "Washington State" allude to the US capital: if included in the sample Washington State would be a huge outlier, with internet frequencies two and a half standard deviations above the mean and occurrence-frequencies two standard deviations below the mean. While becoming slightly more imprecise, our main results are actually robust to the inclusion of these two states: the coefficients (standard errors in parentheses) on inequality, income, and percentage with bachelors degree become 0.68 (.28), 0.62 (.27), and -0.31 (.16)

the government. It is estimated as small positive and non-significant in column 2 and *negative* and significant in column 3.[16]

The results obtained with the survey of house state reporters, column 4, are not dissimilar qualitatively, but many of the parameter estimates are not significantly different from zero. Our internet document-frequency based proxy, hence, appears to be a *better* measure of corruption, in this case, than costly survey data.

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 5 except income inequality. In our Table 6 we report the results of 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 corruptions are remarkably similar.

In sum, we construct a measure of corruption which is both highly correlated with existing measures of corruption and we replicate the findings from a published article assessing the correlates of corruption at the state level.

4.3 City level variation

We now turn to using document-frequency based proxies as information-aggregation mechanism to produce the first assessment of corruption at the city level in

[16] To assess whether we capture variation in corruption in addition to that which is captured by the predictors employed by GS, we estimated a regression equivalent to Column 2 in Table 5 adding internet document-frequency as a predictor. We obtained a positive and significant point estimate (t-stat = 2.66).

the US. We consider all cities with more than 100,000 inhabitants in the 2000 Census, and present a ranking for cities with population over 250,000.

Document-frequency based proxies have an important component of error. Considering that previous research has shown that readers of rankings tend to overweight positional differences over differences in the underlying continuous variables that are used to construct these rankings (Devin G. Pope, 2006), we present the results from our estimation of corruption at the city level in groups of 10 cities, without disclosing the local ranking within groups. The results for the 61 cities with more than 250,000 inhabitants are presented in Table 7. The top-10 cities are consistent with our priors on corruption, including San Diego, New Orleans, Los Angeles, Philadelphia, and Chicago. Conversely, among the bottom-10 we find cities seldom used as examples of corrupt local governments.[17]

***Table 7 ***

Although the rankings map well with our pre-conceived notions of corruption in major US cities, the benchmark we are subjecting the resulting ranking is both subjective and qualitative. A better validation of the index we have created is to assess the extent to which correlates of corruption at the state and country level are correlates of corruption at the city level since most of the theories that predict a relationship with such variables would also predict a relationship at the city level.

[17] It is worth pointing out that, in line with data check #6, we dropped cities whose names are more often used to mean something other than the city in question. In particular we dropped Independence, Washington, Toledo, and Athens. "Toledo," for instance, is much more commonly used to refer to the former Peruvian president than to the city in connection to corruption. For example, in January of 2007, of the first 10 hits for "Corruption NEAR Toledo" in Exalead, nine discussed the former president and only one the Ohioan city.

In Table 8 we report the results of regressing our measure of city-level corruption on several demographic variables. The results indicate that, consistent with the country and state level analyses presented above, poorer cities have more corruption.[18] We find no evidence that larger governments (measured by the share of workers in the public sector) are more corrupt. Larger localities seem more prone to corruption, which might be because of increasing monitoring costs and collective action problems. City corruption, as perceived by internet document-frequency, seems to be a larger problem in the Northeast (consistent with the evidence at the state level), and less so in the South.

Ethnically diverse cities (as measured by the African-American share, and the share of foreign-born individuals) – column 2 - seem to experience more corruption. Blacks and immigrants seem to be more often victimized by corrupt politicians. This pattern of exploitation of minorities and the foreign-born by opportunistic corrupt officials is consistent with various previous findings at the country level (see e.g. Alberto Alesina et al., 2002, Paulo Mauro, 1995). It is also consistent with accounts of the history of corruption in the US (Menes, 2003; Glaeser and Shleifer, 2005), with political machines opportunistically exploiting ethnic divides to extract rents.

We can further test here other hypotheses that previously existing data were not appropriate to test for. For instance, we ask ourselves if declining industrial cities (as in the American rustbelt) tend to experience more corruption. In column 3 we find that, surprisingly, the share of employment in the manufacturing sector is negatively associated with corruption.

[18] Income and education are extremely correlated at the city level, so we cannot meaningfully include both variables in the regression.

We interpret the results from columns 1-3 as supportive of the notion that document-frequency captures meaningful variation in corruption across American cities. In columns 3-6 we conduct various additional analyses that examine alternative interpretations to these results.

First, one may be concerned that the findings we get are spurious, caused by a tendency of people to write more about corruption in cities (or with regards to cities) that have certain characteristics that happen to be associated with corruption at the state and country level. For example, we find that larger cities (in population) tend to be measured as more corrupt. This correlation could possibly be the result of people being more inclined to writing about social issues in general with regards to larger cities (a spurious relationship for the econometrician studying corruption). As suggested in data-check #8 we assess the potential importance of this concern by estimating the document-frequency of a variable that may proxy for the omitted variable in question. We use again the document-frequency of the keyword "socioeconomic" and add this variable as a control in column 4. Although it proves a significant predictor of the document-frequency of corruption, 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.

Another concern about our results has to do with the extent to which they truly capture city rather than state level variation. Although we estimated corruption at the city level by computing document-frequency for searches with city names, it is possible that ultimately we are actually relying on state-level variation in corruption (e.g. by capturing corruption cases taken place in state capitals).

To address this concern we added our document-frequency based measure of corruption at state-level measure in column 5. The results show that, (a) state-level document-frequency of corruption is not a significant predictor of city-level corruption (controlling for city observables), and (b) that more importantly, its introduction in the model does not greatly influence the point estimates of the other independent variables.[19] This strongly suggests we are capturing variation in corruption above and beyond state-level corruption.

Finally, in column 6 we estimate the regression weighting each observation by the population of the city. As suggested by the initial demographic correlations in Table 1, we would expect the internet proxy to be less noisy for the larger cities. We obtain very similar results.

In sum, our document-frequency measure of corruption at the city level both generates a ranking of cities that is consistent with our preconceptions, and it replicates findings of the covariates of corruption at greater levels of aggregation.

6. Conclusions

The internet consists of billions of documents from extremely diverse sources, produced in a decentralized fashion by millions of individuals around the globe. Can we aggregate such massive amounts of information in a meaningful way?

We hypothesized that, *ceteris paribus*, the occurrence of a social phenomenon increases the chances people will publish content about it. In this paper we have demonstrated that using relative measures of internet and newspaper textual frequency in

[19] Results are almost identical if, as in Table 5, we exclude Georgia from the regression (3 cities).

reference to a phenomenon can capture cross-sectional variation of the underlying empirical occurrence-frequencies.

We begun by introducing a 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 social phenomenon. We then validated the technique showing strong, positive, statistically significant correlations between internet document-frequency and empirical data on several major demographic variables.

Focusing on the measurement of corruption at the country, state and city level we also found that document-frequency based measures of corruption were highly correlated with published measures of corruption. Regression analyses utilizing the document-frequency based measures of corruption for countries and states replicated the sign, significance and magnitude of the covariates of corruption from published papers. Strikingly, using data that we obtained from the internet in a matter of hours, we obtain results similar to those arising from data based on expensive surveys or administrative collection processes, illustrating the simplicity and potential power of this approach.

Our results demonstrate that when the requirements put forward in the framework are met, document-frequency's correlation with occurrence-frequency allows researchers to construct proxies for otherwise unobservable variables. This opens the door to studying previously not-measured variables, as we do here with city-level corruption. More importantly, it allows for the creation of proxy variables for suspected omitted variables in settings where the dependent variable *is* observable, an exciting possibility

considering that virtually all non-experimental field studies suffer from potential bias due to omitted variables.

References

Alesina, Alberto; Baquir, Reza and Easterly, William. "Redistributive Public Employment." *Journal of Urban Economics*, 2002, *48*, pp. 219-41.

Antweiler, Werner and Frank, Murray Z. "Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards." *The Journal of Finance*, 2004, *LIX*(3), pp. 1259-94.

Bangerter, Adrian and Heath, Chip. "The Mozart Effect: Tracking the Evolution of a Scientific Legend." *British Journal of Social Psychology*, 2004, *43*, pp. 602-23.

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

Boylan, Richard T. and Long, Cheryl X. "A Survey of State House Reporters' Perception of Public Corruption." *State Politics and Policy Quarterly*, 2003, *3*(4), pp. 420-38.

Clemen, Robert T. "Combining Forecasts: A Review and Annotated Bibliography." *International Journal of Forecasting*, 1989, *5*, pp. 559-83.

Fryer, Roland G. Jr.; Heaton, Paul S.; Levitt, Steven D. and Murphy, Kevin M. "Measuring Crack Cocaine and Its Impact," *NBER Working Paper.* 2005.

Gentzkow, Matthew and Shapiro, Jesse M. "What Drives Media Slant? Evidence from U.S. Daily Newspapers," *NBER Working Paper.* 2006.

Glaeser, Edward L. and Goldin, Claudia. "Corruption and Reform: An Introduction," *NBER Working Paper.* 2004.

Glaeser, Edward L. and Saks, Raven E. "Corruption in America." *Journal of Public Economics*, 2006, *90*(6-7), pp. 1053-72.

Li, Feng. "Do Stock Market Investors Understand the Risk Sentiment of Corporate Annual Reports?," *Available at SSRN: http://ssrn.com/abstract=898181* 2006.

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

Mauro, Paulo. "Corruption and Growth." *Quarterly Journal of Economics*, 1995, *110*(August), pp. 681-712.

Pope, Devin G. "Reacting to Rankings: Evidence From "America's Best Hospitals and Colleges"," *Job Market Paper, University of California-Berkeley, Economics Department.* 2006.

Sinaceur, Marwan and Heath, Chip. "Emotional and Deliberative Reactions to a Public Crisis: Mad Cow Disease in France." *Psychological Science*, 2005, *16*, pp. 247-54.

Svensson, Jakob. "Eight Questions About Corruption." *Journal of Economic Perspectives*, 2005, *19*(3), pp. 19-42.

Tetlock, Paul C. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *The Journal of Finance*, Forthcoming.

Tumarkin, Robert and Whitelaw, Robert F. "News or Noise? Internet Message Board Activity and Stock Prices." *Financial Analysts Journal*, 2001, *57*(3), pp. 41-51.

Wiysocki, Peter D. "Cheap Talk on the Web: The Determinants of Postings on Stock Message Boards," *University of Michigan Business School Working Paper.* 1998.

Wolfers, Justin and Zitzewitz, Eric. "Prediction Markets." *Journal of Economic Perspectives*, 2004, *18*(2), pp. 107-26.

Wooldrige, Jeffrey M. *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press, 2001.

Figure 1: Corruption in the World

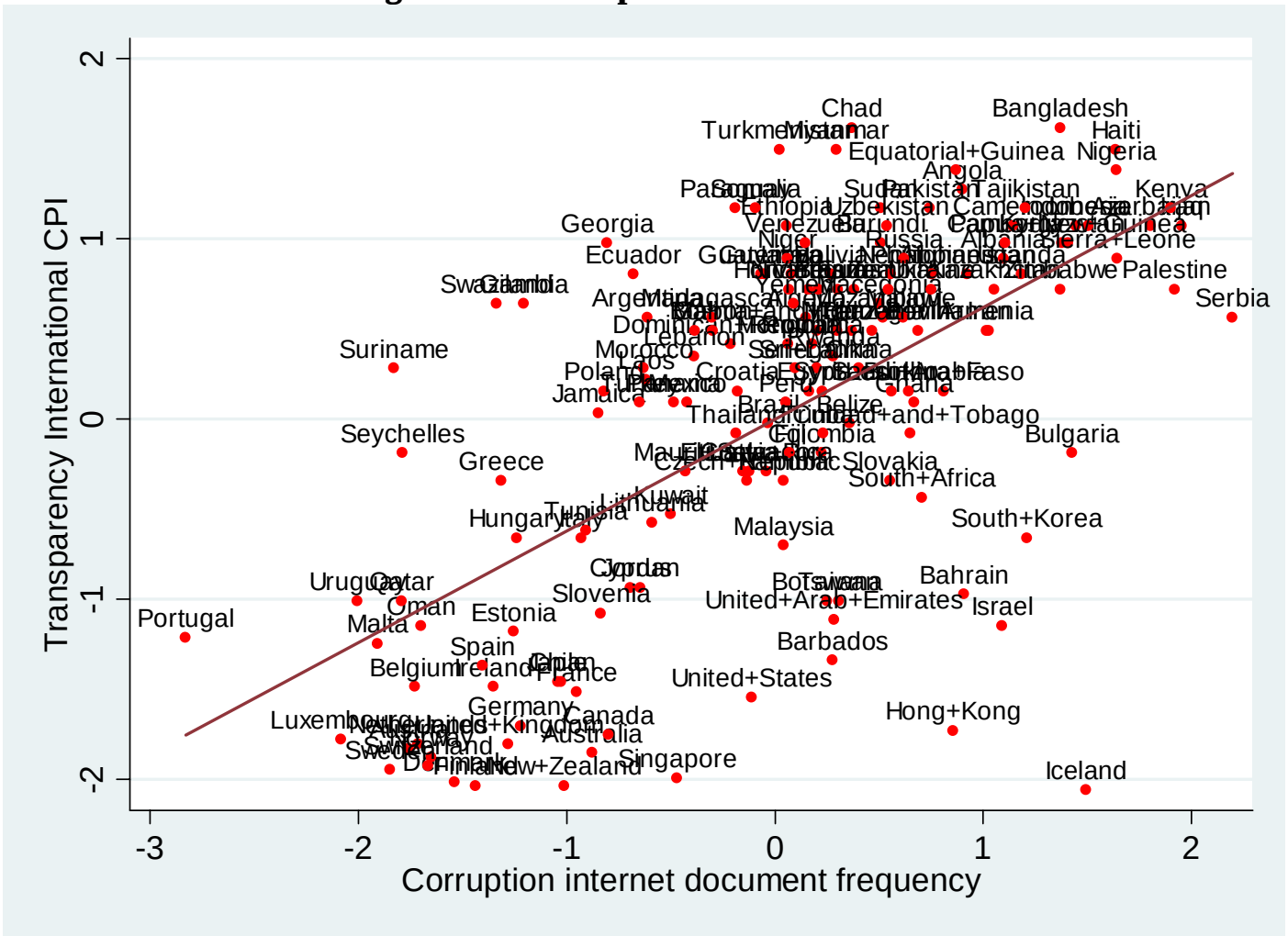


Figure 2: Log-standardizing the Data Sources – African Americans in US Cities

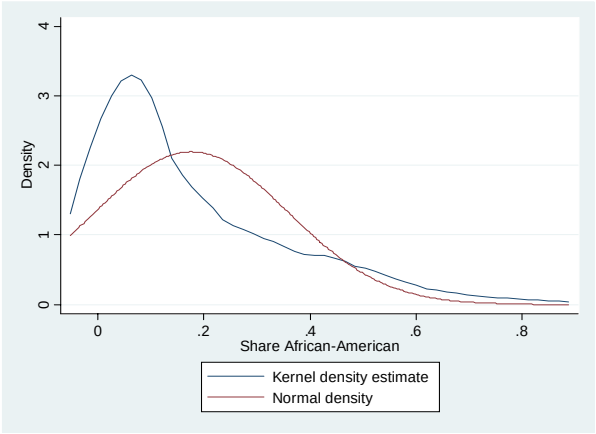
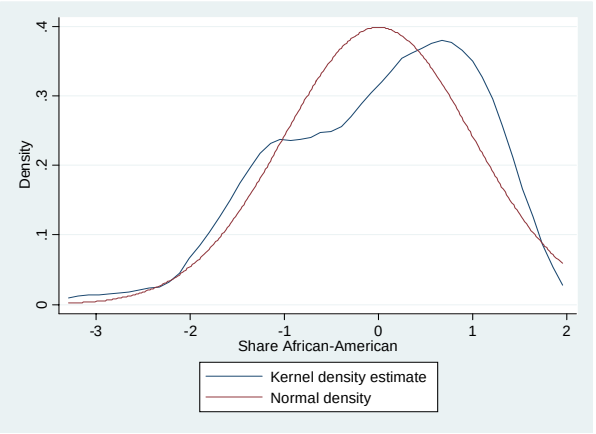
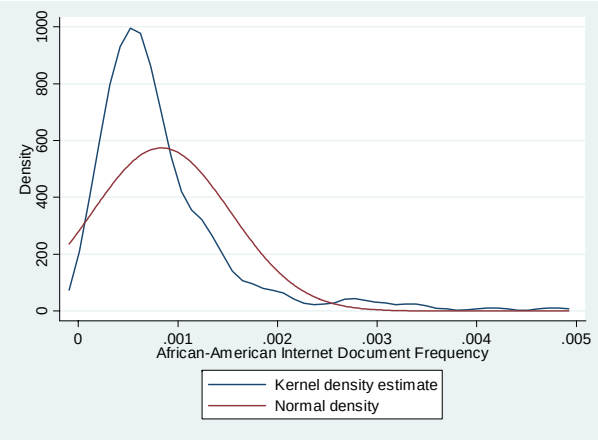
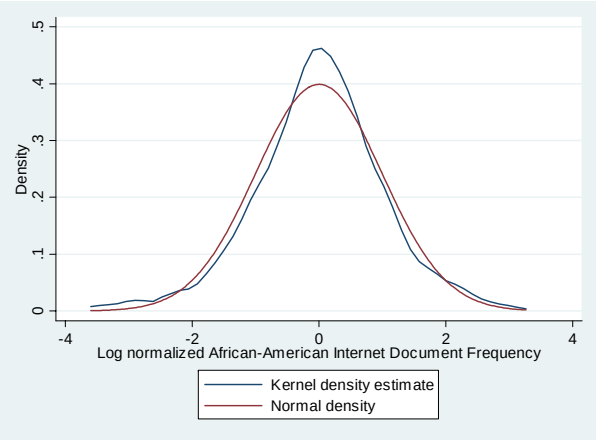
	Raw Data	Log Standardized Data
Empirical Occurrence Frequency		
Internet Document Frequency		

Figure 3: Average Data by Frequency Quintiles (cities)

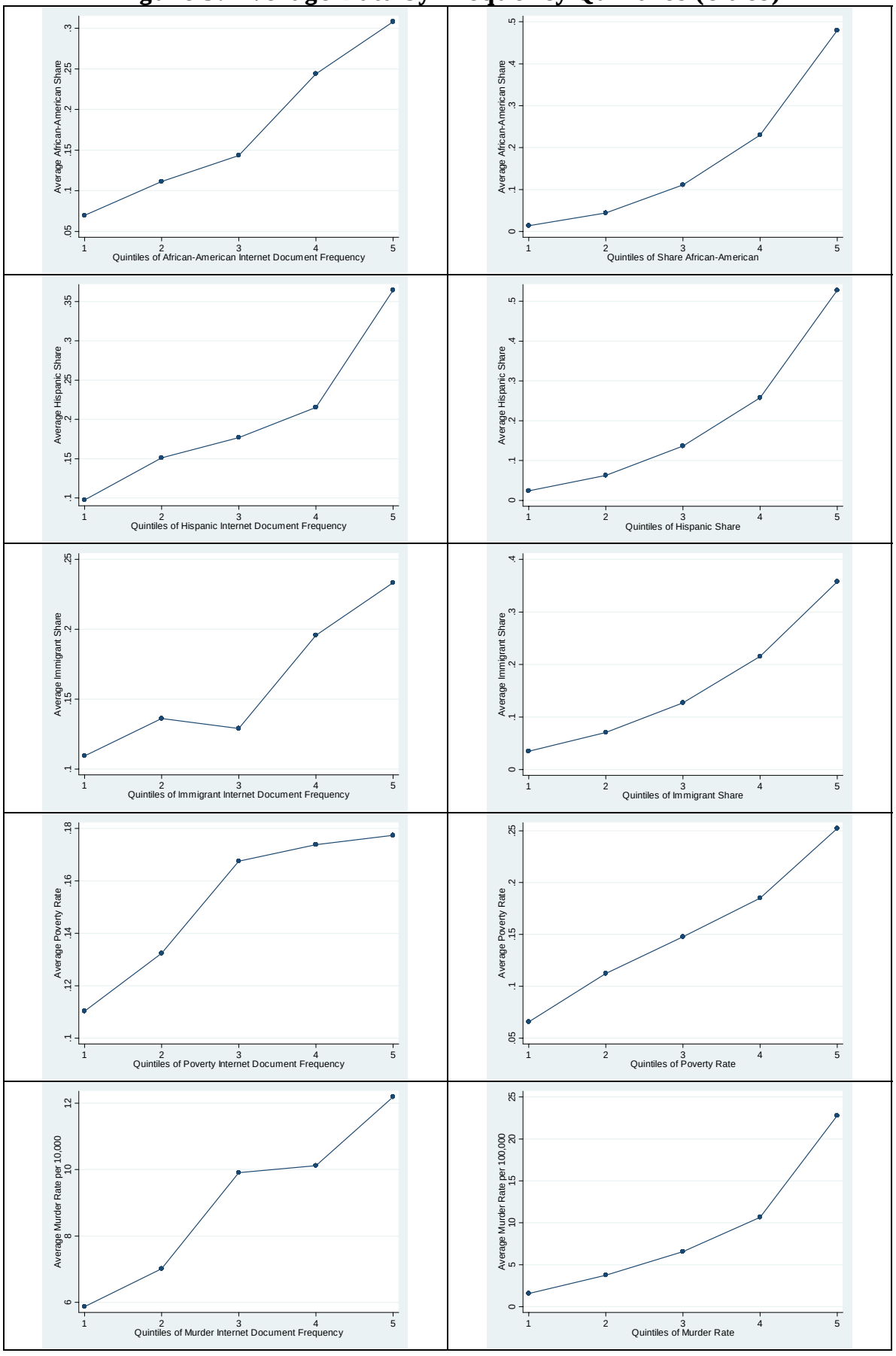


Figure 4: Corruption in the USA

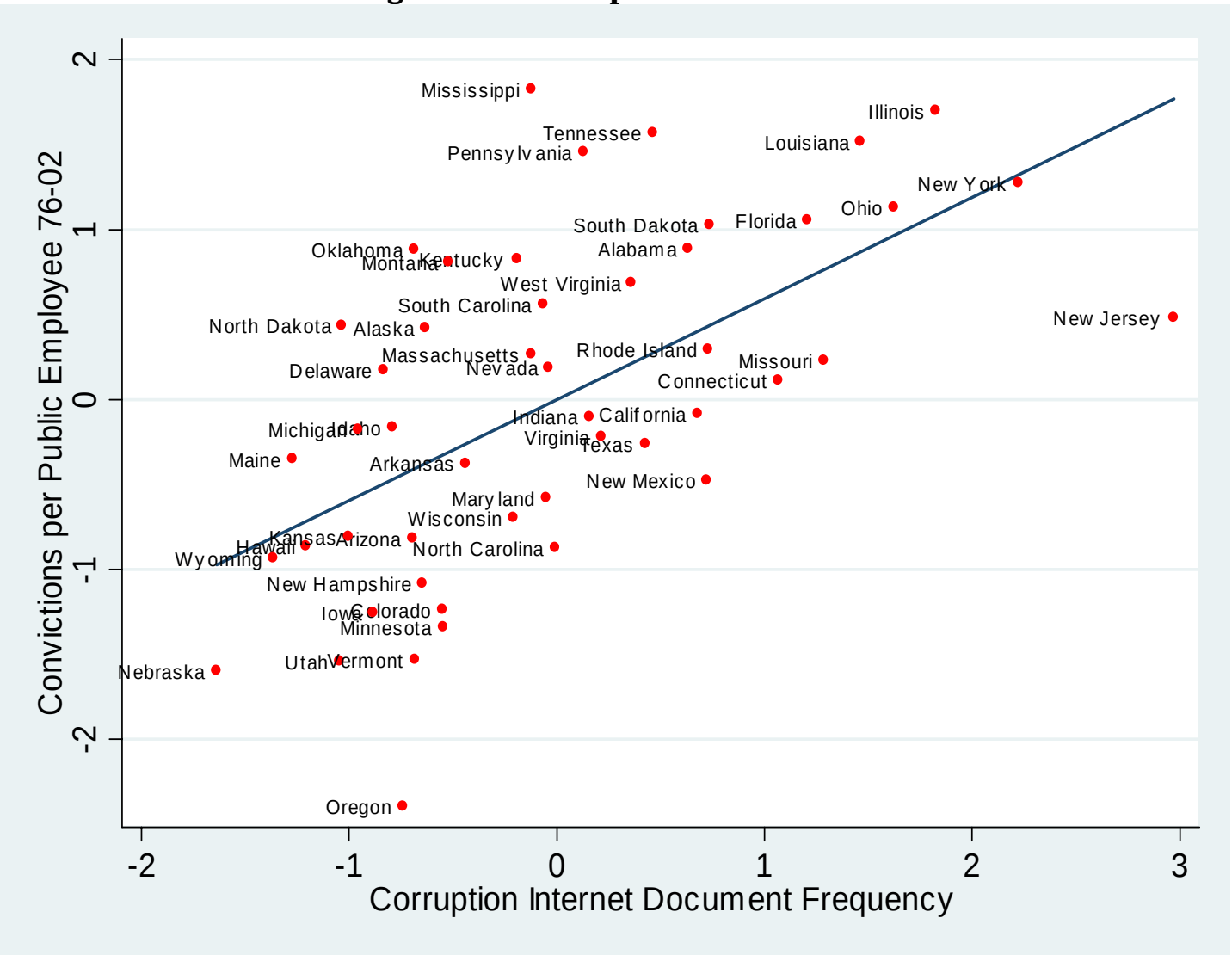


TABLE 1

Correlations Between Occurrence and Document Frequencies

	Internet			***Local Newspapers***		
	US States	Cities		US States	Cities	
		pop>100k	pop>250k		pop>100k	pop>250k
African-Americans[a]	0.70	0.43	0.67	0.82	0.50	0.61
Hispanics[a]	0.50	0.43	0.43	0.74	0.48	0.56
Immigrants[a]	0.51	0.37	0.44	0.69	0.40	0.46
Poverty rate[b]	0.41	0.34	0.31	0.37	0.26	0.20[†]
Murder rate[c]	0.48	0.29	0.26	0.36	0.13	0.02[†]
Average	***.519***	***.375***	***.422***	***.596***	***.354***	***.371***
N	50	227	62	50	227	62

Notes:

Document-frequencies correspond to the number of documents found with a given keyword appearing within 16 words of the keyword for the location, divided by all documents found with the keyword for the location.

All correlations are significant at the 5% level unless stated otherwise.

[†] Not significant at 5%

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

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

[c] Murder rate corresponds to the average murder rate per 10,000 in 2000-2005 according to the FBI's Uniform Crime

TABLE 2

Cross Correlations of Frequencies of African-Americans with Other Frequencies

	(1)	(2)	(3)
	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[a]	0.21	0.47	0.36
Murder rate[b]	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[a]	0.48	0.34	0.38
Murder rate[b]	0.79	0.59	0.51

Notes:

† Not significant at 5%

[a] Poverty rate corresponds to percentage of households below half the median of state income measures income measures.

[b] Murder rate is average for 2000-2005

TABLE 3

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+)[a]	-0.679***	-0.527***	-0.446***	*96*
	(0.100)	*(0.106)*	*(0.093)*	
	Only GPD per capita in 1970 as a predictor			
Log (real GDP in 1970)[b]	-0.761***	-0.618***	-0.483***	*105*
	(0.060)	*(0.074)*	*(0.089)*	
	Only Imports/GDP as a predictor			
Log (average[imports/GDP] 1980-2004)[c]	-0.081	-.0903	-0.196**	*145*
	(0.083)	*(0.072)*	*(0.079)*	
	Only days required to open a new business as a predictor			
Log (days to open new business)[d]	.601***	.265***	.341***	*84*
	(0.055)	*(0.070)*	*(0.110)*	
	All four predictors			
Log(Average education [in years] 1970, adults 25+)	0.072	0.180	0.167	*54*
	(0.096)	(0.167)	(0.137)	
Log (real GDP in 1970)	-0.747***	-0.889***	-0.611***	
	(0.139)	(0.167)	(0.159)	
Log (average[imports/GDP] 2000-2004)	-0.181**	-0.205**	-0.316**	
	(0.075)	(0.078)	(0.120)	
Log (days to open new business)	0.271***	-0.050	0.215	
	(0.075)	(0.091)	(0.130)	

Robust standard errors below parameter estimates

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

Data Sources:

[a] Center for International Development at Harvard (http://www.cid.harvard.edu/ciddata/ciddata.html)

[b] Penn Tables (6.1)

[c] World Development Indicators, World Bank, 2004

[d] From Djankov et al (2002). Available from http://post.economics.harvard.edu/faculty/shleifer/Data/registration_new.dta

TABLE 4

Correlations of Statel 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 & Sak, 2006)

[b] Division of total number of convictions by population, original Glaeser & Sak (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 5

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

	(1)	(2)	(3)	(4)	(5)
Dependent Variable:	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 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 Gov.Employment	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 Urban	0.153	0.255	0.334***	0.660***	0.263*
	(0.188)	*(0.184)*	*(0.118)*	*(0.145)*	*(0.154)*
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-squared	0.54	0.52	0.56	0.5	0.49

Notes:

Robust standard errors in parentheses

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

All variables have been logstandardized

Regressions exclude Washington state and Georgia (see text)

[a] Convictions correspond to Federal Department of Justice convictions on corruption charges of state officials (as used in Glaeser & Sak, 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 6

Additional Predictors of Corruption, Significant at the 5% level in (Glaeser and Sak, 2006)

	Dependent Variable		
Predictors	Convictions per Inhabitant 76-02	Convictions per Public Employee 76-02	Document Frequency of Corruption
Racial Dissimilarity	0.402**	0.343	0.288
	(0.169)	*(0.209)*	*(0.206)*
Share Black	0.381***	0.371**	0.317*
	(0.132)	*(0.155)*	*(0.183)*
Local Share of Gov. Employment	1.112	2.012	1.793
	(-1.368)	*(1.483)*	*(2.102)*
Integrity ranking, 2002	-0.025***	-0.026***	-0.015*
	(0.007)	*(0.008)*	*(0.008)*

Notes:

Shows coefficients for independent regressions. In all regressions (as in Glaeser-Saks, 2006) we also control for 1970 income, education, population, share government employment, urban share, and regional dummies.
Robust standard errors in parentheses.
* significant at 10%; ** significant at 5%; *** significant at 1%
All variables have been logstandardized.
Regressions exclude Washington state and Georgia (see text).

TABLE 7

Document-frequency Based Corruption Measure at the City Level (pop>250,000). Sorted in Groups of 10 from Most to Least Corrupt, Alphabetical Within Group

Group	City Name	Group	City Name
1	Chicago	4	Austin
1	Las Vegas	4	Corpus Christi
1	Los Angeles	4	Fort Worth
1	Miami	4	Honolulu
1	New Orleans	4	Houston
1	New York	4	Long Beach
1	Philadelphia	4	Milwaukee
1	San Diego	4	Sacramento
1	San Jose	4	Santa Ana
1	St. Louis	4	St. Paul
2	Atlanta	5	Anchorage
2	Boston	5	Buffalo
2	Cleveland	5	Cincinnati
2	Detroit	5	Minneapolis
2	El Paso	5	Pittsburgh
2	Newark	5	Portland
2	Oklahoma City	5	Raleigh
2	Phoenix	5	Tampa
2	Riverside	5	Tucson
2	San Francisco	5	Wichita
3	Baltimore	6	Albuquerque
3	Dallas	6	Anaheim
3	Denver	6	Charlotte
3	Fresno	6	Colorado Springs
3	Lexington-Fayette	6	Indianapolis
3	Memphis	6	Jacksonville
3	Oakland	6	Louisville
3	San Antonio	6	Mesa
3	Seattle	6	Nashville-Davidson
3	Virginia Beach	6	Omaha
		6	Tulsa

TABLE 8

OLS Identifying Correlates of Internet Corruption Document-Frequency Based at the City Level

	Benchmark (1)	Adds Racial Heterogeneity (2)	Adds Manufacturing (3)	Adds "Socioeconomic" document-frequency[a] (4)	Adds state-level document-frequency[b] (5)	Weighted by population (6)	Newspaper Document Frequency (7)
Log of Median Household Income	-0.193***	-0.167**	-0.156**	-0.156**	-0.158**	-0.120**	-0.227***
	(0.073)	*(0.077)*	*(0.079)*	*(0.070)*	*(0.070)*	*(0.061)*	*(0.078)*
Share Workers in Public Administration	0.007	0.021	-0.016	-0.042	-0.04	-0.056	0.054
	(0.051)	*(0.053)*	*(0.055)*	*(0.053)*	*(0.054)*	*(0.051)*	*(0.054)*
Log of Population	0.266***	0.232***	0.205***	0.183***	0.185***	0.196***	0.124**
	(0.050)	*(0.050)*	*(0.053)*	*(0.050)*	*(0.049)*	*(0.036)*	*(0.055)*
South	-0.403**	-0.401**	-0.441**	-0.332**	-0.366**	-0.460***	-0.129
	(0.161)	*(0.170)*	*(0.173)*	*(0.164)*	*(0.175)*	*(0.150)*	*(0.167)*
Northeast	0.391*	0.345	0.368	0.295	0.245	-0.15	0.537*
	(0.237)	*(0.231)*	*(0.228)*	*(0.211)*	*(0.222)*	*(0.178)*	*(0.287)*
Midwest	-0.142	-0.043	0.07	0.093	0.072	-0.018	-0.124
	(0.187)	*(0.206)*	*(0.217)*	*(0.215)*	*(0.212)*	*(0.179)*	*(0.181)*
Share African-American	--	0.135*	0.154**	0.11	0.095	0.119**	0.097
		(0.073)	*(0.075)*	*(0.068)*	*(0.074)*	*(0.059)*	*(0.073)*
Share Foreign Born	--	0.163*	0.197**	0.174**	0.164**	0.171***	0.201***
		(0.083)	*(0.085)*	*(0.074)*	*(0.076)*	*(0.056)*	*(0.076)*
Share Workers in Manufacturing	--	--	-0.142*	-0.149*	-0.146*	-0.148*	-0.057
			(0.082)	*(0.079)*	*(0.079)*	*(0.081)*	*(0.082)*
"Socioeconomic" Document Frequency	--	--	--	0.307***	0.312***	0.368***	--
				(0.059)	*(0.061)*	*(0.061)*	
State-Level "Corruption" Document Frequency	--	--	--	--	0.054	0.161**	--
					(0.080)	*(0.065)*	
Observations	224	224	224	224	224	224	224
R-squared	0.18	0.2	0.21	0.3	0.3	0.65	0.19

Robust standard errors in parentheses

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

[a] Controlling for overall tendencies to discuss social issues with respect to the city we control for the document-frequency of "socioeconomic" in proximity to the name of the city

[b] To control for state level corruption we add as a control the document-frequency of "corruption" for the state in which the city is located

Appendix TABLE 1

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

Appendix TABLE 2

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