

Downloading Wisdom from Online Crowds

Albert Saiz

University of Pennsylvania - The Wharton School; Institute for the Study of Labor (IZA)

Uri Simonsohn

University of California, San Diego

May 2007

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.

Prediction markets, where people bet on everything from the likelihood that a movie will be a hit to the chance that a politician will become president to whether the stock market will go up or down, are in vogue. Research papers have been written on their accuracy, and the media likes to write about how these predictions often beat the purported experts.

But they are not perfect. Markets require babysitters. Someone has to set them up and ensure that the traders' money -- in cases where people are, for example, buying shares of stock -- exchanges hands in an orderly way. Wharton professors [Albert Saiz](#) and [Uri Simonsohn](#) have found a cheaper way to deliver some of the same benefits. It's called an Internet search.

Specifically, Saiz, in the real estate department, and Simonsohn, in operations and information management, argue in a new [research paper](#) that the likelihood that a topic is discussed online, in relation to a given location, correlates with its relative prevalence in the real world. "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," they write in a paper titled, "Downloading Wisdom from Online Crowds." For example, they wanted to discern which countries, U.S. states and big U.S. cities people perceived as the most corrupt. So they plugged the appropriate terms into a search engine called Exalead. By assessing how many documents contained the word "corruption" within the same paragraph as the location's name, they came up with corresponding corruption rankings.

The results will surprise no one. Among the countries perceived as most corrupt are Nigeria, Serbia and Haiti. Among the states were New Jersey, New York and Illinois. The cities included Chicago and New Orleans.

Simonsohn points out that there is no way of knowing for sure whether these places are corrupt. What their searches told them was that many online documents referred to the locations and corruption in close proximity.

But people do talk and worry about things in reference to where they are a problem. People fret about hungry alligators in Florida, not Maine. And in fact, alligator attacks are far more prevalent in Florida, where all but one of the country's fatal gator grabbings have been reported since 1948.

As the two scholars put it in their paper: "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 should be correlated with the relative frequency with which the discussed phenomena have occurred."

To create as large a sample as possible, Saiz and Simonsohn didn't restrict their searches to media reports. They searched a wide array of documents and found that no one particular kind dominated their results. "We picked up a lot of news, but we also picked up a lot of government documents," Simonsohn says. In addition, "when we started to search social indicators, like the number of African Americans or Hispanics in a city, we found a lot of documents produced by cultural organizations and museums."

And that's why Simonsohn believes that they are documenting more than just "buzz," that is, the rumors and chatter that often fuel discussion in blogs and chat rooms. "Buzz tends to be short-lived, and the stuff we saw wasn't short-lived," he says. "I thought we would pick up a lot of blogs, but we picked up a lot less than we expected."

Measuring Social Trends

In fact, their paper showed clear and stable patterns with respect to a number of major socio-demographic city and state characteristics. Concretely, Saiz and Simonsohn looked at the number of Internet documents containing the keywords "African American," "Hispanic," "immigrant," "poverty" and "murder" in textual proximity to the names of all major cities and states. Remarkably, they found strong positive correlations between the actual frequencies of a phenomenon in a location; say the percentage of Hispanics in a city, and the frequency of documents on the Internet discussing the phenomenon with reference to each location. The correlations were pervasive at the U.S. city and state levels.

Saiz and Simonsohn thus showed that Internet document frequencies could be used to approximate the relative city and state rankings of major social phenomena that are currently well-measured. But they were intrigued about the possibility of using the technique to measure a variable that is not so readily measurable: corruption.

Simonsohn notes that he and Saiz view their results as demonstrating a useful technique for social scientists and people interested in measuring social trends in cities rather than making a definitive statement about which places have a high number of policemen and politicians seeking bribes. So, one shouldn't sell his or her house in Los Angeles just because it appeared at the top of Saiz' and Simonsohn's corruption list. (But it might be smart to give a donation to the Police Benevolent Association at Christmas.)

The two scholars checked their country corruption Internet results against a prominent annual ranking performed by Transparency International, a Berlin-based nonprofit that conducts surveys of business people and country experts to create its ranking. Transparency International, too, ranks perceptions about countries' corruption, not corruption itself.

Saiz and Simonsohn found that their ranking largely agreed with Transparency International's, with one prominent exception -- Iceland. They ranked Iceland among the countries perceived as most corrupt while Transparency International ranked it as the second least corrupt, behind Finland. "We made our biggest error with Iceland," Simonsohn concedes. "We think it's because it's been one of the least corrupt countries for a number years. People discuss it a lot in terms of corruption but as an example of the best, not the worst."

No group comparable to Transparency International ranks U.S. states and cities, so the two scholars had to find other ways to backstop their method there. For states, they compared their findings with the average number of criminal convictions per public employee. Here again, their list checked out. They ranked Nebraska as the state perceived as least corrupt and found that it also had a very low level of public-employee convictions. New Jersey, in contrast, was perceived as one of the more corrupt on their list and had a relatively high level of convictions. The setting for *The Sopranos* television show, in other words, was no accident.

For the city ranking, Saiz and Simonsohn had to go to even greater lengths to validate their findings, as no single source offered a definitive means of comparison. But that prompted them to dig into demographic and socioeconomic data, where they found correlations that Simonsohn argues are more instructive than a list of which cities are the most corrupt.

"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, 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 researchers write. "The top 10 cities are consistent with our priors on corruption, including San Diego, New Orleans, Los Angeles, Philadelphia and Chicago."

As they explored the data, they found that poorer cities were, by their measure, more corrupt, as were cities in the Northeast. Large cities seemed more corrupt, but cities with larger governments, measured by share of workers in the public sector, didn't.

Launching the New PlayStation

"Ethnically diverse cities (as measured by the African-American share and the share of foreign-born individuals) seem to experience more corruption," they add. "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. It is also consistent with accounts of the history of corruption in the U.S., with political machines opportunistically exploiting ethnic divides to extract rents."

The links between socioeconomic indicators and corruption point to the ways in which people interested in measuring social trends might use Saiz' and Simonsohn's technique. For instance, one could assess the frequency with which the word "pollution" appears next to the name of each Chinese region in that country's websites. It is not clear if current official data sources are a reliable source of information on the issue. Saiz' and Simonsohn's technique could thus yield a measure of the level of concerns about pollution in each area of that country

Their research identifies a recurrent data pattern in situations where massive quantities of textual information are produced by different people in a decentralized fashion. Social scientists are likely to use Internet document frequencies as a proxy for local social trends that are otherwise difficult to measure.

However, other business-oriented applications exist. Simonsohn argues that some carefully worded Internet searches might allow marketers to save money by initially helping them to focus their efforts. A company like Sony might try to assess online buzz when launching a new version of its PlayStation video game console. "When Sony launches a new PlayStation, that's a huge logistics problem," he notes. "Which cities do you ship the most to? Suppose you measured buzz in different cities before you launched, and then you could adjust the number of shipments so the cities with

the most interest got more PlayStations." In fact, firms like Nielsen Buzzmetrics are already using the consumer generated content on the Internet to aid in marketing efforts.

Political consultants -- who, after all, are really just marketers who sell people, not products -- could use the technique, too. They could test the frequency with which people write about their candidates and opponents alongside a variety of desirable and undesirable adjectives. Then they could follow up with surveys or focus groups.

One side benefit of the study was the chance to assess the various online search engines, all of which Saiz and Simonsohn tried to use. Google kicked them off. "It won't allow a single automated search," Simonsohn says. They ended up preferring Exalead, a French search engine, available in English, because they consider it and Ask.com to be the most reliable. "We found Yahoo to be the least reliable," he adds. "If you search for something today and again next week, there can be several million pages of difference in the number of results on Yahoo. I don't think several million new documents were created in a week."

Link to paper: [Downloading Wisdom from Online Crowds](#)