Assessing Public Opinion Trends based on User Search Queries: Validity, Reliability, and Practicality

Jonathan J. H. Zhu[1], Xiaohua Wang[2], Jie Qin[1] and Lingfei Wu[1]

[1] Web Mining Lab, Dept. of Media & Communication, City University of Hong Kong
[2] School of Communication, Shenzhen University, China

Abstract

User search queries (i.e., words entered search engines) were initially regarded as a by-product by the search engine industry to help improve indexing services, but have quickly been recognized as a gold mine of data on user concerns, interests, tastes, etc. Major search engines, such as Google and Baidu, have even published "public opinion trends" based on the queries they receive on a daily basis. The paper aims to assess the validity (i.e., how representative of the general public) and the reliability (i.e., how much random noise is involved) of the query-based trends, by comparing public opinion on selected issues measured by query trends and conventional surveys.

Keywords:

Search Engines, Opinion Polls, Telephone Interviews,

Introduction

Search engines are among the oldest services of the Web 1.0 era, on which the content (i.e., a list of webpages assumed to be relevant to user requests). As such, search engines per se are not a social media of the Web 2.0 era. However, when some search engines make available the aggregated frequency of search queries (i.e., the keywords users enter for their search) through a searchable interface (e.g., Google Trends, Baidu Index, etc.), the new search service (also known as "query of search queries") becomes a social media on which the content (i.e., trends of search) is created by users.

The trends of search queries have quickly been used as indicators of public opinion in the real world. Epidemiologists are the first to exploit this new source of data by predicting the outbreak of influenza and other pandemic diseases based on the rise of relevant queries in Google Trends (Carneiro, & Mylonakis, 2009; Choi & Varian, 2009; Ginsberg et al., 2009; Pelat, et al., 2009; Polgreen, Chen, Pennock & Nelson, 2008; Wilson & Brownstein, 2009). Economists find that the frequency of job related queries in Google and search engines are closely correlated with the rise and fall of unemployment rate (Askitas & Zimmermann, 2009; D'Amuri, 2009; D'Amuri & Marcucci, 2009;) and consumer spending (Kholodilin, Podstawski & Siliverstovs, 2010; Schmidt & Vosen, 2009; Suhoy, 2009). Political scientists use query data to predict the amount of political contributions raised by candidates for U. S. senate (Ellis, Ripberger & Swearingen, 2011), the proportion of votes for issue ballots in the 2008 U. S. presidential election (Reilly, Richey & Taylor, 2012). Most directly related to public opinion research is the examination of agenda-setting effects by correlating the number of news reports by *New York Times* on healthcare, global warming, and terrorism and search queries of the same issues (Ripberger, 2011).

Research Question

Despite the impressive success in using search query data to predict various real-world indicators of public health, the economy, and elections, one critical question remains unknown – the extent to which search queries represent public opinion of the general population. The question arises primarily from a simple fact that not everyone uses search engines on the Internet. Even in the countries with the highest penetration rate of the Internet such as the U. S., about one quarter of the population do not go online. It follows that the trends revealed in search queries reflect the concerns of the younger, more educated, and more active segments of the population, which are likely to be different from those of non-users with the amount of the difference depending on the proportion of non-users in the population.

In addition, there are other reasons, some being substantive others more technical, for us to

The annual conference of the World Association for Public Opinion Research, Hong Kong, June 14-16, 2012. Correspondence should be addressed to j.zhu @ cityu.edu.hk .

question the validity of search queries as measures of public opinion (see the upper panel of Table 1). Even if representative of the general population, search queries measure only one particular aspect – attention, interests, or concerns – of the underlying public opinion. While queries are entered by individual users, search engines display the query trends in an aggregated format, to protect user privacy. As such, it is impossible to know the weight by which each user carries in creating the aggregated queries. It is reasonable to assume that the more active users are likely to enter more queries than the less active ones. If so, the query trends are not the outcome of a democratic process in which each participant carries an equal vote.

Table 1. Comparisons between opinion polls and search queries

	Opinion Polls	Search Queries
Participants	General population	Internet users
Range of opinions	All aspects	Attention
Unit of analysis	Individual	Aggregated
Individual identity	Known	Unknown
Individual weight	Equal	Unknown
Recruitment	Selected	Volunteered
Opinion Source	Solicited	Self-initiated
Unobtrusiveness	Low	High
Cost	High	Low
Sample size	Hundreds	Up to billions
Time frequency	Monthly-annually	Daily-weekly

Despite these limitations, search queries do offer several desirable merits in comparison with public opinion polls (as shown in the lower panel of Table 1). While survey respondents are selected by pollsters, search users are volunteers. As such, the views expressed in polls are solicited, which often lead to the question of whether survey results are genuine opinions or pseudo opinions artificially created under the pressure from survey staff. In contrast, self-initiated search queries represent what the users, who are likely to come from a particular segment of the population, are truly curious or worried about at the moment. The unobtrusive nature of search behavior lends high credibility to the resulting query data, provided that the search engines do not alter the data, which is generally unknown but should and often could be verified by the researcher who uses the query data using cross-validation from multiple search engines (as shown in next section).

On more technical terms, query data are far cheaper (in fact free), of larger size (although usually unknown) and in much smaller time unit (e.g., daily or weekly). These features make it possible for the researcher to carry out more sophisticated analyses, usually based on time series analysis framework, to test dynamic issues in public opinion process.

In summary, search queries present public opinion researchers a gold mine of rich data with the representativeness of the general population remaining unknown. The current study aims to fill the gap by assessing the validity of search queries as indicators of public opinion based on the correspondence between search queries and public concerns over the same set of public issues in Shenzhen, China. A unique feature of the current study is that we are able to validate search queries against the data from a 5-year long tracking survey, in which opinions from both users and nonusers of the Internet are available for the assessment. Therefore, the question of whether search queries represent the general public or only active users of the Internet receives direct testing for the first time in the literature.

With a population over 10 million, GPD per capita above US$15,000 (Shenzhen Statistical Bureau, 2011), and Internet penetration rate over 50% (based on the surveys described below), Shenzhen resembles the "average" city of the world in many aspects, which makes the findings of the current study to be generalizable beyond the locality.

Method

The current assessment study draws on data from two independent sources: online search engines and offline telephone surveys.

Telephone Survey Data

The Public Opinion Survey Lab at Shenzhen University, China, carried out a monthly survey from October 2006 to December 2011 to gauge concerns of the local residents in Shenzhen. In each of the surveys, 200-1,000 randomly selected adults were successfully interviewed by telephone. They were asked a battery of questions about their concerns or satisfactions, with various aspects of their personal and family life in the city. While the average sample size (610) is adequate, about one-fifth of samples fall under 400 each. Therefore, we

aggregate the 63 monthly samples into 21 quarterly samples to ensure an adequate number of cases (mean = 1,650, minimum = 750, and maximum = 2,800) in each of the 21 time points.

For the current study, we choose to focus on three questions that bear explicitly specific keywords that can be easily matched with search queries, including housing conditions ("*juzhu zhuangkuang*" in Chinese), traffic conditions ("*jiaotong zhuangkuang*"), and property crimes (e.g., theft and robbery, or "*touqie qiangjie*" in Chinese).

Search Query Data

Using the keywords (in Chinese) associated with the above survey questions, we searched the relevant query trends collected by two major search engines – Google Trends and Baidu Index – over the same time period. Since both systems provide filtering on location, we carry out the same search twice, one in the national scope (i.e., mainland China, to serve a reference of comparison) and another in the local scope (i.e., Shenzhen in Baidu Index or Guangdong in Google Trends[1]). The resulting query data from Google Trends are in text file format, directly downloadable, whereas the data from Baidu Index are in image file format, which we converted to numeric values using a tailor-made tool of pattern recognition.

In addition to the above keywords, we also carried out a search of "baseline query" using the most frequently used word in Chinese ("*de*", which is a preposition word equivalent to "*of*" in English in both semantics and popularity). The purpose of charting such a trivial word is to obtain a measure that approximates the total search volume on both search engines throughout the period of time under study, which is not publically known. The baseline query helps control for artificial trends in substantive queries under study, which may be caused by mere increase in Internet users (that has indeed the case in Shenzhen and elsewhere in China and most other countries) or interruptive events (e.g., Google's removal of search engine servers from mainland China to Hong Kong).

As shown in Figure 1, while the search volume has been on the rise during the 51-month period for both search engines at both national and local levels (which is in line with the continuous growth of Internet population in China), the trends in Google Trends appear to be more steady over time (which shows that the relocation of Google's servers out of China may cast little impact on the search traffic) and more consistent between the national and local levels (r = 0.99) than the trends in Baidu Index. The finding, coupled with other irregularities found in the trends of various keyword queries, has led us to focus exclusively on Google Trends for the subsequent analyses.

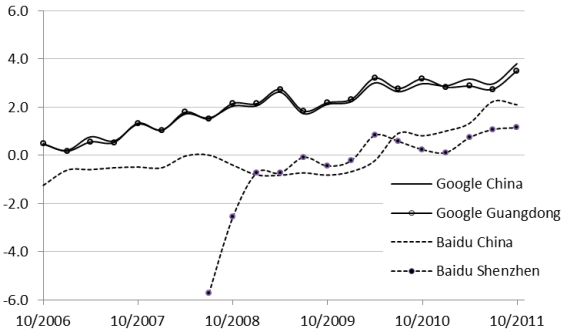

Figure 1. Total volume of baseline query ("*de*") in Baidu Index (in z-score) and Google Trends (in z-score + 2, to be visually distinguishable from the Baidu series).

Data Analysis Strategy

The current study differs from the previous studies reviewed earlier in terms of analysis approach. Aiming to predict real life indicators, most of the previous studies adopt a time series analysis framework, in which elaborated tools such as VAR and ARIMA models are used to control for autocorrelation embedded in time series data. The current study is designed for a different purpose – to assess the similarity between search queries and opinion polls. Although both query data and poll data are of time series nature, we do not remove autocorrelation from either series to preserve the endogeneity (i.e., authenticity) of the data. Instead, we introduce a different type of control (i.e., the baseline query, as described earlier) to the analysis, which in fact substantially reduces the autocorrelation in most series.

In particular, we performed OLS regression with each search query series as the criterion variable, the corresponding survey question series and the baseline query series as two predictors. However, this specification does not suggest a causal relationship between search queries and survey responses, which is not what the current study aims

[1] Guangdong is a province in which Shenzhen is located as province is the smallest locality Google Trends allows to filter.

to test. Instead, it provides a convenient way for us to examine the net correlation (i.e., the standardized regression coefficient) between the queries and the responses while controlling for possible confounding effects of the baseline query.

For each relationship between queries and responses, we repeated the regression analysis three times, the first for the entire sample, the second for the subsample of Internet users (accounting for 52% of the total sample), and the third for the subsample of nonusers (48%). In total, we carried out 30 regressions (= (7 keywords + 3 combinations of the keywords) X (1 total sample + 2 subsamples). For ease of reading, only the correlation (i.e., the standardized regression coefficient) between each pair of query trend and response trend from each of the 30 regressions is presented in Table 2.

Findings

A general pattern emerged from Table 2 is that some of the query series are, individually or collectively, significantly correlated with the corresponding opinion series, with the strength of the association ranging from modest to fair strong. In general, the relationship is the strongest for issue of property crimes and the weakest for the issue of traffic conditions.

Table 2. Standardized regression coefficients of search Queries on survey responses with the baseline query controlled

	Survey Responses		
Query Keywords	Total	Users	Nonusers
Traffic Conditions			
Jam	.26	.19	.38*
Transportation	.22*	.19	.20
Combined	.38*	.33*	.45**
Living Conditions			
Housing	-.09	-.14	-.05
Housing price	.54**	.49*	.42*
Combined	.28	.22	.23
Property Crimes			
Thief	-.13	-.11	-.11
Stealing	.60*	.51*	.58*
Robbery	-.17	-.16	-.19
Combined	.02	.02	.02

*** $p < .001$, ** $p < .01$, * $p < .05$

Another general pattern, perhaps somewhat counterintuitive, is that there is no substantial difference between Internet users and nonusers. That is, the concerns expressed by both users and nonusers in the surveys match with the query trends equally well (or equally poorly, for that matter).

With the two general findings from formal test in mind, we take a closer look at each of the issues with visual inspections, which help illustrate intuitively why some of the correlations are strong and other weak and why little differences are detected between users and nonusers of the Internet.

Traffic Conditions

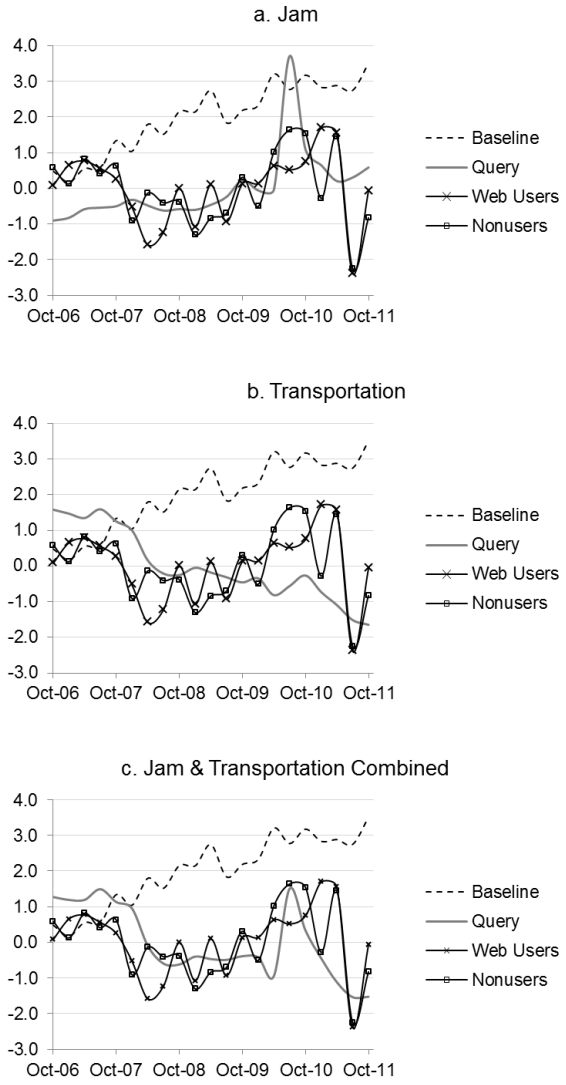

Figure 2. Trends between search queries and survey responses on traffic conditions

Users and nonusers of the Internet display a similar trajectory of rise and fall of concerns over city traffic, which is shown in Figs. 2a-2c. However, the trajectories do not match with the trends represented by two specific query words about traffic (i.e., "jam" or "*duche*" in Chinese, Fig. 2a,

and "transportation" or "*jiaotong*" in Chinese, Fig. 2b). A comparison of Figs. 2a and 2b reveals that the two query series are in opposite directions. When combined into one series, search queries and survey responses become more parallel (Fig. 2c).

Housing Conditions

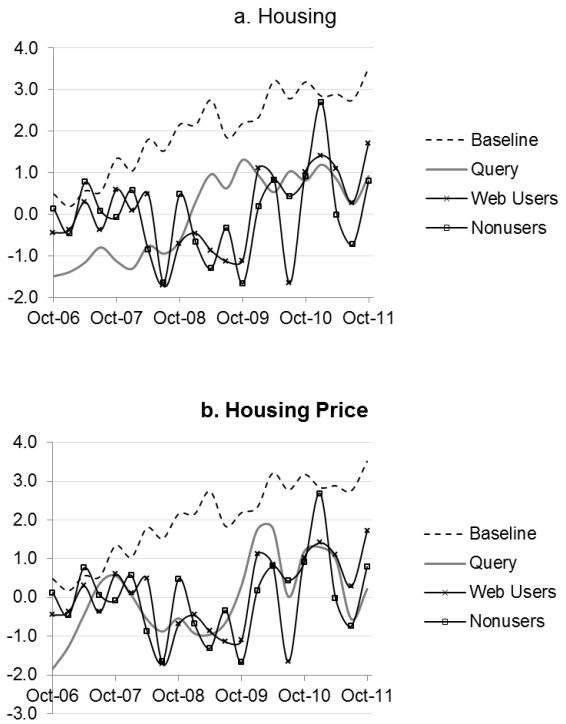

Figure 3. Trends in search queries vs. survey responses on living conditions

Figure 3a illustrates why the search query of "housing" ("*fangzi*" in Chinese) does not match the corresponding public opinion. While the query series is largely parallel to the baseline trend, which suggests the steady increase in the search count of "housing" was probably caused by the mere growth of search activities during the period, the responses registered in the surveys were far more fluctuated, particularly among the nonusers.

The trends in Figure 3b, based on the query of "housing price" ("*fangjie*" in Chinese), look more detectable because the movement of the three trends (query, user opinion, and nonuser opinion) is a lot more synchronized between each other. In particular, all went through the same two cycles of ups and downs from 2009Q4 to 2011Q4, which could hardly be attributed to random fluctuations.

Property Crimes

Finally, users and nonusers of the Internet again demonstrate a high degree of congruency in their responses to survey questions of crimes and social order, as shown in Figs. 4a-4c. However, of the three keywords used to capture public attention to property crimes, only "stealing" (or "*touqie*" in Chinese) follows a trend that resembles the survey responses (Fig. 4b). No such pattern is observed in the other two keyword series, including "thief" (or "*xiaotou*" in Chinese), which had a sharp surge in 2008Q4 but remained flatly low otherwise (Fig. 4a), and "robbery" (or "*qiangjie*" in Chinese), which ran almost in an opposite direction as the trend in survey responses (Fig. 4c).

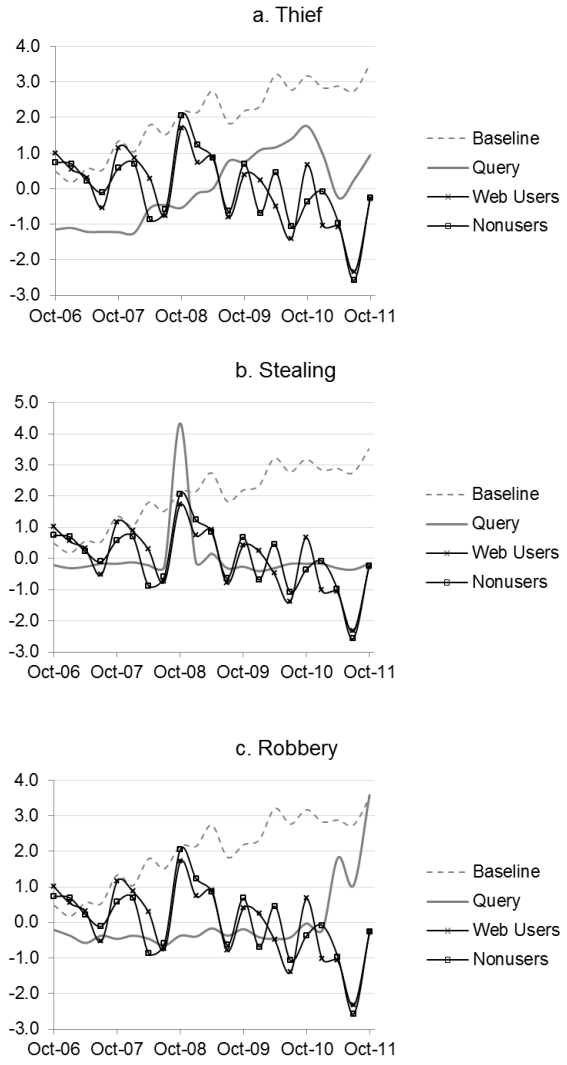

Figure 4. Trends in search queries vs. survey responses on property crimes

Discussion

Thirty years ago, David P. Fan, a seasoned biologist, appeared in conferences on public opinion research to show his ideodynamic model for predicting public opinion based on news coverage (Fan, 1988). He believed that the approach might drive public opinion pollsters out of business because public opinion could be readily predicted from news. While his work has exerted significant impact on public opinion research, including our own studies of media agenda-setting (Zhu, 1992; Zhu et al., 1993), his proposal to use computerized content analysis of news stories to displace opinion polling has yet been embraced probably because of the fundamental differences between news (a product from the profit-seeking media industry) and public opinion.

However, as has been repeated documented, opinion polling has become increasingly expensive and difficult given the continuous rise in labor cost, decline in household subscription to fixed telephone line, elevation in refusal rate, and other troubling trends. Against this backdrop, search query data from search engines emerge to be an attractive and promising alternative to, if not a complete displacement of, public opinion polling. As the current study shows, search queries do capture concerns of the general public, with a varying degree of accuracy across different issues. Of the three issues under study, queries and polls match the best on property crimes, which is arguably the most negative (but not necessarily the most important), but match the least on city traffic, which may be considered by the residents a problem of routine nature. Of course, our conjuncture on the relationship between the severity of the issue and the representativeness of search queries require formal test of a larger set of issues.

In addition to issue characteristics, the validity of search queries as indicators of public opinion also depends on how thoughtful and creative search queries are queried. In other words, a casual search in Google Trends is not guaranteed to yield quality data on the underlying public opinion. As we have learned from the current study, specific keywords without ambiguous boundaries or multiple meaning usually works well. It certainly deserves further investigation, preferably using an experimental design, to explore and identify a white list of "good" keywords as well as a black of "bad" keywords for common issue categories of public interests.

The study does not find any difference between users and nonusers in terms of their concerns over on the three issues under examination, which means that the widespread worry that search queries only represent the interests of Internet users is unfounded. In fact, this is not the first time that empirical studies of longitudinal data produce consistent evidence to refute commonly-held assumptions of fundamental differences between active and passive members of the population. For example, we have found that cognitively sophisticated audiences were equally susceptible as their naïve counterparts to media agenda-setting effects since the trajectories of responses to media agenda from all segments of the population were highly uniform over a 10-year period (Zhu & Boroson, 1997).

As noted earlier, even if representative of the general population, search queries measure only a specific aspect (i.e., attention) of public opinion, leaving other cognitive, evaluative, and affective dimensions of public opinion out of the equation. The latter are better captured by other social media such as posts, tags, comments, votes, and even photos and videos on forums, blogs, social networks, microblogs, and the like. Search queries alone are not likely to displace opinion polling. However, rigorous and creative mining of user generated content across all social media platforms may help realize Fan's 30-year long dream to eventually minimize, if not completely eliminate, the laborious and costly opinion polling.

References

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. *Applied Economics Quarterly*, 55,107-120.

Carneiro, H. A., & Mylonakis, E. (2009). Google Trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases*, 49, 1557-1564.

Choi, H., & Varian, H. (2009). Predicting initial claims for unemployment benefits. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/papers/initialclaimsUS.pdf. Accessed May 31, 2012.

D'Amuri, F. (2009). Predicting unemployment in short samples with internet job search query data. *Working Paper 18403, Munich Personal RePEc Archive.* http://ideas.repec.org/p/pra/mprapa/18403.html. Accessed May 31, 2012.

D'Amuri, F., & Marcucci. J. (2009). Google It! Forecasting the US unemployment rate with a Google Job Search Index. http://ideas.repec.org/p/pra/mprapa/18248.html. Accessed May 31, 2012.

Ellis, W. C., Ripberger, J. T., & Swearingen, C. D. (2011). Examining the impact of public attention on fundraising in U.S. senate elections. Paper presented at the Annual Meeting of the American Political Science Association, Seattle, WA, September 1-4.

Fan, D. P. (1988). *Predictions of public opinion from the mass media: Computer content analysis and mathematical modeling*. New York: Greenwood Press.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457, 1012-1014.

Kholodilin, K. A., Podstawski, M., & Siliverstovs, B. (2010). Do Google searches help in nowcasting private consumption? A real-time evidence for the U. S. *Working Paper 256, KOF Swiss Economic Institute*. http://ssrn.com/abstract=1616453. Accessed May 31, 2012.

Pelat, C., Turbelin, C., Bar-Hen, A., Flahault, A., & Valleron, A. J. (2009). More diseases tracked by using Google Trends. *Emerging Infectious Diseases*, 15:1327-1328.

Polgreen, P. M., Chen, Y. L., Pennock, D. M., & Nelson, F. D. (2008). Using Internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47, 1443-1448.

Reilly, S., Richey, S., & Taylor, J. B. (2012). Using Google Search data for state politics research: An empirical validity test using roll-off data. *State Politics & Policy Quarterly*, 12(2), 146-159.

Ripberger, J. T. (2011). Capturing curiosity: Using Internet search trends to measure public attentiveness. *Policy Studies Journal*, 406, 239-260.

Schmidt, T., & Vosen, S. (2009). Forecasting private consumption: Survey-based indicators vs. Google Trends. *Ruhr Economic Paper 155*. http://ssrn.com/abstract=1514369. Accessed May 31, 2012.

Shenzhen Statistical Bureau (2011). *Statistical yearbook of Shenzhen, 2011*. Beijing: China Statistical Publishing House.

Shimshoni, Y. Efron, N., & Matias, Y. (2009). On the predictability of Search Trends. *Google Technical Report*. http://static.googleusercontent.com/external_content/untrusted_dlcp/research.google.com/en/us/archive/google_trends_predictability.pdf. Accessed May 31, 2012.

Suhoy, T. (2009). Query indices and a 2008 downturn: Israeli data. http://www.bankisrael.gov.il/deptdata/mehkar/papers/dp0906e.htm. Accessed May 31, 2012.

Wilson, K., & Brownstein, J. B. (2009). Early detection of disease outbreaks using the Internet. *Canadian Medical Association Journal*, 180, 829-831.

Zhu, J. H. (1992). Issue competition and attention distraction: A zero-sum theory of agenda-setting. *Journalism Quarterly*, 69(2), 825-836.

Zhu, J. H., & Boroson, W. (1997). Susceptibility to agenda-setting: A cross-sectional and longitudinal analysis of individual differences. In M. E. McCombs, D. L. Shaw, & D. P. Weaver (Eds.), *Communication and Democracy* (pp. 69-83). Lawrence Erlbaum Associates.

Appendix

Wording of the Relevant Questions in Shenzhen Monthly Surveys

Q4. Do you think the crime rate in Shenzhen over the past month is higher/lower than the usual situation in the city? ("*Crimes*")

1. Lower
2. Higher
3. Don't know (excluded from the analysis)

Q12. Are you satisfied with your living conditions in Shenzhen over the past month? (*Living Conditions*)

1. Satisfied
2. Half and half
3. Unsatisfied
4. Don't know (excluded from the analysis)

Q18. Are you satisfied with the traffic conditions during the commuting hours over the past month? (*Traffic Conditions*)

1. Satisfied
2. Half and half
3. Unsatisfied
4. Don't know (excluded from the analysis)