The Future of Prediction:

How Google Searches Foreshadow Housing Prices and Sales

Lynn Wu
MIT Sloan School of Management
50 Memorial Drive, E53-314
Cambridge, MA 02142
lynnwu@mit.edu

Erik Brynjolfsson
MIT Sloan School of Management
50 Memorial Drive, E53-313
Cambridge, MA 02142
erikb@mit.edu

This Draft: Dec 2, 2009

Comments Welcome

Abstract

Most data sources used in economics, whether from the government or businesses, are typically available only after a substantial lag, at a high level of aggregation, and for variables that were specified and collected in advance. This hampers the effectiveness of real-time predictions. We demonstrate how data from search engines like Google provide a highly accurate but simple way to predict future business activities. Applying our methodology to predict housing market trends, we find that a housing search index is strongly predictive of the future housing market sales and prices. Specifically, each percentage point increase in the housing search index is correlated with additional sales of 67,220 houses in the next quarter. The use of search data produces out-of-sample predictions with a mean absolute error of just 0.102, a substantial improvement over the 0.441 mean absolute error of the baseline model which uses conventional data but does not include any search data. We also demonstrate how these data can be used in other markets, such as home appliance sales. In the near future, this type of "nanoeconomic" data can transform prediction in numerous markets, and thus business and consumer decision-making.

Keywords: Online Search, Prediction, Housing Trends

"It's difficult to make predictions, especially about the future"
-- Attributed to Neils Bohr

Introduction

Traditional economic and business forecasting has relied on statistics gathered by government agencies, annual reports and financial statements. Invariably, these are published after significant delay and are aggregated into a relatively small number of pre-specified categories. This limits their usefulness for predictions, especially novel predictions. However, due to the widespread adoption of search engines and related information technologies, it is increasingly possible to obtain highly disaggregated data on literally hundreds of billions[1] of economic decisions almost the instant that they are made. Now, query technology has made it possible to obtain such information at nearly zero cost, virtually instantaneously and at fine-grained level of disaggregation. Each time a consumer or business decides to search for a product via the Internet, valuable information is revealed about that individual's intentions to make an economic transaction (Moe & Faber, 2004). In turn, knowledge of these intentions can be used to predict demand, supply or both. This revolution in information and information technology is well underway and it portends a concomitant revolution in our ability to make business predictions and ultimately a sea change in business decision-making. This new use of information technology is not a mere difference in degree, but a fundamental transformation of what is known about the present and what can be known about the future.

Assisting with predictions has always been a central contribution of social science research. In the past several decades, much of social science research has focused on ever more complex mathematical models, for many types of important business and economic predictions. However, the latest recession has shown that none of the theoretical models was intelligent enough to foresee the biggest economic downturn in our recent history (Krugman, 2009). Perhaps, instead of honing techniques to extract information out of noisy and error-prone data, social science research should focus on inventing tools to observe phenomenon at a higher resolution (Simon, 1984). Search engine technology has precisely delivered such a tool. By effectively aggregating consumers' digital traces and improving data quality by several orders of magnitude, information technology has created a transformation on how we solve the problem of predicting the future. With the observation of billions of consumers and business intentions

[1] Americans performed 14.3 billion Internet searches in March, 2009, which is an annualized rate of over 170 billion searches per year. Worldwide searchers group by 41% between 2008 and 2009.

as revealed by online search, we can significantly improve the accuracy of predictions about future economic activities.

In this paper, we demonstrate how data on Internet queries could be used to make reliable predictions about both prices and quantities literally months before they actually change in the marketplace. We use the housing market as our case example but our techniques can be applied to almost any market where Internet search is non-trivial, which is to say, an increasingly large share of the economy. What's more, by identifying correlations with prices and quantities we can make inferences about changes in the underlying supply and demand. Our techniques can be focused on particular regions or specific cities, or the nation as a whole, and can look at broad or narrow product categories. Search not only precedes purchase decisions, but in many cases is a more "honest signal" (Pentland, 2008) of actual interests and preferences since there is no bargaining, gaming or strategic signaling involved, in contrast to many market-based transactions. As a result, these digital traces left by consumers can be compiled to reveal comprehensive pictures of the true underlying economic intentions and activities. Using aggregation of query data collected from the Internet has the potential to make accurate predictions about areas as diverse as the eventual winners of standard wars, or the potential success of product introductions.

The Real Estate Market

We use the real estate market to demonstrate how online search can be used to reveal the present economic activities and predict future economic trends. Studying the real estate market is especially important at the awake of the recent burst of the real estate bubble that has triggered the current economic downturn in the US and the rest of the world. In turn, when the housing market becomes healthy again, the recession may also come to an end (New Work Times Editorial, 2009). Economists, politicians and investors alike are pouring over government data released every month to assess the current housing market and predict its recovery and subsequently the end of the current recession. However, government data are often released with a lag of months or more, rendering a delay in assessing the current economic conditions. We propose a different way to predict the future housing price via the frequency of online search terms. Analyzing consumers' interests as revealed by their online behaviors, we are able to uncover sales trends before they appear in published data.

The Internet is a valuable research tool and can provide critical information to make purchase decisions (Horrigan, 2008). As the Web becomes ubiquitous, more shoppers are using the Internet to gather information and narrow down the number of selections, especially for products that require a high level of financial commitment, such as buying a home. According to the 2008 Profile of Home Buyers and Sellers by National Association of Realtors (NAR), 87% of home buyers used the Internet to search for a home in 2008 (NAR, 2008). Similarly, a report, written by California Association of Realtors in 2008, shows that 63% of home buyers find their real estate agent using a search engine (Appleton-Young, 2008). To explore the link between search and actual sales, we analyze billions of individual searches from five years of the Google Web Search portal[2] to predict housing sales and housing prices. Using these fine-grained data on individual consumer behaviors, we built a comprehensive model to predict housing market trends.

We found evidence that queries submitted to Google's Search Engine are correlated with both the volume of housing sales as well as a house price index—specifically the Case-Shiller Index. The Case-Shiller index is a predominant housing index and is widely used in most government reports. We find that the search term frequency can be used to predict future housing sales. Specifically, we find that a one percentage point increase in search frequency about real estate agents is associated with selling an additional 67,700 future quarterly housing sales in the average US state.

Similarly, we also examine the relationship between housing price and housing related searches online. Using house price index (HPI) from Federal Housing Finance Agency,[3] we find a positive relationship between the housing related online queries and the present house price index. This appears to reflect an increase in housing demand, driven by home buyers who search for houses online prior to actually buying. Interestingly, the house price index is negatively correlated with housing queries three months prior. We infer this to correspond to an increase in the supply of available houses in the market. Sellers "move first" in this marketplace, surveying the competition and assessing market conditions before making a decision to sell. As more sellers reveal their intentions, more houses eventually become available for sale. In turn, the listing price is likely to fall, driving down the overall house price index.

---

[2] http://www.google.com/insights/search/#

[3] http://www.fhfa.gov/

In turn, we also find evidence that the total volume of houses sold is correlated with consumers' intention to purchase home appliances. We use the search frequency of home appliances to approximate their consumers' interests (Moe and Fader, 2004). We find that every thousand houses sold are correlated with a 1.23 percentage point increase in the frequency of search terms that are related to home appliances. This highlights the linkages between home sales and other parts of the economy that may complement home sales.

Literature Review

In the past decades, much of the social science research has focused on refining increasingly complex mathematical models to predict social and economic trends. However, the alternative of collecting high quality data at a much finer-grained levels has mostly eluded in social science research.

Today, advances in information technologies, such as the Internet search technologies, e-mail, smart sociometric badges, offer remarkable detailed records of human behaviors. Recently, researchers have started to take advantage of real-time data collected from these new technologies. For example, deploying sociometric badges to measure moment-to-moment interactions among a group of IT workers, Wu et al. (2008) has uncovered new social network dynamics that are only possible by accessing accurate data at micro-level. Similarly, Aral et al. (2007) used email data to capture real-time communication patterns of a group of people over several years. They were able to examine work behaviors, such as multitasking, and their impact on long-term work performance. Lazer et al (2009) provided various examples of how high quality data produced by novel technologies are transforming the landscape of social network research. Similarly, firms have also leveraged the massive amounts of data collected online to make predictions, such as consumer preferences, supplies and demands for various goods as well as basic operational parameters such as inventory level and turnover rate. The ability to collect and efficiently analyze the enormous amount of data made available by information technology has enabled firms, such as Amazon, Harrah's and Capital One, to hone their business strategies and to achieve tremendous gain in profitability and market shares (Davenport, 2006).

Our work follows a similar stream in demonstrating the power of using fine-grained data to predict underlying social and economic trends. Unlike previous research and businesses that have primarily used proprietary data, we leverage free and public available data from Google to accurately forecast economic trends. Research has shown that online

behaviors can be used to reveal consumers' intention and predict purchase outcomes (e.g. Moe and Fader, 2004; Kuruzovich et al. 2008). We believe that we can rely on digital traces left by trillions of online search to reveal consumers' intentions and examine their power to predict underlying social and economic trends. Using such fine-grained data to study individual buying or selling decisions could be called nanoeconomics.

Our methodologies are similar to a recent analysis on flu outbreaks using Google Flu Trends (Ginsberg et, al., 2009) and also parallel, but unpublished research by Choi and Varian (2009) where the authors also correlate housing trends in the US using search frequencies. While Choi and Varian (2009) mainly focus using search frequencies to reveal the current economic statistics, our work attempts to predict the *future* economic trends, such as forecasting price and quantity of houses sold in the future. Our work also use more fine-grained data at the state level instead of at the level of the whole nation to provide a more nuanced prediction of real estate market which often varies greatly depending on geographical locations.

Economics of Real Estate

Our work also contributes to the literature of real estate economics. There are two types of forecasting methodologies for predicting real estate market trends. The first is the technically analysis that is used to predict stock market trends. The main assumption for this type of analysis is that the key statistical regularities for the underlying housing price do not change. The trending behaviors are therefore more likely to exhibit long-term reversion to the mean but with short-term momentum (e.g. Case & Shiller, 1989). Glaeser and Gyourko (2007) found evidence of long-term reversion in housing price. They found that, ceteris paribus, when regional prices go up by an extra dollar over one five-year period, the regional price on average would drop by 32 cents over the next five years. The second approach to predicting housing market trends is to use fundamental economic analysis. Housing price should depend heavily on the cost of construction, the interest rate to finance the housing purchase, the regional income as well as the January temperature (Glaeser, 2009). In principle, this suggests that regions with steady building costs and relatively stable income level should have a steady housing price. However, these economic variables do not seem to fully capture housing price trends. For instance, in Dallas, an example of a region with steady fundamentals, the housing price has been increasing despite the predictions of fundamental analysis.

Some dynamic housing demand models try to incorporate both approaches to predict housing trends (Glaeser and Gyourko 2007, Han 2008, Han 2009). Using dynamic rational expectation to model housing price, Glaeser and Gyourko (2007) detects mean-reverting mechanism but they cannot explain serial correlation or price changes in most volatile markets. Glaeser (2009) suggests this may reflect sentiment or even "irrational exuberance" in some housing markets, generating a bigger boom and bust cycle than what are predicted by the model (Glaeser 2009).

With the ability to gather billions of search queries over time, Google Insight is essentially aggregating all the honest signals of decision-makers intentions to capture the overall level of "sentiment". This provides unprecedented opportunities to improve predictions in housing markets. Using very simple regression models, we demonstrate that Google search frequencies can be used as a reliable predictor for the underlying housing market trends both in the present and in the future.

Data Sources

Google Search Data

We collected the volume of Internet search queries related to real estate from Google Trends, which provides weekly and monthly reports on query statistics for various industries. It allows users to obtain a query index pertaining to a specific phrase such as "Housing Price". Since 2004, Google Trends has systematically captured online queries submitted to the Google Search Engine and categorized them into several predefined categories such as Computer & Electronics, Finance & Business and Real Estate. As Nielsen NetRatings has consistently placed Google to be the top search engine, which processed more than 60% of all the online queries in the world (Nielsen Report, 2008), the volume of queries submitted to Google has the potential to approximate people's interests over time. In fact, recent work using Google Search can accurately predicts flu outbreaks few days before it actually happens (Ginsberg et al, 2008). We believe that the search volumes can also be used to predict future economic indicators.

Figure 1: Search Index for "Real Estate Agencies. It is a normalized measure of search volume ranging from 0 to 100.

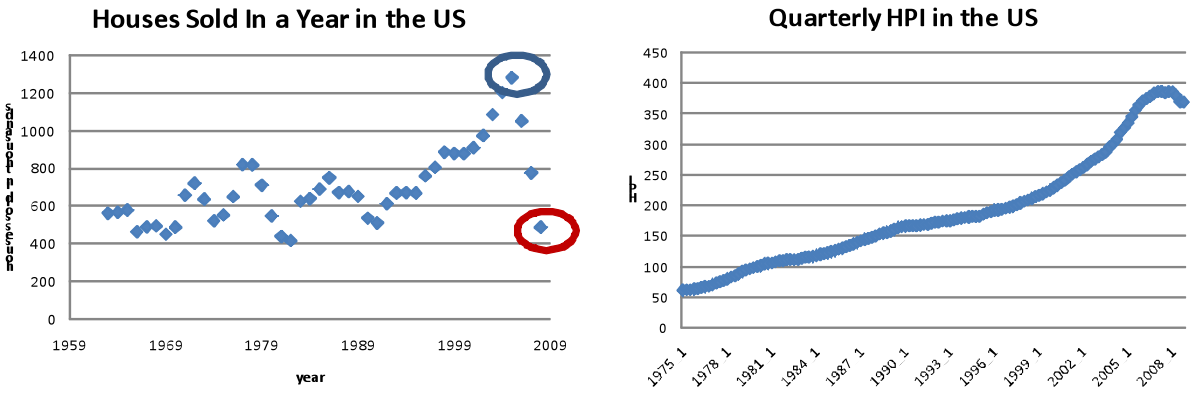


Figure 2: Housing and Prices of New House Sold in the US. (a) Number of New Houses Sold Annually. (b) Quarterly House Price Index.

Google Trends provides a search index for the volume of queries based on geographic locations and time. The search index is a compilation of all Internet queries submitted to Google's search engine since 2004. The index for each query term is not the absolute level of queries for a given search phrase. Instead, it reports a query index measured by query share, which is calculated as the search volume for each query in a given geographical location divided by the total number of queries in that region at a given point in time[4]. Thus, the index is always a number from 0 to 100. The reports on search index are also much more fine-grained than most government reports.

[4] http://www.google.com/support/insights//bin/answer.py?answer=87285

Typically, Google calculates the query index on a weekly or a monthly basis and can be disaggregated down to country, state/province and city levels around the world. For example, in the US, a query index can be calculated at the state level. A more detailed query index at the MSA level can also be computed by specifying the appropriate sub-regions within a state. Figure 1 shows the overall interest in real estate agencies using online searches. From the graph, interests in housing price peaked in 2005 and fell through 2009.

Our analysis uses a predefined category in Google Trends, "Real Estate" to approximate the overall interest for housing. This category aggregates all online search queries that are related to real estate. We also collected more fine-grained search index for several subcategories, such as the "Real Estate Agencies." We hypothesize that these housing related search index are correlated with the underlying conditions of the US housing market. To test this hypothesis, we gather housing market indicators, such as the volume of houses sold and the house price index.

Housing Market Indicators

We collect housing market indicators such as housing sales and the house price index (HPI) to examine their relationships with online search. The volume of housing sales is collected from National Association of Realtors (http://www.realtor.org/research) for all 50 states in the US and the District of Columbia from the 4th quarter of 2007 to the 2nd quarter of 2009. We also obtain the house price index for the same period at the Office of Federal Housing Enterprise Oversight (http://www.ofheo.gov/), where housing prices for nine Census Bureau divisions are collected. The Office of Federal Housing Enterprise Oversight calculated the HPI for each state in the US on a quarterly basis since 1975[5]. Detailed calculations of the HPI can be found at http://www.fhfa.gov/.

As shown in Figure 2(a), the number of houses sold in the US peaked at around 2005 and then declined precipitously soon after, reaching a historically low at the beginning of 2009. The HPI also increased gradually and reached a peak in 2007, two years after the housing sales peak (Figure 2b), and began to fall shortly after. Comparing housing market indicators (Figure 2) to their associated online search indices (Figure 1) shows that they appear to be closely correlated. As shown in Figure 1, housing related search peaked at 2005 and gradually declined

[5] http://www.fhfa.gov/Default.aspx?Page=81

to its lowest point in early 2009, mirroring the volume of houses sold in Figure 2(a). This provides some evidence that the search index may be correlated with the number of houses sold and the house price index.

Empirical Methods

We use a simple seasonal autoregressive (AR) model to estimate the relationship between the search index and the housing market indicators— the volume of housing sales, and the house price index. A single explanatory variable is studied: the search index for housing related queries for each state in the US. We first estimate the baseline model to predict the current housing sales using the past home sales and the HPI.

$$HomeSales_{it} = \alpha + \beta_1 HomeSales_{i,t-1} + HPI_{it,-1} + \Sigma S_i + \Sigma T_t + \varepsilon_{it} \qquad (1)$$

Then, we add a housing-related search index as an explanatory variable to see if it can predict housing sales. We incorporate both the present and the past search index into the regression model.

$$HomeSales_{it} = \alpha + \beta_1 HomeSales_{i,t-1} + HPI_{i,t-1} + \beta_2 SearchFreq_{it} + \beta_3 SearchFreq_{i,t-1} + + \Sigma S_i + \Sigma T_t + \varepsilon_{it} \qquad (2)$$

Next, we examine whether the housing related search index can be used to forecast the future housing sales. Since at any point in time, only the past, but not the present housing statistics are available, we can only use the past housing statistics to predict the future housing trends. In contrast, search frequencies are available in real-time. We can incorporate both the present and the past search index to predict future housing sales, as shown in the following model.

$$HomeSales_{it+1} = \alpha + \beta_1 HomeSales_{i,t-1} + HPI_{i,t-1} + \beta_2 SearchFreq_{it} + \beta_3 SearchFreq_{i,t-1} + + \Sigma S_i + \Sigma T_t + \varepsilon_{it} \qquad (3)$$

Similarly, we use the same approach to predict the current and the future HPI. In the baseline model, we use the past HPI and past housing sales to predict the current HPI.

$$HPI_{it} = \alpha + \beta_1 HPI_{i,t-1} + \beta_2 HomeSales_{i,t-1} + \Sigma S_i + \Sigma T_t + \varepsilon_{it} \qquad (4)$$

We then incorporate the current and past search indices into the baseline model to predict the present HPI.

$$HPI_{it} = \alpha + \beta_1 HPI_{i,t-1} + \beta_2 HomeSales_{i,t-1} \beta_3 SearchFreq_{it} + \beta_4 SearchFreq_{i,t-1} + \Sigma S_i + \Sigma T_t + \varepsilon_{it} \quad (5)$$

Lastly, we predict the future HPI using past HPI from periods earlier as well as the present and past search indices.

$$HPI_{it+1} = \alpha + \beta_1 HomeSales_{i,t-1} + HPI_{i,t-1} + \beta_2 SearchFreq_{it} + \beta_3 SearchFreq_{i,t-1} + + \Sigma S_i + \Sigma T_t + \varepsilon_{it} \quad (6)$$

We apply a state level fixed-effect specification to all our models in order to control for the influence of any time-invariant properties, such as the demographics of a state, and any statewide policies that may affect real estate purchase decisions.

We also examine whether housing related Internet search also spurs future economic activities for industries that complement home-buying activities. For example, if online search can indeed reveal consumers' intentions (Moe & Fader, 2004), we may also expect a surge in Internet queries about home appliances, shortly after observing a rise in home sales. Since new homeowners may plan to purchase appliances to furnish their property, tracking their online search behavior allow us to detect their intention to purchase home appliances. Accordingly, we correlate housing sales with the search index for home appliances. If search index for home appliance can translate into actual purchases, we would expect a rise in search frequencies for home appliances, spurred from home sales, to indicate a rise in their future demands as well.

$$HomeApplianceSearch_{it} = \alpha + \beta_1 HouseSold_{it} + \beta_2 HouseSolds_{i,t-1} + \varepsilon \quad (7)$$

Empirical Results

Predicting Home Sales Using Online Search

Using the search index captured by Google's search engine, we find a positive relationship between housing sales and online queries related to housing (Table 1). All models in Table 1 are based on a seasonal autoregressive (AR) model, which assumes that the sales for the current period are related to sales from the previous period. We see a broad support for this as the lagged sales are strongly correlated with the contemporary sales. We also applied a fixed-effect specification to each model to eliminate influence from any time-invariant properties. In addition, we included seasonality controls for time-specific changes by creating a set of dummy variables for each quarter of the

year. To capture online interests for purchasing real estate properties, we use the search index for a predefined category in Google Trends – "Real Estate" – that contains all queries pertaining to real estate. However, we realize this category may be too broad to infer the underlying home sales, since some of the queries, such as those related to property management, are irrelevant to buying or selling a home. We mitigate this by using the search index of a subcategory of real estate, "Real Estate Agencies", to approximate consumers' interests in actually purchasing or selling a home. We assume people who are looking for real estate agents online are more likely participate in a real estate transaction than those who search for property management.

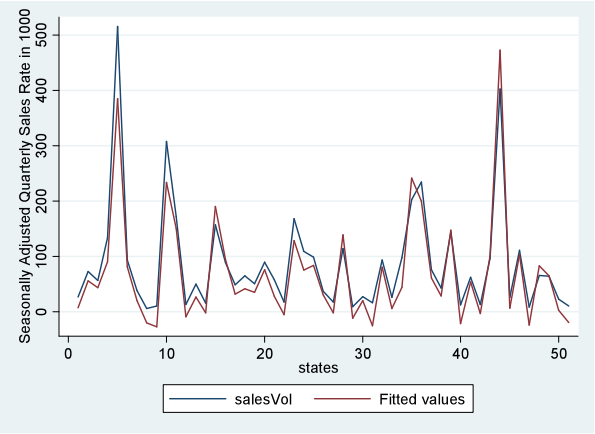
Table 1. Linear Regression to Predict the Present Home Sales Using Search Frequency

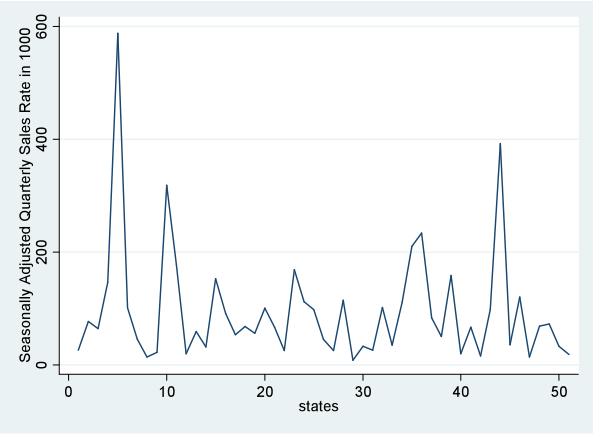
Dependent Var.	Quarterly Sales	Quarterly Sales	Quarterly Sales	Quarterly Sales	Quarterly Sales	Quarterly Sales
	(0)	(1)	(2)	(3)	(4)	(5)
$Sales_{t-1}$	0.491*** (0.0764)	0.440*** (0.0700)	0.429*** (0.0699)	0.424*** (0.0689)	0.445*** (0.0688)	
HPI_{t-1}	-0.751*** (0.0887)	-0.845*** (0.0822)	0.909*** (0.0840)	-0.901*** (0.0828)	-0.889*** (0.0822)	
Index "Real Estate Agencies" at t		119.7*** (18.65)		67.22** (25.52)	71.64** (33.25)	67.45** (32.00)
Index "Real Estate Agencies"--(t-1)			109.1*** (16.58)	67.21*** (16.84)	119.8*** (34.57)	44.43 (27.86)
Index "Real Estate" at t					-64.49 (52.36)	-57.01 (47.67)
Index "Real Estate" – (t-1)					-123.3 (63.40)	-115.7*** (40.05)
Obs.	304	304	304	304	304	304
Controls	Quarters	Quarters	Quarters	Quarters	Quarters	Quarters
States	51	51	51	51	51	51
Adjusted R^2	0.973	.980	.981	.982	.983	.970
F	29.55	36.52	37.12	33.77	27.71	4.16

*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's

First, we correlate the current home sales with the past housing statistics. As shown in the baseline AR(1) model (Model 0), the past housing price and quantities are correlated with the current housing sales. We then explore the effect of incorporating search frequencies into the baseline model as shown in Model 1-5. Overall, the contemporaneous and the past search index for the category "Real Estate Agencies" are found to have predictive power to forecast the current housing trends. As shown in Model 1, a one-percentage increase in the search index for the category "Real Estate Agencies" is associated with 119,700 additional houses sold in contemporaneous

quarter. Similarly, in Model 2, a one-percentage increase in the search index from the previous quarter is correlated with 109,100 houses sold in the next quarter. While both provides explanatory power to predict the contemporaneous home sales, the past search index provides a slightly better fit than the present search index, as indicated by the higher adjusted R^2 in Model 2. This provides some evidence that online search behavior has the predictive power to forecast economic activities.





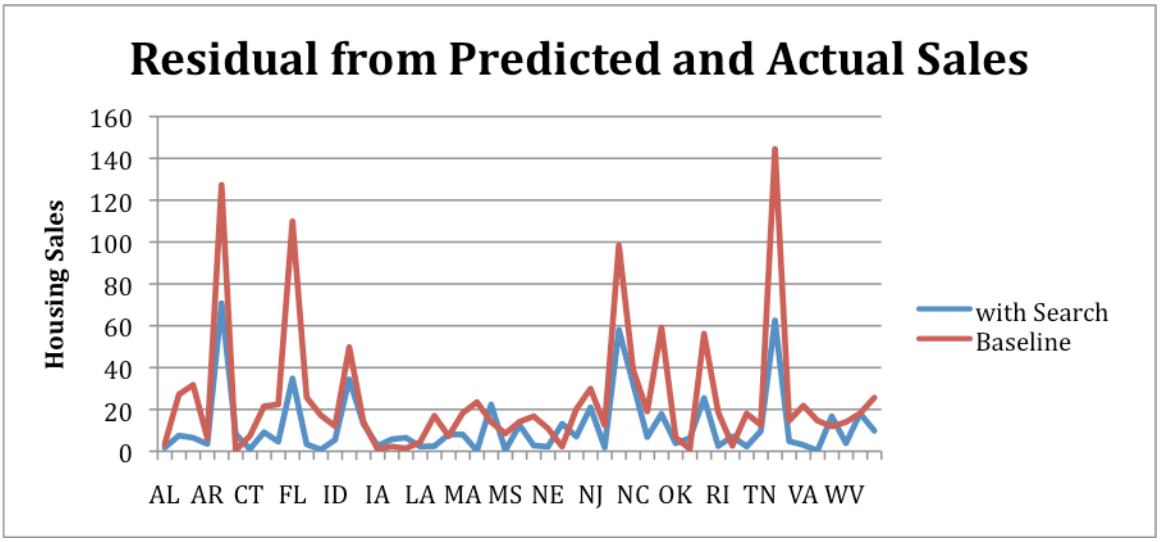


Figure 3: Seasonally Adjusted Quarterly Sales vs. Predicted. The state: a) Fitted or Predicted Value vs. the Actual Sales Volume for the 2nd Quarter in 2009, b) Predicted Value for the 3nd Quarter of 2009. c) Residuals of Predicted Housing Sales for 2009 Q3

Next, we examine both the current and past search indices and their correlations to housing sales. As shown in Model 3 in Table 1, both the current and the past search index for "Real Estate Agencies" continue to be positively correlated with the contemporaneous housing sales, demonstrating that past search frequencies could be used to predict future housing sales. The past search index ($\beta = 67.22$, $p < .001$), however, has a higher explanatory power than the current search index ($\beta = 67.22$, $p < .05$). In Model 4, we explore the effect of using both search indices for "Real Estate" and "Real Estate Agencies" in the same model. The past search index for "Real Estate Agencies" is again positively correlated with sales. However, we find that the indices for "Real Estate" are not correlated with the housing sales. This may be because that the "Real Estate" category contains queries that are not necessarily related to actual home sales. For example, some of these queries are related home insurance or property management, which are irrelevant to buying or selling a home. In Model 5, we only use the search indices to predict housing sales and we find similar results as what is shown in Model 4, where past sales and HPI are included. This suggests that using online search frequency alone can be used to predict future sales, even without knowing any housing statistics.

To examine whether our model can actually predict the contemporaneous home sales, we created a one-quarter-ahead prediction using data from the previous year to forecast the present housing sales. The actual values and the predicted values of quarterly housing sales for the second quarter of 2009 are well correlated, as shown in Figure 3a. We use mean absolute error (MAE) to evaluate the accuracy of the predictions. It is defined as

$$MAE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y_t - \hat{y}_t}{y_t}\right|.$$

The mean absolute error (MAE) using Model 4 of Table 1 is reduced to 0.102 from 0.441 (using the baseline model without search terms), which is about 77% better than the baseline (Model 0).

Figure 3b shows the predicted housing sales for the 3rd quarter of 2009 using Model 4 of Table 1. On August 26, 2009, we predicted that in the third quarter of 2009, nationwide average home sales would increase approximately 6.4% compared to the third quarter of 2008, with specific growth estimates for each of the states ranging from -35% (for the state of North Dakota) to 92% (for the state of Nevada). In November 25 2009, the housing statistics for the 3rd quarter of 2009 was released. The national average home sales increased 6.1% compared to the 3rd quarter of

2008. We find that the MAE for our prediction is 0.102 while the MAE for the baseline model (Model 0) is 0.442. This is a 75% increase in accuracy from the baseline model.

Table 2. Linear Regression to Predict the Future Home Sales Using Search Frequency						
Dependent Var.	Quarterly $Sales_{t+1}$	Quarterly $Sales_{t+1}$	Quarterly $Sales_{t+1}$	Quarterly $Sales_{t+1}$	Quarterly $Sales_{t+1}$	Quarterly $Sales_{t+1}$
	(0)	(1)	(2)	(3)	(4)	(5)
$Sales_{t-1}$	-.071 (.084)	-.127 (.0775)	-.149** (.075)	-.153** (.074)	-.133* (.075)	
HPI_{t-1}	-.835*** (.098)	-.939*** (.091)	-1.035*** (.089)	-1.029*** (.089)	-1.018*** (.089)	
Index "Real Estate Agencies" at t		131.8*** (20.63)		48.49* (27.55)	54.65 (36.03)	36.95 (29.96)
Index "Real Estate Agencies" at (t-1)			136.7*** (17.72)	106.5*** (24.57)	154.7*** (37.46)	63.58** (28.67)
Index "Real Estate" at t					-65.27 (56.77)	-96.59** (40.16)
Index "Real Estate" at (t-1)					-114.2* (58.46)	-60.87 (39.29)
Obs.	254	254	254	254	254	254
Controls	Quarters	Quarters	Quarters	Quarters	Quarters	Quarters
States	51	51	51	51	51	51
Adjusted R^2	0.971	.976	.978	.978	.979	.965
F	15.33	22.33	26.69	23.56	19.23	4.23
*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's						

Next, we apply our methods to predict the future housing trends with the data available today. Thus, we can only use the past housing statistics from the previous quarter since the present housing statistics would not be released yet. Unlike housing statistics, which is always released with a lag, we can obtain the contemporaneous search index from Google Insight, allowing us to incorporate these real-time search behaviors in predicting the future real estate trends. We show our results in Table 2, which largely supports our hypothesis that search indices can be used to predict future housing sales in the next quarter. The search category "Real Estate Agencies" continues to have a positive and statistically significant correlation with future home sales. As shown in Model 4, a one-percentage increase in the search index for "Real Estate Agencies" during the previous quarter is associated with additional 154,700 units of future sales two quarters later. We use Model 4 to predict the housing sales for the 3rd quarter of 2009 using the available housing statistics for the 1st quarter of 2009 as well as the search index in the first two quarters of 2009. The MAE for this model is 0.37 while the MAE for the baseline model (Model 0) is 0.47. This is a 21.3% improvement in MAE. Figure 4 shows the residuals from the prediction using Model 4 as well as the baseline

model. As shown in the Figure, the predictions from Model 4, which includes search indices as independent variables, has lower residuals than the baseline model that does not use the search indices.

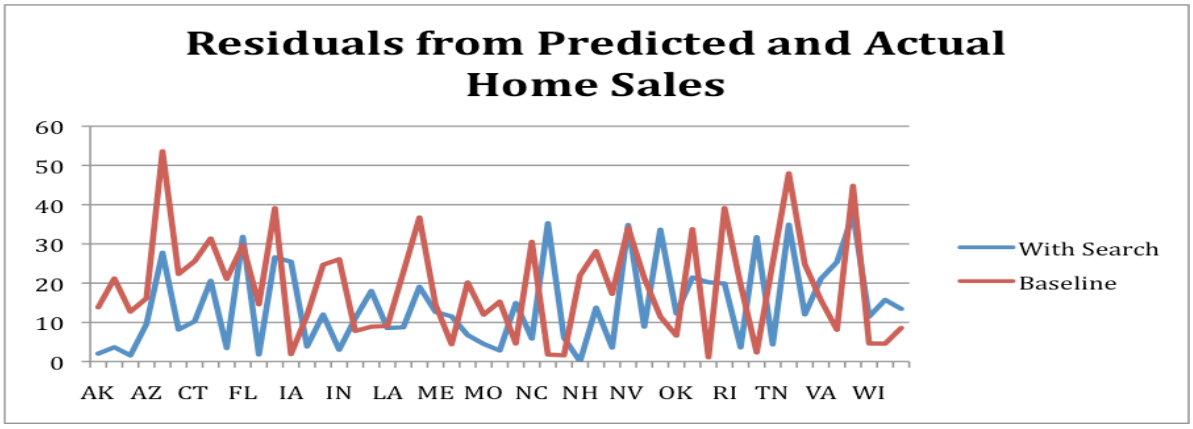


Figure 4: Residuals from Predicted and Actual Sales for Q3 in 2009 using prior housing statistics in Q1 of 2009.

Predicting the House Price Index Using Online Search Data

In Table 3, we explore the relationship between the housing related search index and the house price index, which is calculated based on a modified version of the weighted-repeat sales (WRS) methodology proposed by Case and Shiller (1989). Similar to models in Table 1, all the models in Table 3 use a fixed-effect specification on an AR model with seasonality controls. As expected from the AR model (Model 0), the lagged HPI and lagged sales are positively correlated with the present HPI. In Model 1, we estimate the correlation between the current search index for "Real Estate Agencies" and the HPI and find that a one-percentage increase in search index is associated with an increase of 15.06 points in HPI. However the past search index on "Real Estate Agencies" from the previous quarter is no longer statistically significant as shown in Model 2. Next, we introduce both the current and the past indices of for "Real Estate" in Model 4. We find that the current search index for "Real Estate" is positively correlated with the contemporaneous HPI. Specifically, a one-percentage point increase in the search index for "Real Estate" is associated with 32.97 HPI points. Then, we introduced the present and the past search index for both "Real Estate" and "Real Estate Agencies" in Model 4. Again, we find that only the present search index for "Real Estate" is statistically significant. This appears to reflect an increase in housing demand, driven by homebuyers who search for properties online prior to actually buying. We find that a one-percentage point increase in the search index for

"Real Estate" is associated with 35.93 points or a 10% increase compared to the average HPI from 2008 to 2009. As buyers are more likely to conduct search online prior to buying, past search indices can be used to predict future housing price.

Table 3: Linear Regression of HPI on the Search Index Related to Real Estate and Real Estate Agencies						
Dependent Var.	HPI_t	HPI_t	HPI_t	HPI_t	HPI_t	HPI_t
	(0)	(1)	(2)	(3)	(4)	(5)
HPI_{t-1}	.922*** (.015)	.894*** (0.016)	.889*** (0.017)	0.890*** (0.016)	0.884*** (0.016)	
$Sales_{t-1}$	.068*** (.010)	.056*** (0.011)	.053*** (0.011)	0.065*** (0.011)	.060*** (0.011)	
Index "Real Estate Agencies"		15.06*** (3.551)	11.30** (5.040)		-4.215 (6.601)	-21.23 (22.97)
Index "Real Estate Agencies" at (t-1)			4.767 (4.53)		9.149 (5.876)	41.20** (20.91)
Index "Real Estate"				32.97*** (7.169)	35.93*** (9.603)	88.94*** (30.67)
Index "Real Estate" at (t-1)				13.00* (7.034)	2.773 (9.258)	58.68** (28.27)
Obs.	356	356	356	356	356	356
Controls	Quarters	Quarters	Quarters	Quarters	Quarters	Quarters
States	51	51	51	51	51	51
Adjusted R^2	0.990	.991	.991	.991	.991	.982
F	1024.78	905.34	776.44	814.60	635.81	25.30
*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's						

Next we predict the future HPI in the next quarter. First we predicts the future quarterly HPI from existing housing statistics, which is always released with a lag of a quarter. As shown in Model 0 of Table 4, the future HPI is correlated with the past HPI two quarters prior, however the past sales from two quarters ago is not longer correlated with the current HPI. In Model 1, we incorporated the current search index for "Real State Agencies", and find that the current HPI is correlated with the future HPI in the next quarter ($b=131.8$, $p<0.001$). Similarly we find the past search index from the previous quarter is also correlated with the future HPI, two quarters later (Model 2). In Model 3, we incorporate both the current and the previous search indices to predict the future HPI. We find while both indices are correlated with the future HPI, the search index in the previous quarter has more explanatory power to predict the future HPI ($b =106.5$, $p<0.001$). Lastly, we used the present and past search indices for "Real Estate Agencies" and "Real Estate" in Model 4. The past search index for "Real Estate Agencies" remains highly correlated with the future search index two quarters later. In column 5, we used only the search index to predict the

future HPI. While they are still statistically significant, the overall fit of the model is worse than Model 4 where the past housing statistics are used.

Table 4: Linear Regression of HPI on the Search Index Related to Real Estate and Real Estate Agencies						
Dependent Var.	HPI_{t+1}	HPI_{t+1}	HPI_{t+1}	HPI_{t+1}	HPI_{t+1}	HPI_{t+1}
	(0)	(1)	(2)	(3)	(4)	(5)
$Sales_{t-1}$	-.071 (.084)	-.127 (.078)	-.149** (.075)	-.153** (.074)	-.133* (.075)	
HPI_{t-1}	-.835*** (.098)	-.939*** (.091)	-1.035*** (.090)	-1.029*** (.089)	-1.018*** (.089)	
Index "Real Estate Agencies" at t		131.8*** (20.63)		48.49* (27.55)	54.65 (36.03)	36.95 (29.96)
Index "Real Estate Agencies" at (t-1)			136.7*** (17.72)	106.5*** (24.57)	154.7*** (37.46)	63.58** (28.67)
Index "Real Estate" at t					-65.27 (56.77)	-96.59** (40.16)
Index "Real Estate" at (t-1)					-114.2* (58.46)	-60.87 (39.29)
Obs.	254	254	254	254	254	254
Controls	Quarters	Quarters	Quarters	Quarters	Quarters	Quarters
States	51	51	51	51	51	51
Adjusted R^2	0.971	.976	.978	.978	.979	.965
F	15.33	22.33	26.69	23.56	19.23	4.23
*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's						

We explore how using search frequencies can distinguish the supply from the demand in the housing market by refining the search categories to differentiate searches that are buyer specific from those that are seller specific. For example, home buyers are more likely to look for loans than sellers whereas sellers are more likely to hire a staging company to make the property more appealing to the highest number of potential buyers. We would therefore expect that an increase in search frequencies related to financing and loans to shift the demand curve while a similar increase for searches related to home staging is more likely to shift the supply curve for housing. We see some evidence that home financing are positively correlated with HPI, suggesting it may be shifting the demand outward. However, we have not found a set of queries that could shift the supply curve.

Predicting the Demand for Home Appliances

Lastly, we explore trends in home appliance sales. We expect that housing sales would spur interests in buying home appliances, increasing their demand in the future. To gauge the overall interests in home appliances, we use the search index for the "Home Appliance" category from Google Trends and show its relationship with home sales (Table 3). We observe that the current home sales are not correlated with the contemporaneous search index for home appliances (Model 1, Model 4). But after 6 months, each one thousand houses sold previously is correlated with a 1.14 percentage point increase in the search index for home appliances. Since buyers move into their new properties first before making major purchases (and often researching such purchases), it is natural that the number of online searchers for home appliances would only increase after a consumer has already bought a house. Thus, we may expect the online search for home appliances to lag behind housing sales. The actual demand for home appliance may rise after this increase in the appliance search index if some of the online searches translate into future sales. Similarly, we correlated the housing real estate related search index with the home appliance search index and we find that they are also positively correlated (Column 2 Table 5) This highlights the linkages between home sales and other parts of the economy that may complement real estate purchases.

Table 4: Linear Regression of HPI on the Search Index Related to Real Estate and Real Estate Agencies		
Dependent Var.	HPI_{t+1}	HPI_{t+1}
	(0)	(1)
Home Sale Vol	.188 (0.359)	
Home Sale Vol – lagging 1 quarter	-0.627 (0.393)	
Home Sale Vol – lagging 2 quarters	1.14*** (0.427)	
Search index "Real Estate Agencies—t-1		.051** (023)
Search index "Real Estate Agencies—t-2		.036 (028)
Obs.	306	306
Controls	Quarters	Quarter
States	51	51
Adjusted R^2	.973	.982
Obs.	306	306
*p<.1, **p<.05, ***p<.01, Huber-White robust standard errors are shown in parentheses. Quarterly Sales are in 1000's		

Implications

Twenty-five years ago, Herbert Simon (1984) observed:

> "In the physical sciences, when errors of measurement and other noise are found to be of the same order of magnitude as the phenomena under study, the response is not to try to squeeze more information out of the data by statistical means; it is instead to find techniques for observing the phenomena at a higher level of resolution. The corresponding strategy for economics is obvious: to secure new kinds of data at the micro level"

Today, advances in information technology in general, and Internet search query data in particular, are making Simon's vision a reality. Who could have imagined that we would be observing literally billions of consumer and business intentions to buy or sell before they even occur in the marketplace? Yet, that is what search query data does. What's more, we can do so at nearly zero cost, virtually instantaneously and at remarkably fine-grained levels of disaggregation. These data are increasingly available to ordinary consumers, business people and researchers of all types.

We have found that analyzing online search data with relatively simple models can yield more accurate predictions about the housing market than were previously possible. If online search patterns can be construed as a broad indicator of interest within a group, it can also be used as a reliable predictor to forecast economic activity. Analyzing housing market trends, we find evidence that the housing search index is correlated with both housing sales and the house price index. This correlation lends support to the hypothesis that Web search can be used to predict present and future economic activity. For example, housing-related search can be used to predict the recovery of the currently embattled housing market and potentially, when the economy may recover from the current recession.

Timely and accurate predictions about the housing market can benefit a wide array of industries, such as construction and home appliances, as well as individuals, such as homebuyers and sellers. Since buying a home is the single biggest expenditure and one of the biggest financial decisions for most people, obtaining accurate and timely information can help them make informed decisions and potentially save tens of thousands of dollars for the average family.

Similarly, businesses that depend on the housing market can benefit from this simple use of Internet search data. Currently, economists and investors primarily rely on housing data released from the government and trade groups such as the National Association of Realtors, to understand the current housing market and forecast future market trends. However, government and trade group data are released with a delay and often with pending revisions. Furthermore, they do not provide fine-grained reports at the town level that is crucial for buyers and sellers to make informed decisions. With easy access to billions of online search frequencies, it is now possible to use a simple technology to cheaply collect timely, accurate and fine-grained analysis about the housing market. Not only does Google Trends provide weekly reports on the volume of housing related queries, it also offers a detailed regional analysis at country, state and city levels. By leveraging micro data collected from Google Trends, investors can obtain deeper insight about the housing market in order to make informed decisions.

Accurate predictions on the housing market can also have strong ripple effects on other sectors of the economy, especially for its complementary goods. For example, timely and accurate forecasts of housing demand allow the construction industry to improve future plans for developments and thus reduce the probability of experiencing the housing boom and bust cycles. Similarly, accurate housing market forecasts can also help the home appliances industry to manage its inventory.

Other applications of this research

While we show promising predictions about the housing market using Google Trends, it can be also used in many other contexts to predict future economic activities, for example, the technology sectors. In particular, Google trends can be used to predict the outcome of the standards war between HD DVD and Blue-Ray. If user interests for Blue Ray grow overtime relatively to HD DVD, we may expect Blue Ray to win the standards war. Similarly, we can also use search frequency to predict the market share of an electronic product or an operating system such as Macintosh. Instead of paying a premium for industry reports, Google Trends can be used to predict if a particular technology would gain market shares.

Conclusions, Limitations and Future Work

Four hundred years ago, the microscope was invented. For the first time, scientists could see individual microbes in a drop of water and blood cells that traversed through the body. The result was a revolution in biology and medicine, including the germ theory of disease and ultimately, new vaccines, medicines and treatments. Today, due to advances in IT and IT research, we are gaining the capability to observe micro-behaviors online. Rather than rely on painstaking surveys and census data, predefined metrics and backward-looking financial reports, social science researchers can use query data to learn the intentions of buyers, sellers, employers, gamers, gardeners, lovers, travelers and all manner of other decision-makers even before they execute their decisions. It is possible to accurately predict what will happen in the market place days, weeks and even months in the future with this approach. Search technology has revolutionized many markets, and it is now revolutionizing our research.

This is an exploratory study investigating whether online search behavior from Google Search can predict underlying economic activity. Using housing sales data, we find evidence that search terms are correlated with sales volume and also with the house price index, lending credibility to the hypotheses that Web search can be used to predict future economic activity, for example when the economy may recover from the current recession. We are aware of the fact Google search queries do not represents all the online housing search activities. As some consumers may bypass the search engine all together and go directly to certain websites such as Realtor.org when considering buying and selling a home. Approach using Google search alone would miss this type of consumers. However, despite missing these consumers, we can still predict the housing sales and housing price using only online search captured by Google, demonstrating the power of online queries in forecasting economic trends.

Ultimately, micro data collected using Google Trends may prove one of the most powerful tools for helping consumers, businesses and government officials make accurate predictions about the future so that they can make effective and efficient decisions. It distills the collective intelligence and unfiltered intentions of millions of people and businesses at a point in their decision-making process that precedes actual transactions. Because search is generally not strategic, it provides honest signals of decision-makers intentions. The breadth of coverage, the level of disaggregation and the speed of its availability is a radical break from the majority of earlier social science data. Even simple models can thus be used to make predictions that matter. Of course, there are many obstacles yet to overcome and refinements to be made. For instance, paradoxically, as businesses and consumers come to rely on query data for their decision-making, as we expect they will, there will be incentives for opposing parties to try to

degrade the value of the data, perhaps by generating billions of false or misleading queries. This will in turn call for counter-measures and perhaps the golden age of simple models using these data will be brief. Nonetheless, as these data and methods become more widely used, we can only conclude that the future of prediction is far brighter than it was only a few years ago.

References

Appleton-Young, L., 2008, "State of the California Housing Market 2008-2009, California Association of Realtors

Aral, S., Brynjolfsson, E., & Van Alstyne, M. 2006. "Information, Technology and Information Worker Productivity: Task Level Evidence." Proceedings of the 27th Annual International Conference on Information Systems, Milwaukee, Wisconsin.

Calhoun, C.A, (1996) "OFHEO House Price Indexes: HPI Technical Description", OFHEO, http://www.fhfa.gov/webfiles/896/hpi_tech.pdf

Case, K.E. and Shiller, R.J. (1987). "Prices of Single Family Real Estate Prices," New England Economic Review. 45-56.

Case, K.E. and Shiller, R.J. (1989). "The Efficiency of the Market for Single-Family Homes," The American Economic Review. 79, 125-137.

Davenport, T. (2006) "Competing on Analytics" Harvard Business Review Article, Jan, 2006

Choi, H., Varian, H., 2009 "Predicting the Present with Google Trends, April 2009", http://www.google.com/insights/search/#

Horrigan, "The Internet and Consumer Choice", Pew Internet and American Life Project, May 2008. http://www.pewInternet.org/~/media/Files/Reports/2008/PIP_Consumer.Decisions.pdf.pdf

Glaeser & Gyourko, 2007, "Housing Dynamics", NBER Working Paper

Glaser, 2009 "Housing Prices in the Three Americas", http://economix.blogs.nytimes.com/2008/09/30/housing-prices-in-the-three-americas/

Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant, "Detecting influenza epidemics using search engine query data", Nature vol. 457, November 2008.

Han, L (2008) Hedging House Price Risk in the Presence of Lumpy Transaction Cost, *Journal of Urban Economics (64)* , February 2008, 270-287

Han, L., (2009) "The Effects of Price Uncertainty on Housing Demand: Empirical Evidence from the U.S. Markets", Working Paper

Krugman, P., "How Did Economists Get It So Wrong?", New York Times, September 2, 2009, http://www.nytimes.com/2009/09/06/magazine/06Economic-t.html?em

Kuruzovich, J., Viswanathan, S., Agarwal, R., Gosain, S. and Weitzman, S. 2008. "Marketspace or Marketplace? Online Information Search and Channel Outcomes in Auto Retailing," Information Systems Research 19:2, pp. 182-201.

Moe, W. W., and Fader, P. S. 2004. "Dynamic Conversion Behavior at E-Commerce Sites," Management Science 50:3, pp. 326-335.Pentland, A., "Honest Signals: How They Shape Our World", The MIT Press: London, 2008

National Association of Realtors, "Profile of Home Buyers and Sellers", 2009

Nielsen Report, 2008, http://www.polepositionmarketing.com/emp/august-2008-search-3/

New York Times Editorial, "Unemployment Rising", http://www.nytimes.com/2009/04/05/opinion/05sun1.html, April 4, 2009,

Pentland, A., "Honest Signals", MIT Press, 2008

Simon, Herbert A. "On the Behavioral and Rational Foundations of Economic Dynamics." Journal of Economic Behavior and Organizations, Vol. 5, (1984), pp. 35-66.

Wu, L., Waber, B., Aral, S., Brynjolfsson, E., & Pentland, A. ""Mining Face-to-Face Interaction Networks Using Sociometric Badges: Predicting Productivity in an IT Configuration Task", International Conference on Information Systems, Paris, France, December 14 – 17, 2008.