Using Social Media to Measure Labor Market Flows

Dolan Antenucci[a]
Michael Cafarella[b]
Margaret C. Levenstein[c]
Christopher Ré[d]
Matthew D. Shapiro[e]

University of Michigan

March 2014

Revised May 2, 2014

This project is part of the University of Michigan node of the NSF-Census Research Network (NCRN) and is supported by the National Science Foundation under Grant No. SES 1131500. We are grateful to Tomaz Cajner, Michael Elsby, Jeremy Nalewaik, Linda Tesar, Kenneth West, Justin Wolfers, and seminar participants at the NBER Summer Institute, Vanderbilt University, the University of Wisconsin, the University of Michigan, and the Federal Reserve Board for helpful comments and to Mark Fontana for excellent research assistance.

[a]Department of Electrical Engineering and Computer Science, University of Michigan
[b]Department of Electrical Engineering and Computer Science, University of Michigan
[c]Survey Research Center and Ross School of Business, University of Michigan
[d]Department of Computer Science, Stanford University
[e]Department of Economics and Survey Research Center, University of Michigan and NBER

Using Social Media to Measure Labor Market Flows

ABSTRACT

Social media enable promising new approaches to measuring economic activity and analyzing economic behavior at high frequency and in real time using information independent from standard survey and administrative sources. This paper uses data from Twitter to create indexes of job loss, job search, and job posting. Signals are derived by counting job-related phrases in Tweets such as "lost my job." The social media indexes are constructed from the principal components of these signals. The University of Michigan Social Media Job Loss Index tracks initial claims for unemployment insurance at medium and high frequencies and predicts 15 to 20 percent of the variance of the prediction error of the consensus forecast for initial claims. The social media index is shown to have a greater signal to noise ratio than initial claims in providing information about the true state of job loss in the economy and is, therefore, of value to policy makers in need of high frequency, real-time indicators. The social media indexes provide real-time indicators of events such as Hurricane Sandy and the 2013 government shutdown. Comparing the job loss index with the search and posting indexes indicates that the Beveridge Curve has been shifting inward since 2011.

The social media job loss prediction is posted on the web each week at http://econprediction.eecs.umich.edu/.

JEL: C81, C82, E24, J60

Dolan Antenucci
Department of Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109-2121
dol@umich.edu

Michael Cafarella
Department of Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109-2121
michjc@umich.edu

Margaret C. Levenstein
Ross School of Business and Survey Research Center
University of Michigan
Ann Arbor, MI 48106-1248
maggiel@umich.edu

Christopher Ré
Department of Computer Science
Stanford University
353 Serra Mall
Stanford, CA 94305-9025
chrismre@cs.stanford.edu

Matthew D. Shapiro
Department of Economics and Survey Research Center
University of Michigan
Ann Arbor, MI 48109-1248
and NBER
shapiro@umich.edu

This paper develops new measures of flows in the labor market using social media data. Specifically, we use Twitter data to produce and analyze new weekly estimates of job flows from July 2011 to early November 2013. We present methods for validating such novel economic measures and articulate principles for assessing the usefulness of time series derived from social media (Section I). We do this first by comparing our estimates with official data. Our Twitter-derived job loss index tracks initial claims for unemployment insurance (UI) and carries incremental information relative to both lagged UI data and the consensus forecast (Section II). We also propose social media indexes to measure concepts with weaker analogues in official statistics—job search and job posting—and then use these measures to study shifts in the relationship between posting and job loss (Section III). Finally, we show that our high-frequency social media index tracks events that affect the job market in real time, such as Hurricane Sandy and the federal government shutdown (Section IV). We show that the index has greater signal to noise ratio than the official initial claims series in providing information about the true state of job loss in the economy. It is, therefore, of potential value to policy makers in need of such indicators, particularly in formulating policies when there is uncertainty about whether the economy is at a turning point.

Social media provide an enormous amount of information that can be tapped to create measures that potentially serve as both substitutes and complements to traditional sources of data from surveys and administrative records. The use of social media to construct economic indicators has a number of potential benefits. First, social media data are available in real time and at very high frequency. Such timely and high-frequency data may be useful to policymakers and market participants who often need to make decisions prior to the availability of official indicators. The fine time-series resolution may be particularly helpful in identifying turning points in economic activity. Second, social media data are potentially a low-cost source of valuable information, in contrast to traditional surveys that are costly for both the respondent and the organization collecting the data. Third, social media offer a distinctive window into economic activity. They represent naturally-occurring personal communication among individuals about events in their everyday lives without reference to any particular economic concept. Like administrative data, but unlike surveys, social media challenge economists to map the observed information into the economic concept being measured. Fourth, social media can be used to answer questions we would have liked to ask in surveys had we known about events in advance. In ordinary survey design, we frame the questions and then collect the data. Social media allows us to reverse this order and generate *ex post* "surveys." For example, we use the indexes to examine the impact of two shocks to the labor market, Hurricane Sandy in October 2012 and the October 2013 government shutdown.

This paper implements social media indexes for job flows. Why do we focus on job flows? Substantively, job flows are of central interest to economists, market participants, and policymakers. Practically, the weekly frequency of the official UI claims data makes them a good benchmark for testing the performance of our social media measures. We have Twitter data for only 28 months, so there is insufficient time-series variation against which to compare national aggregates such as GDP or employment. Given that the UI series is available at high frequency and without sampling error, one might ask what the Twitter signal has to add. We chose the unemployment flows concept as the case study for this paper precisely because of the availability of a high quality, frequent series against which to compare it. This comparison should give researchers confidence to use the techniques developed in this research to study domains that are not as well-covered by official statistics.

Official UI data and our job loss index track related but not identical phenomena. Our aim is therefore to track the official index with our social media index, but not perfectly so. Since they are designed to measure the same general economic concept, they should certainly have strong co-movements. Yet they should not be perfectly correlated because of differences in population, timing, and the underlying data generation process. Indeed, one of the promises of social media for measuring economic concepts is that it will provide incremental information relative to official statistics. We find that the social media index not only does a very good job of tracking the official data, it also has important independent movements that we show—both statistically and anecdotally—carry incremental information.

I. Twitter Data

Twitter is a social media service through which individuals and enterprises can post short, 140-character messages of any subject of their choosing. These messages are known as Tweets.[1] Unless restricted by the user, they are available publicly. These messages can be read on the web through internet browsers and by a variety of other software. Individuals can subscribe to the Tweets of particular users, or subscribe to them by topic (denoted by a hash tag, i.e., a keyword with a "#" prefix). A common use of Twitter is to communicate news about life events to a community of friends. These can be mundane ("I am standing at 3rd and Elm waiting for a bus"), communicate plans or whereabouts ("Let's meet at Showcase Cinema at 7:15 to see the new Bond film"), or momentous ("George and I are pleased to announce the birth of Polly, 7lb, 8oz"). The following Tweet contains a job loss phrase of the type we analyze.

[1] The length of a Tweet derives from the 160 character limit on an SMS text message. Twitter reserves 20 characters for the identifier.

> 2011 was interesting. I ended an engagement, got laid off, started a small biz, and it looks like I'll be moving this year too. Whew!

Our analysis is based on a roughly 10 percent sample of all Tweets between July 2011 and early November 2013. The dataset contains 19.3 billion Tweets and is 43.8 terabytes (TB) in size.

Web search queries, an alternative source of naturally-occurring web data, have also been examined for their economic content. Web search queries are framed very differently from social media data and contain different types of information. Our approach based on social media thus is both similar and complementary to approaches based on web searches (Choi and Varian 2009a, b, Scott and Varian 2013). There are several differences in technique between our system and that of Choi and Varian. First, web search queries and social media data likely capture different kinds of information. Social media data capture communications among individuals about their lives, while web search queries reflect individuals trying to find information on the web. These datasets capture phenomena that likely overlap, but are not identical. For example, an individual may be likelier to announce a new job to friends via social media, but may be more willing to reveal personal health information via a web search query. Second, Twitter messages are tied to a user, who exists in a public social network. User meta-data can be used, for example, to classify messages by demographic groups or geography. Potentially, user information could be used to relate Twitter messages across users. As of this writing, Google Trends does not give information about the person who generated the search query, so controlling for individual characteristics or following a user over time is not possible. (The current version of Google Trends does give country-level geographic distributions of the users who generated the search queries.) Last, but not least, raw Google search queries are not public and so cannot be analyzed directly; in contrast, Tweets are public. Google's web search query data are made public only via the Google Trends tool. It currently does not reveal actual frequencies for search terms, but instead places frequencies on a 0-100 scale, making some uses of the data difficult or impossible. The techniques used to collect and prepare the data are not public. In contrast, the methods we propose here are transparent, so researchers can more easily inspect and reproduce analyses performed on the resulting signal data.

Use of social media to construct time-series indicators has recently come under criticism along several dimensions (see Lazer et al., 2014). These criticisms include lack of transparency of the statistical procedures, lack of replicability because of the use of proprietary data, and instability of time series. Specifically, the well-known Google Trends index for the flu suffers from these problems. As we will discuss, our Twitter based approach was designed not to have these issues. We explicitly describe the statistical procedure. We post the signals used to construct the index on the Web. Our index is the result of a collaboration between economists and computer scientists focused on the underlying economic phenomena. Because we select signals from Twitter completely independently from series we are trying to track, our procedure is immune from the overfitting problem raised by Lazer et al.

A. Strategies for Converting Social Media into Data

A core challenge in this work is to develop a rigorous methodology to convert the corpus of social media texts into time-varying signals that have both predictive and explanatory power for labor market flows. We can convert a given set of relevant Tweets into a signal by first counting their frequency in each 24-hour period in the sample period and then compiling these daily counts into a single time-varying signal. Obtaining *m* signals amounts to choosing *m* relevant sets of Tweets. Clearly, the power of our social media index depends on how well we choose these sets.

Given the very large number of Tweets, automated statistical techniques for choosing predictive features—in our case, sets of Tweets—are appealing (Guyon and Elisseeff, 2003). These techniques, however, pose serious challenges for our task. First and most basic, the technique for enumerating all the possible signals in the Tweet collection—i.e., modeling the feature space—is not obvious. One approach is to create a Tweet set for each unique Twitter author; another is to compose a Tweet set for every *k-gram*, in which a *k-gram* is a sequence of *k* or fewer consecutive ordered words found in the Tweet corpus. Considering a restrictive set of features may make feature selection easier, while a larger set of features may enable creation of an index with better predictive power. Second, even a relatively restrictive set of potential features, such as *k-grams* with $k \leq 4$, yields vastly more features than we have time-series macroeconomic data.[2] (There are roughly one billion *4-grams* that occur at least 20 times in our data.) Some features with high correlations will in fact be entirely spurious and thus carry no predictive power. Other features may be predictive but not causal (e.g., Lysol as a feature for flu). Such features can be useful for the predictive model but may be logically opaque to human observers (that is, the Tweets in a set will not have an obvious common thread or will have a common thread that appears to be nonsensical).

The computer science members of the research team are exploring the problem of feature selection when applied to social media and any macroeconomic or similar topic.

[2] In their development of a flu index using Google web search queries, Ginsberg, *et al.* (2009) followed a variable ranking strategy that chose signals that were highly correlated with a target signal. In preliminary work, we considered the correlation of phrases found in Tweets with weekly unemployment. None of the top 100 most correlated phrases had any plausible connection to unemployment (Antenucci *et al.* 2013a).

Macroeconomic tasks tend to offer very small in-sample datasets in comparison to other data-intensive *trained system* tasks in computer science. For example, web search engines can exploit billions of human judgments about web page relevance, derived by observing users' clicks on search engine result pages. In the absence of large datasets that can automatically validate feature selections, the techniques under development would have the researcher describe feature preferences (that is, provide "domain knowledge") and then observe a set of features suggested by the system. The researcher would then reject features that violate real but implicit researcher preferences. (We give several examples of this procedure from our own experience in the section below.) One early version of this automated system offered suggestions based on a combination of user-suggested terms, thesauri, and statistics derived from web text (Antenucci et al. 2013a,b). Ideally, this system would give results in interactive timeframes, but the massive number of possible feature combinations (roughly 2.7 x 10^{103} when choosing 10 features from among *4-grams* in our corpus) makes known suggestion techniques infeasible. Solving these problems is the subject of ongoing research, and could substantially lower the researcher burdens associated with applying social media techniques to any novel topic.

In this paper, we solve this problem by first limiting our analysis to *k-grams*, which are essentially repeated cross sections of Tweets, aggregated first to days and then weeks. For the economics task at hand, we chose signals from a feature space in which each feature corresponds to a *k-gram* where $k \leq 4$. Second, we narrow the feature space further by using domain knowledge to select signals that we strongly believe are causally connected to job loss. Specifically, the research team identified terms that it believes are indicative of the phenomenon being measured, based on knowledge and expertise in the area. We describe this procedure in more detail in the next section.

This approach has drawbacks. First, we may unwittingly add bias during our selection of phrases. Second, some Tweet sets cannot be described at all (e.g., because we restrict ourselves to *4-grams*, we cannot characterize the set of all Tweets that contain the five-word phrase, “my Mom no longer works”). Finally, the feature space does not automatically group phrases that are textually distinct but semantically similar (e.g., “I got fired” and “my boss canned me” express the same idea but are not identical *k-grams*). Our choice of feature modeling has the benefits of being easy to describe and enabling many Tweet sets that are understandable to the user (e.g., all Tweets that contain the phrase, “I lost my job”). Moreover, despite its restrictiveness, our design is sufficient to demonstrate that social media data contain genuinely useful information about labor flows.

B. Implementation

To implement the domain knowledge strategy, the research team developed a list of phrases related to job loss and unemployment that it expected to be found in Tweets that carried information about job loss in general and initial claims for unemployment insurance in particular. The phrases we use to aggregate signals of job loss and unemployment are listed in Table 1. A space is denoted as “|” and a wildcard as “*” as in the detailed descriptions in Appendix Table 1. The process for generating the list of phrases includes the following steps:

- *A priori* specification of terms such as “lost job,” “laid off,” and “unemployment” that we expect to be contained in Tweets of interest.
- Expansion of the specification of the target phrases to include plausible misspellings and wildcards to capture variants such as “lost my job” or “lost his job.”
- Deletion of phrases where—upon inspection—it becomes clear that the *a priori* specified phrases have little to do with the labor market.

The first column of Table 1 gives the ten job loss and unemployment signals that we will analyze. We allow for variants in spelling and spacing. The variants we consider include the 27 search phrases listed in Appendix Table 1.

There are some terms one might expect to include *a priori,* but which we exclude or include only in combination with other words. For example, we do not include a search for the words "fired," "benefits," and "insurance" alone because each was used much more frequently in unrelated contexts (e.g., fired up). Note that "unemployment benefits" or "unemployment insurance" are captured because we do include any *k-gram* including the word "unemployment." In general, singleton terms can be problematic. We originally included the term "sacked" but eliminated the signal from further analysis because its frequency in the data—several orders of magnitude greater than other employment-related terms—suggested that its use referred to other linguistic meanings. Similarly, we eliminated "let go" because it appeared much more frequently than other employment-related phrases and seemed to have other plausible meanings.

In the case of the phrase "lost * work," inspection of the matched *k-grams* clearly indicated nearly universal non-employment related concepts. Many phrases referred to computer problems such as "lost all my work" and "lost my #$% work," as well as happier references such as "lost in my work" and "lost Beethoven work." As a consequence, we excluded all candidates related to this signal in the creation of the job loss measure. Having the wildcard in this search was critical for revealing that the "lost work" phrases were not about employment.

One concern about the use of social media to measure economic activity is that it will capture comments on releases of official statistics rather than provide independent measures of activity.[3] We did not see evidence that there is a lot of Tweeting about the Department of Labor's release of initial claims data, but the monthly Employment Situation release does get a lot of attention and might account for a significant number of mentions of "unemployment." Indeed, the Bureau of Labor Statistics (BLS) plans to use Twitter as an official release channel, so re-Tweets of the unemployment report may be a significant confound in the future. To check for the importance of the unemployment report *per se* on Tweets about unemployment, we estimate a linear regression with the unemployment signal as the dependent variable and a dummy for weeks containing the unemployment report as the regressor. The estimated relationship is

$$\underset{}{r\_unemp_t} = \underset{(1.9)}{49.4} + \underset{(4.1)}{17.4}\, emp\_sit_t + u_t$$

where *r_unemp* is the unemployment signal, *emp_sit* is the Employment Situation dummy, and *u* is the residual.[4] Tweets about unemployment are about a third higher in an Employment Situation week than average, so we purge Tweet-derived signals containing "unemployment" of the Employment Situation effect using a regression as shown above.

Table 1 presents summary statistics of the signals. The signals are expressed as weekly rates per million Tweets. While the signals derived from the selected phrases are fairly rare—between 0.5 and 54 per million Tweets—there are so many Tweets that the signals still provide a rich dataset. Of the 19.3 billion Tweets reflected in Table 1, there are 2.4 million associated with job loss and unemployment. The signals have roughly comparable coefficients of variation, so there is potentially information in each of them. The correlation matrix in Table 2 shows that the

[3] There is evidence that a substantial amount of communication over social media consists of links to internet sites of content creators (CNN, Justin Bieber's Tweets, etc.). See Goel, Watts, and Goldstein (2012).

[4] Standard errors are in parentheses.

signals from the selected phrases are positively correlated (with the exception of "pink slip"), so they do appear to be picking up related phenomena in the Tweets.

In order to preserve degrees of freedom while extracting as much information as possible from the Twitter signals, we perform a principal components analysis on the ten signals. Table 3 reports the factor loadings and variances. Not surprisingly, given the positive and fairly uniform correlation structure reported in Table 2, the first factor has fairly uniform coefficients across the signals and accounts for 43 percent of the variance. The next four factors each account for about 10 percent of the variance. Figure 1 plots the factors estimated over the entire sample from July 2011 through early November 2013. The first panel shows only the first factor, and the second panel shows the first four. The first factor is fairly volatile in the second half of 2011 and has a noticeable downward trend into 2012, when it flattens. This pattern is interrupted in late 2012. In 2013, the downward trend evident throughout the period resumes. The next panel adds the factors 2, 3, and 4. By construction, they have signal less and have spikes that might be suspect. Factor 2 has a spike in late 2011 that matches a spike in Factor 1, so it might be genuinely related to job flows.

On the other hand, note that Factor 3 is dominated by the "pink slip" signal (see Table 3). There is evidently a spike in that signal in December 2012. Without it, Factor 3 would not have emerged from the principal components analysis as having significant variance, so unless we are prepared to believe this spike is job related, it should be discounted.[5]

[5] We suspect that the spike was driven by Tweets about the November 19, 2012 launch of a marketing campaign titled "Pink Slip" featuring football player Tom Brady. See Business Wire, November 19, 2012 "UGG for Men Launches New 'Pink Slip' Integrated Campaign for Holiday 2012 Featuring Tom Brady."

C. Relating Social Media Data to the Economy and Economic Data

These signals from social media and the factors that summarize them are new measures of economic activity. They are not based in any way on standard measures using conventional sources of data. It is natural to ask how they relate to a standard measure of economic activity: initial claims for unemployment insurance (UI). The initial claims data are well-suited for evaluating the social media signals. First, they are available at weekly frequency. Given that we have just over two years of Twitter data, a high-frequency economic indicator for comparison is very important. Second, initial claims for UI are a direct measure of transitions in the labor market. Hence, they are likely to have much more high-frequency variation than variables that measure stocks (e.g., the unemployment rate). We expect that social media data will be useful precisely for measuring such high-frequency changes in activity.

Figure 2 shows initial claims for UI (left scale) and the first factor from the Twitter job loss and unemployment signals (right scale). *The social media series is estimated completely independently from the new claims data.* The relationship between these two indicators of job loss is quite strong—both in the general trend and in some notable spikes. Over the sample period from July 2011 to early November 2013, initial claims have a general downward trend in new claims. They flatten in the first half of 2012 and then resume the downward trend in 2013. The social media series has a very similar pattern.

There are also some high-frequency changes in new claims—notably the spike in late fall 2012. Our indicator also captures that spike in job loss. We will investigate this spike, associated with Hurricane Sandy, in some detail below.

Note also that the fit of the social media series to the new claims series is not perfect. Aside from period-by-period variation, the social media series has a spike in 2011 that is not in the initial claims data. More interestingly, it does not indicate the slowdown in job loss seen in the new claims data in September 2013. The social media information contains independent information about the job market. Indeed, as we will discuss below, the drop in initial claims in September relates to a processing problem in California.

Not all job loss is associated with applications for UI, so we are not seeking simply to predict UI. Nonetheless, the high- and low-frequency association of the series with the official data is reassuring.

We can test the association of the social media signal with the UI initial claims data statistically. Table 4 presents regressions of initial claims on the social media series and for comparison, lagged initial claims and the consensus forecast.[6,7] The social media series is not as good a predictor of new claims as are the lagged dependent variable or the consensus, though of course there is no reason to expect or hope it to be. Nonetheless, it is strongly predictive of new claims and remains significant in the regressions that include the lagged dependent variable and the consensus.

II. A Real-time Predictor from Social Media Data

The social media series for job loss successfully tracks official data at both high and low frequency. This section constructs a real-time index for predicting initial claims for unemployment insurance, and evaluates its ability to provide a real-time indicator of economic activity. In contrast with the previous section, which sought to estimate time series from social

[6] The consensus forecast is produced by Bloomberg Surveys (various dates). It is the median forecast of a panel of approximately 50 economists.

[7] For this set of regressions, the social media index is normalized to make the regression coefficients comparable, normalizing so it has the same mean and standard deviation as the dependent variable. This normalization, of course, has no effect on the *t*-statistics or fit of the regression.

media and show that they are related to economic activity, this section aims to construct a predictor that is feasible in real-time.

A. Constructing the Real-time Predictor

To construct the University of Michigan Social Media Job Loss Index, we estimate a model relating initial claims for unemployment insurance to social media signals recursively, using only data that are available at the point of the prediction. The Twitter data are available almost immediately, so we can construct a prediction of the current week's new claims with virtually no lag. The procedure is as follows:

1. Estimate the factors on the social media signals from the beginning of the sample through the current week.
2. Estimate the University of Michigan Social Media Job Loss Index by regressing real-time initial claims data on the factors. The regression coefficients are updated each week.
3. Construct the prediction as the fitted value for the current week from that regression.
4. Update the data weekly and repeat this procedure.

Precise details of the procedure are provided in the appendix.

We carry out this procedure recursively over periods ending July 7, 2012 through November 2, 2013. The starting period of the estimation is always July 16, 2011.[8] We consider various specifications for the regression in step 2. Table 5 reports the estimates of these specifications for the final period. Table 6 reports the root mean squared error of these different specifications using the predictions estimated recursively. The specification with a constant and

[8] We experimented with various alternatives to having a fixed starting period. These included estimating over a rolling, one-year window and using the whole period, but with exponentially declining weights on older observations. The results were quite similar, so we report the simpler specification using OLS estimated over all the data available in real time.

the first factor yields a strongly significant coefficient and an adjusted R^2 of 59 percent.[9] Adding factors 2 through 4 adds little to the fit of the regression. Table 6 shows that the RMSE of the specification with one factor is the lowest, so that is our preferred specification based both on goodness of fit and parsimony.

Table 5 also includes specifications with two additional explanatory variables. Specification 5 includes the seasonal factor for initial claims as an additional explanatory variable to evaluate whether there is discernible seasonality in the relationship between the social media index and initial claims. Flows into unemployment are highly seasonal with peaks in December/January and the summer. The Twitter data may also exhibit seasonality, but with less than three years of data, we cannot seasonally adjust it. Using the new claims seasonal factor implicitly seasonally adjusts, assuming the same seasonality in both series. The seasonal factor is small and insignificant, so we do not include it in our preferred specification. Given that job loss is indeed seasonal, it is interesting to note that the social media mentions of job loss do not have the same spike as the official data. An interpretation of this finding is that a predictable job transition relating, for example, to the end of a seasonal spell of employment, is not something that one would mention in a Tweet using the phrases we use to construct the signals. The absence of such predictable transitions is not necessarily a problem for the social media index—indeed for some purposes it might be an advantage—but it needs to be kept in mind for the use and interpretation of the indicator.

The last column of Table 5 considers whether the announcement of the Bureau of Labor Statistics unemployment data affects the index. As we describe in Section I, the signals mentioning “unemployment” are already purged of this announcement effect. The estimate in

[9] Note that the estimate in the first column of Table 5 is the same, apart from normalization, of that in the third column of Table 4B.

column (7) checks whether this processing is sufficient for removing the announcement effect from the index. The dummy for weeks that the unemployment rate is released is insignificant, so the procedure discussed in Section I does appear to suffice.

B. Analyzing the Real-Time Social Media Job Loss Index

Figure 3 shows our preferred specification for the social media index with a constant and Factor 1. It is plotted against the initial claims data. The shaded area is the first year of data. Because it is not feasible to estimate the factors and perform the regression recursively, as described in steps 1 and 2 above, they are estimated over the whole period. The balance of the data shown in Figure 3 is estimated recursively, as described in the previous section. The social media index tracks the official data closely, both in overall trend and in some of the movements. On the other hand, it carries independent information about job loss, for example, indicating a spike in 2011 not present in the official data and failing to show the decline in job loss in September 2013 at the end of our sample. This drop in reported initial claims in the official data in September 2013 relates to a data processing issue in California;[10] this is an example of where the social media index does not suffer from measurement error encountered by the official data, and thus may more closely track the true state of the economy.

Additionally, the social media index tracks increases in job loss evidently associated with the government shutdown during the first two weeks of October 2013. The index rises noticeably in the first half of October and declines by about the same amount in the second half

[10] Employment and Training Administration's Unemployment Insurance Weekly Claims Report, issued September 19, 2013, reports a decrease of 25,412 UI claims in California "due to Labor Day holiday and computer system updates." http://www.dol.gov/opa/media/press/eta/ui/eta20131889.htm The following week the Employment and Training Administration's Unemployment Insurance Weekly Claims Report, issued September 26, 2013, reported that a comparable increase in UI claims in California "reflects return to 5 day workweek and a full week of processing after computer system updates." See http://www.dol.gov/opa/media/press/eta/ui/eta20131953.htm

of the month. Initial claims have a similar pattern (after accounting for the rebound from the resolution of the processing issues in California).[11]

While the UI series and social media index generally move together, they are certainly not perfectly correlated. This is to be expected, since they measure different things. While part of the proof of concept is to show that the social media index moves with the official data, the aim is not to replicate the official data perfectly. For myriad reasons relating to the concept being measured, the coverage and take-up of unemployment insurance benefits, and the makeup of the samples, the social media index measures something different from the official series. Nonetheless, our findings that they are related do provide evidence that the social media index is a meaningful measure of economic activity.

C. Assessing the Information in the Real-Time Social Media Job Loss Index

Next we ask whether, from the perspective of predicting the state of the economy in real-time, there is incremental information in our social media index. The results in the previous section suggest this might be so. We know from column 6 of Table 4 that the consensus forecast is a very good predictor of the initial claims data, but that the social media factor has incremental explanatory power. In order to address the question of incremental information, we compare our Social Media Job Loss Index to the consensus forecast on the eve of the initial claims announcement. This consensus forecast is based on a survey of market experts several days prior to the release of initial claims for UI. Table 7 reports the results of this analysis. First, we examine the preliminary report of new UI claims. We subtract the consensus estimate from the preliminary UI claims report to calculate the error in the consensus view. We then compare these

[11] Federal workers apply to a different unemployment insurance system. They are not included in the preliminary initial claims data used to construct the real-time job loss index.

errors to the Social Media Job Loss Index, which we construct based on information available in real time as described above.

To assess the incremental information in the Social Media Job Loss Index, we examine the regression of the error (preliminary initial claims minus consensus) on the Social Media Job Loss Index minus the consensus (Table 7, Column 1). The social media index carries incremental information. It is statistically significant and explains about 15 percent of the variation in the surprise relative to the consensus. In Table 7, Column 2 we report an estimate that separates the impact of the University of Michigan Social Media Job Loss Index and the consensus. The University of Michigan Social Media Job Loss Index remains a significant predictor of the error in the consensus, while the coefficient on consensus itself is roughly equal and in opposite sign to the coefficient on the University of Michigan Social Media Job Loss Index minus consensus in the first estimate (the *p*-value of the test of equal and opposite coefficients is 0.16). These results are included to show that the correlation of the consensus with the surprise is not driving the result. Finally, in Column 3 we include a lagged index. It has a very small coefficient that is not significantly different from zero, suggesting that, after one week the information content in the Tweets had been incorporated into the consensus view. Indeed, there is little evidence of any lags in the relationship between the social media signals and the UI data.

In the second part of Table 7, we compare our social media index to the UI claims, revised one week after the initial numbers. The results are similar to those for the preliminary UI claims, except that the explanatory power of the social media index increases to 19 percent of the variance, suggesting that the social media index is better at predicting the true, revised UI number than it is at predicting the original estimate. This finding suggests that the social media index is capturing information about the true state of the job market that is not captured in either the consensus or the preliminary UI claims estimate. The incremental information in the social media index is relevant relative to both the preliminary and revised data. Policymakers and forecasters will be more interested in information about the revised data. Market participants may be more interested in the incremental information for the preliminary announcement.

Figure 4 shows the incremental information in the social media index on a week-by-week basis. The figures show the surprise (initial claims minus consensus) and the part predicted by the University of Michigan Social Media Job Loss Index (that is, fitted value of the regression of the surprise on the University of Michigan Social Media Job Loss Index minus consensus) for the preliminary and revised initial claims data. Again, we do not aim to track all the surprise and indeed account for 15 to 20 percent of it. Much of the surprise is serially uncorrelated noise with no intrinsic interest. The social media index does capture some of the unexpected increase in initial claims in early 2012 and some of the swing from positive to negative surprises in the last quarter of 2012 to the first quarter of 2013.[12]

D. Providing a Real-Time Economic Indicator from Social Media

This research project has implemented the creation of the University of Michigan Social Media Job Loss Index in real time. At the end of each week ending Saturday, our automated computer program processes the latest Tweets, recalculates the job loss index based on the one-factor model described in the previous section, and updates the prediction. We post the prediction on the web each week. We also post the underlying daily signals used to construct the index.[13] In

[12] Choi and Varian (2009b) use Google search queries to predict initial claims. For "Welfare and Unemployment" (though less so for "Jobs") Google Trends captures the increase in unemployment at the onset of the Great Recession.

[13] See http://econprediction.eecs.umich.edu/ for the predictions and http://ringtail.eecs.umich.edu/unempSignalsDailyUpdate.csv for the daily signals. Making the

this way, we are able to provide policymakers, forecasters, and other interested parties with a useful high-frequency economic indicator with virtually no lag between availability of the source data and availability of the indicator. Such virtually contemporaneous information should be useful to policymakers and market participants who need to make decisions in real time with incomplete information.

E. Demographics

A concern about the use of social media data is that those who participate in social media are not representative of the population.[14] We can assess this concern by estimating demographic characteristics of Tweeters. For a subset of signals, we can probabilistically estimate the age and sex of the sender based on attributes of the Tweet. By examining the distribution of word choices in a set of Tweets written by people with known age and gender, we can train statistical models to predict age and gender for the author of a novel Tweet. The training set for the age predictor includes up to 3,200 Tweets for each of 24,000 users, while the training set for gender includes 12,500 users' Tweets. We identify users with known age and gender by searching for Tweets that contain self-admission of demographic details, for example, "I'm 30 years old now, but still live with my mom" or "I'm a strong woman." The statistical technique we employ is a randomized decision tree classifier (Breiman 2001).

We use six age brackets: 14-18, 19-21, 22-24, 25-34, 35-44, and 45-64. Classifier accuracy on held-out data is 47.3 percent for age, and 82.4 percent for gender. Appendix Table 3

inputs into our index available addresses concerns about transparency and replicability of indexes from social media raised by Lazer et al. (2014). These inputs are raw counts of Tweets as described in the text. We cannot redistribute individual Tweets.

[14] See, for example, Mislove et al. (2011). Note that the authors use a sample of Twitter data that ends in 2009, arguably well before Twitter became a mass phenomenon. Consequently, though their method is of interest, their demographic findings represent early adopters who are likely less representative than the current population of Twitter users.

shows the fraction of job-related signals by age and sex for the subset of signals for which we can estimate demographic characteristics.[15] Though the distribution of age and sex does not match the population, the use of social media to communicate about job-related issues is much more evenly spread across demographic groups than one might have expected. In particular, middle-aged and older individuals are over-represented in the job-related signals, in comparison to how frequently they Tweet overall. Note that senders of a signal need not be Tweeting about themselves: for example, messages by a teenager could be commenting on a job transition for a parent. Even so, it is reassuring that a substantial majority of our job-related signals are from the working-age population.

III. Job Search and Job Posting

The previous section focuses on job loss—a concept that has a good high-frequency analogue in official data. In this section we broaden the range of labor market flows to include job search and posting. The first section constructs indexes and the following section uses them to study shifts in the Beveridge curve over the recovery from the Great Recession.

A. Indexes

We create and describe two additional series related to search, matching, and labor market equilibrium. Specifically, we examine Tweets containing phrases indicating that the Tweeter is searching for a job (e.g., "find," "look," "need," "search," or "seek," each followed by "job" or "work") and others that suggest that the Tweeter is searching for an employee ("hiring," and "job" or "work" followed by "opportunity" or a phrase indicating location or job type). The

[15] The estimates are based on data only through June 2013. After that time, Twitter changed its public API, thereby reducing our ability to gather large numbers of Tweets for specific individuals. Further work is required to quantify how this change in data availability will affect the accuracy of demographic classification, and if accuracy is reduced, what novel methods can be used to improve quality.

signals for job search and job posting are listed in Table 8 and the detailed phrases are given in Appendix Table 2.

Signals reflecting a job posting are much more frequent than those reflecting job search. Search signals are comparable in their frequency to those reflecting job loss (compare to Table 1). Table 9 presents the correlation matrix of all the job search and job posting terms. While "find" is not closely correlated with any other terms, the other search terms are positively and similarly correlated. As expected, the posting terms are more closely correlated to one another than to the search terms. The "seek" term is correlated across search and posting terms, and is syntactically related to both, so it is included in both sets of terms.

There are analogues to the Twitter signals for search and posting in conventional data sources:

- The unemployment rate is a measure of search activity, especially since the BLS requires a modicum of job search activity as part of the CPS definition of being unemployed.
- Help wanted advertising has been a traditional source of data on vacancies.[16]
- The BLS JOLTS data provide a survey measure of job openings.[17]

For the job loss index developed in the previous section, UI claims were a particularly good analogue in official data. New claims for UI are high frequency. Moreover, both the "lost job" Tweets and the new claims data are direct measures of flows. In contrast, the match of the

[16] The Conference Board formerly produced a monthly Help Wanted Advertising Index based on print advertisements, but discontinued it in 2008. It currently produces a comparable series, drawn from internet postings of job advertisements, the Conference Board Help Wanted OnLine data series http://www.conference-board.org/data/helpwantedonline.cfm

[17] In the JOLTS, an opening needs to meet three criteria: A specific position exists; work could start within 30 days; and the firm is actively seeking workers from outside its location to fill the position. See Bureau of Labor Statistics, "Job Openings and Labor Turnover Report" (2013). The JOLTS data also have data on separations that can be compared to our job loss index. The social media index is more frequent and more timely than the JOLTS data. The JOLTS data are produced monthly and are available about two months after the reference week of the survey.

search and posting terms with the unemployment rate and vacancies is not as clear-cut. First, it is not obvious which side of the market is generating the search and posting terms. Second, unlike in the job loss analysis, there may be a mismatch between stocks and flows when comparing the social media signals to unemployment or vacancies. There has been imaginative recent work on addressing the measurement issues relating to stocks versus flows in the labor market (see Davis, Faberman, and Haltiwanger (2013) and Barnichon et al. (2012)). Given the less than perfect analogy between our search and matching measures and potential official benchmarks, we do not pursue an econometric analysis along the lines of the previous section. We do, however, discuss our series in the context of recent findings from the JOLTS.

To construct the search and posting indexes, we do a factor analysis as discussed in the previous section. Table 10 presents the factor loadings for the first factor for the search and posting signals. As expected from the correlation matrix, the "find" signal has a smaller loading. Aside from the "find" signal in the first set of loadings and the "hiring" signal in the second, the loadings are fairly even across the signals. Figure 5 shows the indexes based on the first factors reported in Table 10. Since we are not benchmarking against an official index, there is no re-normalization of this index. Hence, the index is not measured in meaningful units: it is the change in the index that has meaning. We use the same recursive procedure as before: after the starting period shaded in gray, the indexes are estimated recursively.

Overall, there is a downward trend similar to that found in the job loss index. There are also spikes at various points, for example, December 2012, but not the previous December. The dip in October 2013 associated with the government shutdown is discussed in Section D.

Note that the downward trend in our posting and search indexes is not seen in the JOLTS job openings rate over the same period. The JOLTS job opening rate was 2.4 percent in July

2011 (the start of our sample). In 2012 and 2013, it was higher than in 2011, but with no discernible trend. It reached 2.8 percent in early 2012 and—with small ups and downs—was still at 2.8 in late 2013 (see Bureau of Labor Statistics 2014, Chart 1). In contrast, our search and posting indexes have a substantial movement down in the second half of 2011 and a slight downward trend in 2012 and 2013. These differences suggest that the social media indexes are measuring something different from the JOLTS openings, perhaps because of stock/flow considerations.[18]

Demographics of search and posting. Note that the demographics of search and posting Tweets are quite different. The modal group sending the job search Tweets are teenagers with the balance of search Tweets evenly spread across age groups. On the other hand, the majority of the job posting Tweets is from the 45 and older age group. This age pattern provides external support for the validity of our coding of the search and posting signals in that one would expect those posting jobs to be substantially older.

B. Beveridge Curves

We use our labor market indexes to study the relationship between job loss, search, and posting akin to the Beveridge Curve. Figure 6A shows the Search/Job Loss Beveridge Curve and Figure 6B show the Posting/Job Loss Beveridge Curve. Of course, these indicators are not identical to vacancies and unemployment in the standard Beveridge Curve, but they have potential to shed light on labor market equilibrium. Note that the relationship between the variables is mainly

[18] The Conference Board HWOL series has a similar flattening in 2013 after a recovery from the 2009 trough. The JOLTS separation rate (see Chart 2 of BLS (2014)) have a similar pattern to the job openings rate: moving slightly up from 3.2 percent to 3.4 percent from mid-2011 to the beginning of 2012, then bouncing between 3.1 and 3.4 in 2012 and 2013 with no discernible trend. Note that we are comparing our series to JOLTS for the period where we have data, beginning in 2011. The JOLTS openings rate has a strong upward movement from its trough in 2009 at the depth of the Great Recession.

positive, especially in the posting/job loss figure that is most analogous to the traditional curve. Taken at face value, this finding suggests that over this period (July 2011 to early November 2013), inward shifts of the Beveridge Curve dominated movement along the Beveridge Curve. This finding is consistent with recent work by Barnichon, et al. (2012) and Hobijn and Şahin (2013) that shows that there were significant outward shifts of the Beveridge Curve, as measured by JOLTS data, with the onset of the Great Recession.[19] There are various explanations of the outward shift in the Beveridge Curve that started at the onset of the Great Recession relating to deterioration in matching jobs to the unemployed, especially those unemployed for long durations. Our social media indexes suggest a reversal of this deterioration of labor market conditions at least since mid-2011.[20,21]

[19] In Section A, we noted that the JOLTS openings and separation rates both shift up from 2011 to 2012 and then exhibit no trend. In contrast, as discussed in the previous sub-section, our job loss and search and posting indexes both have opposite shifts over the same period. The work cited concerning the JOLTS focuses on the bigger shifts in the Beveridge curve surrounding the 2009 trough. Unfortunately, our Twitter data does not encompass the Great Recession.

[20] In a related application, Baker and Fradkin (2013) use Google Trends search data to construct an index measuring searching for a job. They link this index at the MSA level to administrative data on unemployment insurance claimants in Texas. They find that job-search related use of Google search increases across time and localities as the fraction of claimants approach the expiration of their benefits. Garthwaite, Gross, and Notowidigdo (2013) use Google Trends for "job openings" to show that search activity increased when beneficiaries lost their eligibility for TennCare (Medicaid) coverage.

[21] An obvious application of these data is to test for changes in search behavior surrounding the expiration of unemployment benefits. We have investigated this possibility. There is only a slight and not statistically significant increase in our social media search index after the expiration of extended benefits in late 2013. Note however that the expiration of extended benefits—though a large subject for political debate and news coverage—is not discernible in the official national, continuing claims data. Indeed, the official index of continuing claims *increases* after the termination of long-term benefits. (There is sharp January seasonality in this series that makes analysis of policy changes around the first of the year quite difficult.) The aggregate impact of the policy is small relative to other effects on labor market flows, such as seasonality, so it is not surprising that we cannot find effects in the social media index.

IV. Using Social Media Indicators as High-Frequency Indicators

In this section we assess the usefulness of the social media indexes as economic indicators for business conditions and policy analysis. The social media indexes provide contemporaneous indications of the effects of shocks that potentially disrupt the labor market. Using a signal extraction technique, we show that the social media index of job loss has superior measurement properties relative to new claims.

A. Labor Market and Hurricane Sandy

One of the potential benefits of analyses using social media data is that the researcher may examine the impact of unexpected events as they happen without relying on recall or chance surveying during such events.[22] We examine signals related to Hurricane Sandy to demonstrate this type of analysis. Figure 7 shows all Tweets (measured in thousands) that include the words "Sandy" or "hurricane." Unlike our previous analysis, carried out at weekly frequency, Figure 7 shows daily data. Additionally, we simply present raw counts rather than rates or a statistical index, because we have no historical baseline.[23] Not surprisingly, the number of such Tweets increases sharply as Hurricane Sandy headed toward the northeast coast of the United States in late October 2012. The series peaks on October 29 when Hurricane Sandy hit New York and New Jersey. Figure 7 also shows the subsets of signals that include either our job loss or search and posting terms. (Note that the scale differs by a factor of 1,000 from that of the total.) We can see that search and posting terms spike simultaneously with Sandy's arrival, while job loss

[22] For example, Kimball et al. (2006) had a survey in the field when Hurricane Katrina hit; they used it to study the hurricane's effect on psychometric measures of happiness.

[23] We do see increased mentions of hurricanes in the signals during the storm seasons in our data.

references increase just after the hurricane arrived. The job market related signals continued at elevated rates well after mentions of the storm *per se* peaked.[24]

B. Government Shutdown

The effects of the government shutdown in October 2013 are clearly evident in the job loss index (Figure 3) and in the search and posting indexes (Figure 5). All have pronounced falls in the first two weeks during the shutdown and equal bounce backs in the weeks following the shutdown. These results imply that the shutdown had a significant effect on labor market activity, but that the effect was short lived.

Interestingly, beginning in September 2013, there are changes in search and posting relative to job loss that are consistent with movements along the Beveridge Curve (Figure 6). These observations are dominated by the effects of the government shutdown and reopening that are evident in the time series. Hence the labor market disruption associated with the government shutdown appears to be a classic demand shock instead of a disruption of the matching function.

This episode illustrates the usefulness of social media for measuring and analyzing the impact of unexpected events. Our social media indexes provide high-frequency and contemporaneous information that is not available in conventional sources.

C. Monitoring the Economy: What Series Should the Fed Follow?

The Federal Reserve is called on to take policy actions based on imperfect and lagging indicators of the economy. There are relatively few high-frequency indicators of real economic activity. Social media indexes may prove to be an important addition to its data arsenal especially during period of economic volatility or at turning points. In her first press conference, Federal Reserve

[24] There is a sharp peak on December 7, 2012 in Tweets that mention "hurricane" or "Sandy" and "unemployment," presumably reflecting the release of the first Bureau of Labor Statistics Employment Situation report to reflect unemployment data after Sandy. This example illustrates the importance of controlling for data releases when analyzing the social media signals.

Chairman Janet Yellen stressed the importance of monitoring labor market data beyond the unemployment rate.[25] Though initial claims is the earliest available indicator of labor market activity, there is reluctance to give it much credence because of its volatility. In this section we use a very simple statistical model to suggest that the social media index carries more signal about the true state of job flows.

To assess the signal in the social media index we estimate a *very* simple statistical model. Suppose that the job loss x_t^* is unobserved. Suppose further that we have two indicators of job loss, initial claims denoted y_{1t} and the social media index denoted y_{2t}. Suppose further that the true state of job loss follows a random walk with innovation ε_t and that the two indictors are noisy measures of the true state with serial and mutually uncorrelated errors ν_{1t} and ν_{2t}. Hence, suppressing constants,

$$\begin{aligned}\Delta x_t^* &= \varepsilon_t \\ \Delta y_{1t} &= \varepsilon_t + \nu_{1t} - \nu_{1t-1} \\ \Delta y_{2t} &= \varepsilon_t + \nu_{2t} - \nu_{2t-1}\end{aligned}$$

describes the joint behavior of the observable indicators. Table 11 shows the estimated standard deviations of the disturbances, σ_ε, σ_{ν_1}, and σ_{ν_2}.[26] Recall that, except for scaling, the two

[25] "Now, the Committee has never felt that the unemployment rate is a sufficient statistic for the labor market. I think if I had to choose one indicator of the labor market, the unemployment rate is probably as good a one as I could find. But in assessing the real state of slack in the labor market and ultimately of inflationary pressures that might—or deflationary pressures that could result from that—it's appropriate to look at many more things. And that's why the Committee now states we will look at a broad range of information. So the closer we get as we narrow in on coming closer to the target we want to achieve, we will be carefully considering many indicators of how close are we to our targets." (Federal Reserve Board 2014). While Yellen is clearly referring to official indicators such as labor market participation and flows from the JOLTS, we expect that the Fed would be interested in social media indicators provided they can be shown to carry information.

[26] Estimation is via method of moments on demeaned data. The sample period is July 16, 2011 through November 2, 2013. Note that the identification of the parameters derives both from the

indicators are totally independent of each other. Given the short sample available, even this very parsimonious model is not that well pinned down by the data. As a result, the variance of the unobserved true state of the economy is not that precisely estimated. The specification of the model in differences rather than levels also, of course, reduces the fit. The estimated standard deviation of the social media indicator is much smaller than that of initial claims, suggesting that the social media index provides a more precise estimate of the true state of the underlying job loss process.

The model formalizes what is apparent in the plots of the series: The initial claims data are much more volatile, so they have a much lower signal-to-noise ratio than the social media index. Figure 8 makes this concrete by showing the initial claims and social media indicators plotted against the implied estimate of the underlying unobserved state of job loss using the parameters in Table 11. The unobserved state is estimated using the Kalman filter. The shaded area (after November 2, 2013) is entirely out of sample, in that the parameters in Table 11 that generate this series are estimated on data prior to November 2.

Figure 8 shows graphically what is implied by the estimates reported in Table 11: the social media index is a better indicator of the underlying state of the job market than is initial claims. Note that during the "out of sample" period, initial claims has been very volatile, reflecting processing problems, weather fluctuations, and possibly changes in program benefits. The initial claims series also appeared to be falling, suggesting improvements in the labor market, that are not present or as strong in the social media index. In the latest data, the initial claims index has increased, reinforcing the view from the social media index that job loss had

cross correlation of the indicators and the serial correlation of each of them. That is, given the stochastic model, the univariate representation of each indicator is MA(1) in differences. Consistent with the specified model, the indicators do indeed have negative MA(1) coefficients univariately.

not declined. Had the Fed been monitoring the social media index during this period, the index would have reinforced its decision to maintain monetary support in a way that the initial claims series did not.

V. Conclusion

This paper addresses the challenge of turning the vast output of social media into data that can be used to create meaningful measurements. Doing so requires processing a very large dataset, coding social media signals for analysis, and using statistical methods to transform them into economic data. This paper accomplishes these tasks. It creates a social media signal of job loss that closely tracks initial claims for unemployment insurance. Despite obvious differences in the underlying processes generating unemployment insurance claims and Tweets about job loss, the social media index tracks the official data at both high and medium frequencies. We construct a real-time index and show that this index has information about initial claims not reflected in either the consensus forecast or the lagged data. The indexes shed light on specific events such as Hurricane Sandy and the government shutdown.

Based on an admittedly short sample period, a simple statistical analysis suggests that the social media index is a better indicator of the true state of job loss than is initial claims. There is no other official statistic that is available at as high frequency and in close to real time for assessing job flows. Given the volatility of the official series, including in early 2014 when the Federal Reserve was beginning to contemplate a change in policy direction, having additional informative real-time, high-frequency measures of job loss should be valuable to policy makers.

We began our analysis with a concept—job loss—that has a relatively well-measured analogue in high-frequency official data. Having shown that a social media index can track a concept that is relatively well-measured, we turn to concepts that are less well measured. In particular, we construct indexes of job search and job posting, concepts of keen interest to analysts of the labor market, but less well measured in official statistics. We apply these series to show that the Beveridge Curve appears to be shifting inward since mid-2011—reversing outward shifts that other researchers identified during the Great Recession.

Longer time series and further analysis are needed to confirm the usefulness of social media in constructing indicators of economic activity. Nonetheless, this paper has demonstrated that it is both feasible and useful to infer information about the state of the labor market from postings on social media that are generated by individuals in the normal course of their social interactions. Variables derived from social media can be both substitutes and complements to data generated from surveys and administrative records by statistical agencies and the private sector. They have the promise of providing measurements at relatively low cost, with high frequency, and virtually in real time, so they have potential advantages over traditional data sources. That is not to suggest, however, that social media could supplant official statistics. Official statistics provide necessary benchmarks for understanding even the best measured variables from social media. In practice, the rapid evolution of the use of social media could make the relationship between the measurement and the underlying fundamental being measured unstable. Our recursive procedure in constructing the University of Michigan Social Media Job Loss Index is one approach to addressing this potential instability. As we accumulate longer time series, research using methods described in this paper is necessary to evaluate the extent to which social media data do track activity. Nonetheless, as with the search and posting series constructed in this paper, social media data provide an opportunity to track hard-to-measure components of economic activity by capturing information that has been previously neglected or is difficult to measure in traditional sources.

References

Antenucci, Dolan, Michael J. Cafarella, Margaret C. Levenstein, Christopher Ré, and Matthew D. Shapiro. (2013a). "Ringtail: Feature Selection for Easier Nowcasting." *WebDB 2013*.

Antenucci, Dolan, Erdong Li, Shaobo Liu, Bochun Zhang, Michael J. Cafarella, and Christopher Ré (2013b). Ringtail: A Generalized Nowcasting System. *Proceedings of the VLDB Endowment*.

Baker, Scott and Andrey Fradkin (2013). The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data. Unpublished working paper, Stanford University.

Barnichon, Regis, Michael Elsby, Bart Hobijn, and Ayşegül Şahin (2012). "Which Industries are Shifting the Beveridge Curve" *Monthly Labor Review* 135, 25-37. http://www.bls.gov/opub/mlr/2012/06/art2full.pdf

Bloomberg Surveys (various dates). *U.S. Economic Forecasts*. Downloaded from http://www.bloomberg.com/markets/economic-calendar/ and Bloomberg terminal on November 8, 2013.

Brieman, Leo (2001). "Random Forests." *Machine Learning* 45(1), 5-32.

Bureau of Labor Statistics (2013). "Job Openings and Labor Turnover Report." [Survey instrument downloaded February 5, 2014. http://www.bls.gov/jlt/jltc1.pdf]

Bureau of Labor Statistics (2014). "Job Opening and Labor Turnover – November 2013." [JOLTS new release (January 17, 2014) downloaded February 5, 2013. http://www.bls.gov/news.release/pdf/jolts.pdf]

Conference Board, The (2014). "The Conference Board Help Wanted Online (HWOL)." [Downloaded on February 5, 2014 from http://www.conference-board.org/data/helpwantedonline.cfm]

Choi, Hyunyoung and Hal Varian (2009a). "Predicting the Present with Google Trends." Technical Report, Google. http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf.

Choi, Hyunyoung and Hal Varian (2009b). “Predicting Initial Claims for Unemployment Insurance Using Google Trends.” Technical Report, Google. http://research.google.com/archive/papers/initialclaimsUS.pdf.

Davis, Steven J., R. Jason Faberman, and John C. Haltiwanger (2013). The Establishment-Level Behavior of Vacancies and Hiring.” *Quarterly Journal of Economics* 128 (2), 581-622.

Garthwaite, Craig, Tal Gross, and Matthew J. Notowidigdo (2013). Public Health Insurance, Labor Supply, and Employment Lock. Unpublished working paper.

Ginsberg, J., M. H. Mohebbi, R. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant (2009). “Detecting Influenza Epidemics Using search Engine Query Data.” *Nature* (February).

Goel, Sharad, Duncan J. Watts, and Daniel G. Goldstein (2012). “The Structure of Online Diffusion Networks” Proceedings of the 13th ACM Conference on Electronic Commerce (EC 2012). http://5harad.com/papers/diffusion.pdf.

Hobijn, Bart and Ayşegül Şahin (2013). “Beveridge Curve Shifts across Countries since the Great Recession.” Federal Reserve Bank of San Francisco Working Paper 2012-24 http://www.frbsf.org/economic-research/files/wp12-24bk.pdf.

Kimball, Miles, Helen Levy, Fumio Ohtake, and Yoshiro Tsutsui (2006). “Unhappiness after Hurricane Katrina.” NBER Working Paper No. 12062.

Federal Reserve Board (2014). “Transcript of Chair Yellen’s Press Conference March 19, 2014.” [FOMCpresconf20140319.pdf downloaded from http://www.federalreserve.gov/mediacenter/media.htm on April 29, 2014.]

Lazer, David, Ryan Kennedy, Gary King, and Alessandro Vespignani (2014). The Parable of Google Flu: Traps in Big Data Analysis *Science* (14 March 2014) 1203-1205.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and J. Niels Rosenquist (2011). Understanding the Demographics of Twitter Users. *ICWSM* 11.

Scott, Steven L. and Hal R. Varian (2013). “Bayesian Variable Selection for Nowcasting Economic Time Series” NBER Working Paper No 19567. http://www.nber.org/papers/w19567

Table 1. Summary Statistics for Signals: Job Loss and Unemployment (Weekly Rate per Million Tweets)

Signal	Mean	Standard Deviation	Coefficient of Variation
Axed	3.25	1.51	0.46
Canned	8.86	3.42	0.39
Downsized	0.49	0.25	0.51
Outsourced	2.11	1.35	0.64
Pink slip	1.34	1.31	0.98
Lost job	3.21	0.86	0.27
Fired job	27.45	6.67	0.24
Been fired	15.19	6.76	0.45
Laid off	15.70	3.59	0.23
Unemployment	53.33	20.07	0.38

Note: Sample period is July 16, 2011 through November 2, 2013 (weeks ending Saturday). Sample is 19.3 billion total Tweets of which 2.4 million are job loss and unemployment related.. See Appendix Table 1 for detailed descriptions of phrases for signals.

Table 2. Correlation of Job Loss and Unemployment Signals

	Axed	Canned	Downsized	Outsourced	Pink slip	Lost job	Fired job	Been fired	Laid off	Unemployment
Axed	1									
Canned	0.37	1								
Downsized	0.34	0.29	1							
Outsourced	0.18	0.31	0.34	1						
Pink slip	-0.05	0.00	-0.10	-0.12	1					
Lost job	0.45	0.49	0.46	0.40	0.02	1				
Fired job	0.48	0.46	0.28	0.18	-0.08	0.52	1			
Been fired	0.45	0.36	0.04	0.01	-0.03	0.30	0.65	1		
Laid off	0.43	0.52	0.46	0.43	-0.05	0.59	0.63	0.24	1	
Unemployment	0.35	0.40	0.50	0.44	-0.14	0.54	0.47	0.21	0.66	1

Note: Sample period is July 16, 2011 through November 2, 2013. The Unemployment signal is purged of the Employment Situation effect.

Table 3. Factor Loadings on Job Loss and Unemployment Signals

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
Signal:										
Axed	0.65	0.29	-0.06	0.39	0.45	0.13	-0.33	0.07	0.03	-0.04
Canned	0.68	0.11	0.14	-0.36	0.21	-0.56	-0.02	0.11	0.04	-0.05
Downsized	0.60	-0.42	0.04	0.51	0.01	-0.18	0.36	0.14	-0.12	0.01
Outsourced	0.52	-0.54	0.02	-0.40	0.33	0.37	0.13	0.09	-0.09	-0.03
Pink slip	-0.11	0.22	0.95	0.04	-0.03	0.13	0.03	0.11	0.01	-0.02
Lost job	0.78	-0.07	0.19	0.03	0.04	0.00	0.08	-0.56	0.12	0.02
Fired job	0.77	0.41	-0.11	-0.05	-0.27	0.11	0.05	-0.02	-0.24	-0.28
Been fired	0.51	0.72	-0.17	-0.08	0.04	0.14	0.32	0.11	0.11	0.20
Laid off	0.83	-0.13	0.08	-0.07	-0.26	0.00	-0.29	0.03	-0.26	0.25
Unemployment	0.76	-0.29	-0.04	0.01	-0.32	0.09	-0.10	0.18	0.42	-0.06
Variance of factor	4.27	1.40	1.02	0.72	0.60	0.55	0.46	0.42	0.36	0.19
Cumulative fraction of variance	0.43	0.57	0.67	0.74	0.80	0.86	0.90	0.95	0.98	1.00

Note: Sample period is July 16, 2011 through November 2, 2013 (weeks ending Saturday). Principal component factors calculated based on the correlation matrix of signals shown in Table 2.

Table 4. Predicting Initial Claims: Consensus, Social Media Factor, and Lagged Dependent Variable

A. Preliminary Data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	90.24	30.16	98.46	31.57	43.00	19.87	23.60
	(21.98)	(23.09)	(22.93)	(23.56)	(21.51)	(22.48)	(22.72)
Lagged initial claims	0.75			0.05	0.48		0.17
	(0.06)			(0.16)	(0.07)		(0.15)
Consensus forecast		0.92		0.86		0.67	0.48
		(0.06)		(0.18)		(0.10)	(0.21)
Social media: factor 1 (scaled)			0.73		0.40	0.27	0.29
			(0.06)		(0.07)	(0.08)	(0.09)
Adjusted R^2	0.57	0.64	0.53	0.63	0.65	0.66	0.66

B. Revised Data

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Constant	74.73	46.33	103.43	46.56	43.36	34.17	34.22
	(20.27)	(20.75)	(20.02)	(20.55)	(18.81)	(19.56)	(19.28)
Lagged initial claims	0.80			0.27	0.50		0.29
	(0.05)			(0.15)	(0.07)		(0.14)
Consensus forecast		0.88		0.61		0.59	0.30
		(0.06)		(0.16)		(0.08)	(0.16)
Social media: factor 1 (scaled)			0.73		0.38	0.32	0.33
			(0.05)		(0.07)	(0.07)	(0.07)
Adjusted R^2	0.64	0.67	0.59	0.67	0.71	0.71	0.72

Note: Sample period is July 16, 2011 through November 2, 2013. Standard errors in parentheses.
Dependent variable: Initial Claims for Unemployment Insurance (preliminary data in panel A, revised data in panel B).
Regressors: Lagged dependent variable, consensus forecast, and social media factor 1 scaled to have same units as initial claims.

Table 5. Constructing the Social Media Job Loss Index

	(1)	(2)	(3)	(4)	(5)	(6)
Constant	368.26	368.25	368.23	368.23	365.77	368.81
	(1.52)	(1.50)	(1.50)	(1.48)	(9.11)	(1.73)
Factor 1	4.70	4.70	4.71	4.71	4.69	4.69
	(0.36)	(0.35)	(0.35)	(0.35)	(0.36)	(0.36)
Factor 2		-2.35	-2.35	-2.35		
		(1.07)	(1.07)	(1.06)		
Factor 3			-1.52	-1.52		
			(1.48)	(1.47)		
Factor 4				3.76		
				(2.05)		
Seasonal factor for initial claims					2.54	
					(9.16)	
Employment Situation week						-2.43
						(3.66)
Adjusted R2	0.59	0.60	0.60	0.61	0.59	0.59

Note: The dependent variable is the Department of Labor Initial Claims for Unemployment Insurance (thousands, seasonally adjusted). The independent variables are the job loss and unemployment factors. The Social Media Job Loss Index is based on regressions re-estimated each week using real-time data available as of the prediction period, as described in text. This table presents the estimates for the final week in the sample. Sample period is July 16, 2011 through November 2, 2013. Standard errors in parentheses.

Table 6. Prediction Errors of Social Media Job Loss Index

	Root Mean Squared Error	
Specification	Preliminary Data	Revised Data
(1) Factor 1	21.9	19.2
(2) Factor 1,2	22.7	20.0
(3) Factor 1,2,3	23.7	21.4
(4) Factor 1,2,3,4	22.6	20.6
(5) Factor 1, Seasonal Factor	22.1	19.4
(6) Factor 1, Employment Situation Week	22.1	19.3

Note: Table gives the root mean squared error (RMSE) of the Social Media Job Loss Index for initial claims for unemployment insurance (preliminary data and revised data). The models and RMSEs are estimated recursively, using data from July 16, 2011 forward, for weeks ending July 16, 2011 through November 2, 2013.

Table 7. Incremental Information in Social Media Job Loss Index

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable	Preliminary Initial Claims − Consensus			Revised Initial Claims − Consensus		
Constant	-9.41	-55.49	-10.57	-6.78	-75.54	-7.29
	(3.16)	(53.78)	(3.43)	(2.92)	(49.19)	(3.18)
Social Media Job Loss Index – consensus	0.72		0.63	0.75		0.71
	(0.20)		(0.22)	(0.18)		(0.21)
Social Media Job Loss Index		0.85			0.94	
		(0.25)			(0.23)	
Consensus		-0.72			-0.75	
		(0.20)			(0.18)	
Lag of			0.19			0.09
Social Media Job Loss Index – consensus			(0.22)			(0.20)
Adjusted R^2	0.15	0.15	0.15	0.19	0.20	0.18

Note: Sample period is July 16, 2011 through November 2, 2013 (recursive sample). Standard errors in parentheses.
Dependent variables: Columns (1)-(3), Preliminary initial claims minus consensus; Columns (4)-(6), Revised initial claims minus consensus.

Table 8. Summary Statistics for Signals: Job Search and Job Posting (Weekly Rate per Million Tweets)

Category	Signal	Mean	Standard Deviation	Coefficient of Variation
Search	Find	25.23	4.96	0.20
	Look	25.31	8.28	0.33
	Need	90.74	19.79	0.22
	Search	2.13	0.93	0.44
Search and posting	Seek	0.64	0.29	0.45
Posting	Hiring	220.81	212.79	0.96
	Job	161.14	59.97	0.37
	Work	384.37	67.66	0.18

Note: Sample period is July 16, 2011 through November 2, 2013 (weeks ending Saturday). Sample is 17.2 billion Tweets (116 weeks, 148.21 million Tweets per week on average). See Appendix Table 2 for detailed descriptions of phrases for signals.

Table 9. Correlation of Job Search and Job Posting Signals

	Find	Look	Need	Search	Seek	Hiring	Job	Work
Find	1							
Look	0.20	1						
Need	0.34	0.85	1					
Search	0.14	0.45	0.45	1				
Seek	0.25	0.33	0.29	0.45	1			
Hiring	-0.01	0.05	0.21	0.13	0.21	1		
Job	0.26	0.24	0.13	0.30	0.46	0.24	1	
Work	0.39	0.50	0.52	0.31	0.50	0.19	0.56	1

Note: Sample period is July 16, 2011 through November 2, 2013.

Table 10. Job Search and Job Posting Factors: Loadings of First Factor, Alternative Sets of Signals

	Search	Posting
Find	0.45	
Look	0.86	
Need	0.87	
Search	0.70	
Seek	0.61	0.77
Hiring		0.46
Job		0.81
Work		0.82
Variance	2.57	2.14
Fraction of variance	0.51	0.53

Note: Table shows the factor loading for the first factor for the selected signals. The bottom two rows report the variance of the first factor and the fraction of the overall variance accounted for by the first factor.

Table 11. Unobserved Components Model of Job Loss

Component	Parameter	Estimate
True	σ_{ε}	3.0 (2.2)
Noise, Initial Claims	σ_{ν_1}	11.2 (1.0)
Noise, Social Media Index	σ_{ν_2}	5.0 (0.7)

Note: Sample period is July 16, 2011 through November 2, 2013. Estimates by method of moments as described in text. Standard errors in parentheses.

Appendix Table 1. Social Media Signals: Job Loss and Unemployment

Category	Signal	Phrase	Number of distinct matched phrases
Job loss	Axed	axed	1
	Canned	canned	1
	Downsized	downsized	1
		down|sized	1
	Outsourced	outsourced	1
	Pink slip	pinkslip	1
		pink|slip	1
	Lost job	lost|*|job	45
	Fired	fired|*|job	28
		fired|*|work	16
		fired|from	1
		fired|lol	1
		get|fired	1
		got|fired	1
		just|fired	1
	Been fired	been|fired	1
		being|fired	1
		be|fired	1
		was|fired	1
	Laid off	laidoff	1
		laid|off	1
		layed|off	1
		layoff	1
		lay|off	1
Unemployment	Unemployment	unemploy	1
		unemployed	1
		unemployment	1

Note: The signals are counts of Tweets that contain 4-grams with the indicated phrase where “|” denotes a space and “*” is a wildcard. The last column indicates the number of distinct phrases found in the database of Tweets matching the target phrases with wildcards. See text for details.

Appendix Table 2. Social Media Signals: Job Search and Job Posting

Category	Signal	Phrases	Number of distinct matched phrases
Search	Find	find|*|job	242
	find|	*|work	178
	Look	look*|*|job	237
	look*|	*|work	497
	Need	need|*|job	398
	need|	*|work	515
	Search	search*|*|job	93
	search*	|*|work	38
Search and Posting	Seek	seek*|*|job	30
		seek*|*|position	11
	seek*|	*|work	29
Posting	Hiring	hiring|*	17278
	Job	job|opportunities	1
	job|	opportunity	1
	jobs|	in|*	3040
	jo	bs|near|*	36
	job|	in|*	4397
	job|	near|*	18
	Work	work|in|*	10163
	wo	rk|near|*	32
	work|	opportunities	1
	work|	opportunity	1

Note: See Appendix Table 1.

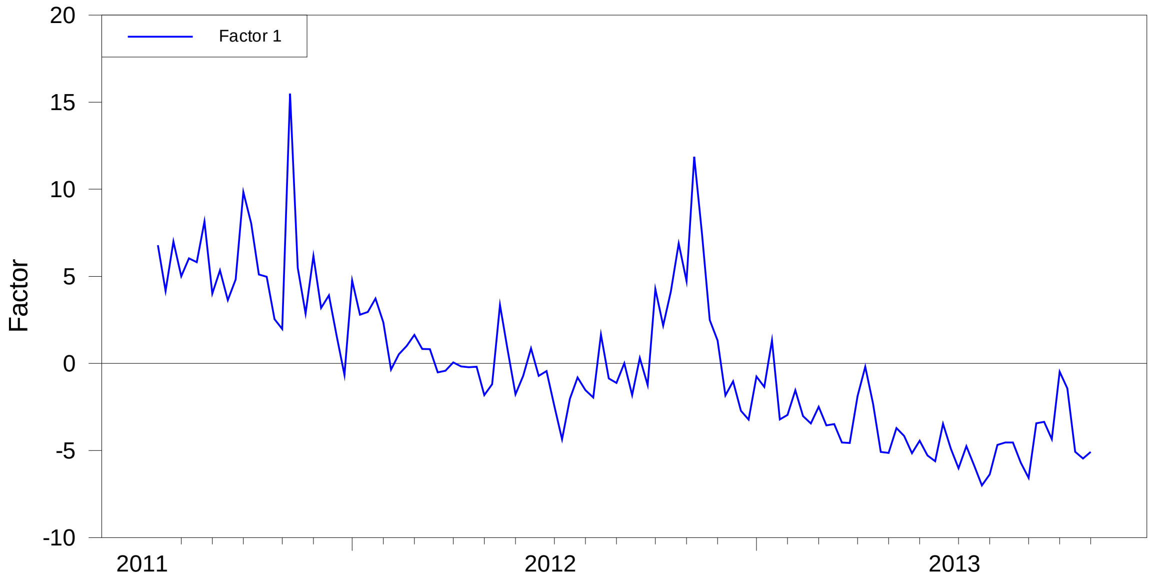
Appendix Table 3. Signals by Age and Sex

		Fraction of Signals (percent)			
		All	Job loss	Search	Posting
Age	14-18	20.9	12.3	37.1	9.6
	19-21	8.2	5.6	13.0	5.4
	22-24	10.1	9.0	12.4	6.8
	25-34	14.5	15.2	14.0	11.0
	35-44	13.1	15.9	8.3	14.5
	45-64	33.2	42.1	15.1	52.7
	Total	100.0	100.0	100.0	100.0
Sex	Male	60.6	66.7	50.5	59.8
	Female	39.4	33.3	49.5	40.2
	Total	100.0	100.0	100.0	100.0

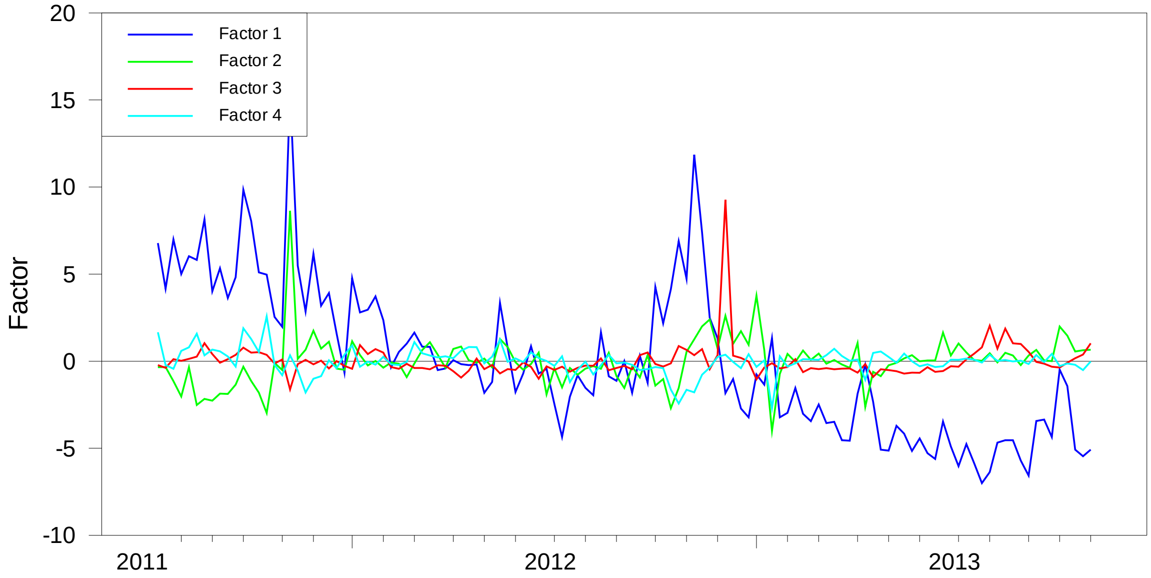
Table shows fraction of job-related signals by age and sex of sender. The demographics are estimated probabilistically and are coded for only a subset of signals. Because of changes in the API, this sample ends June 15, 2013.

Figure 1. Twitter Job Loss and Unemployment Signals

A. Factor 1

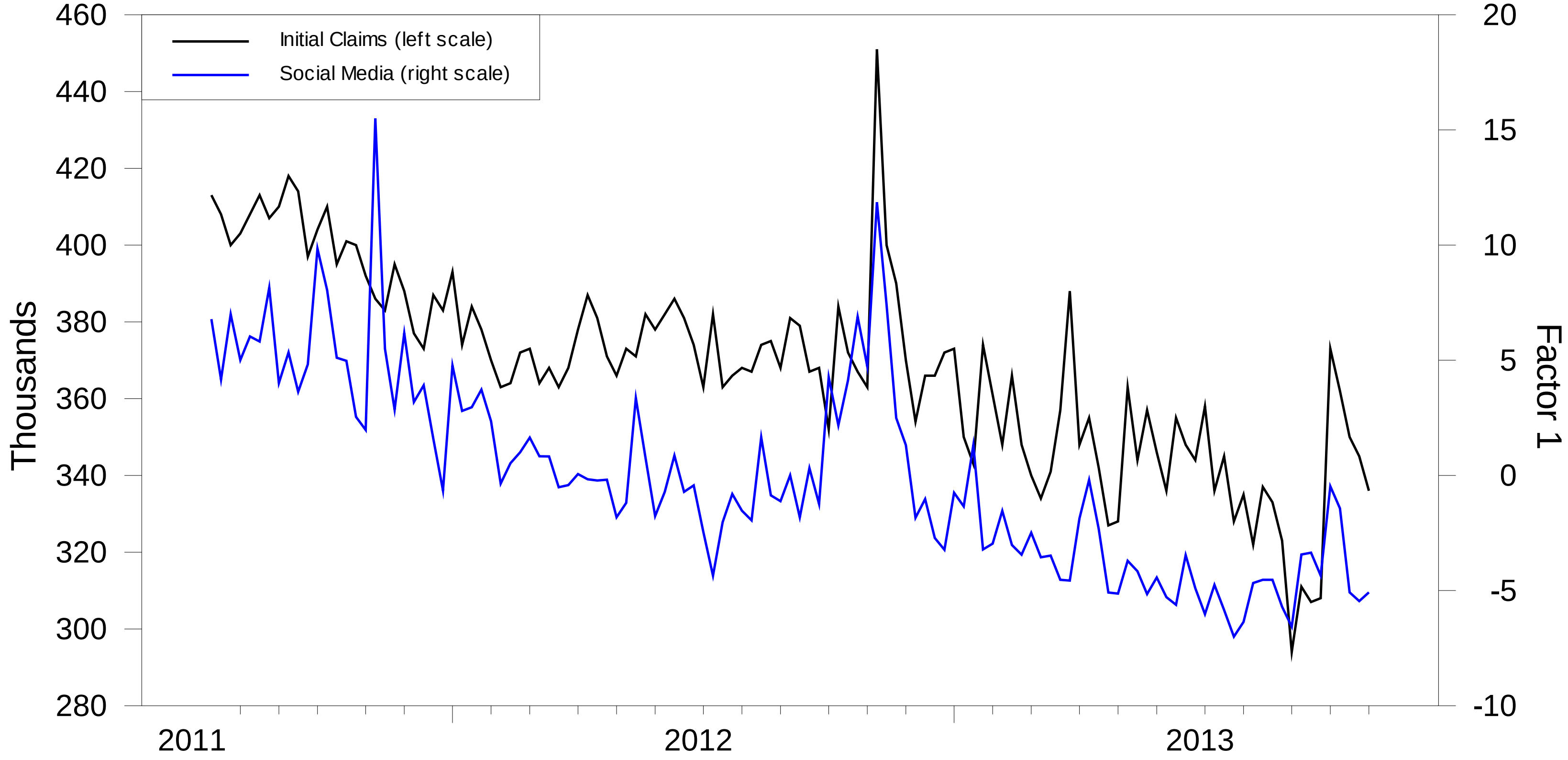


B. Factors 1 – 4



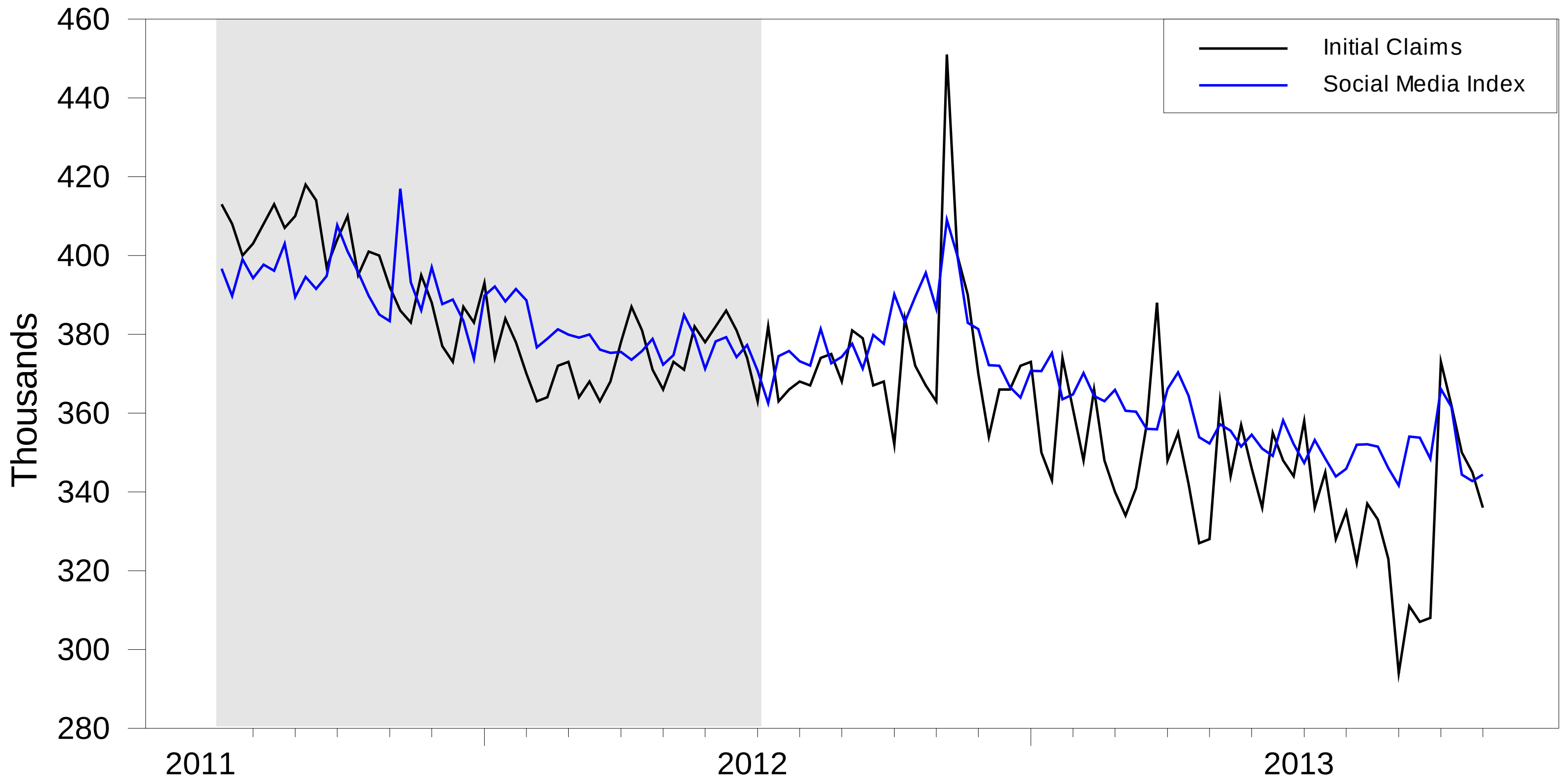
Note: Sample period is July 16, 2011 through November 2, 2013 (weeks ending Saturday). Principal component factors calculated based on the correlation matrix of signals shown in Table 2.

Figure 2. Initial Claims for Unemployment Insurance and Job Loss and Unemployment Factor 1



Note: Figure shows the Department of Labor's Initial Claims for Unemployment Insurance (left scale, revised data, seasonally adjusted) and the Social Media Factor 1 (right scale). The factor is estimated as described in the text and is no way fit to the initial claims data.

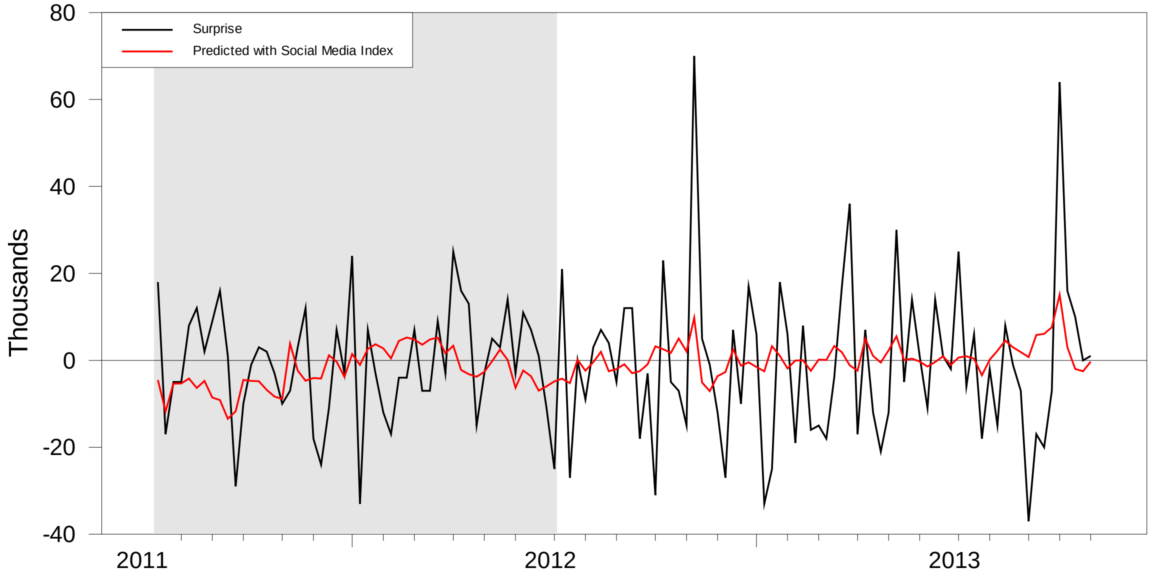
Figure 3. Initial Claims for Unemployment Insurance and the Social Media Job Loss Index



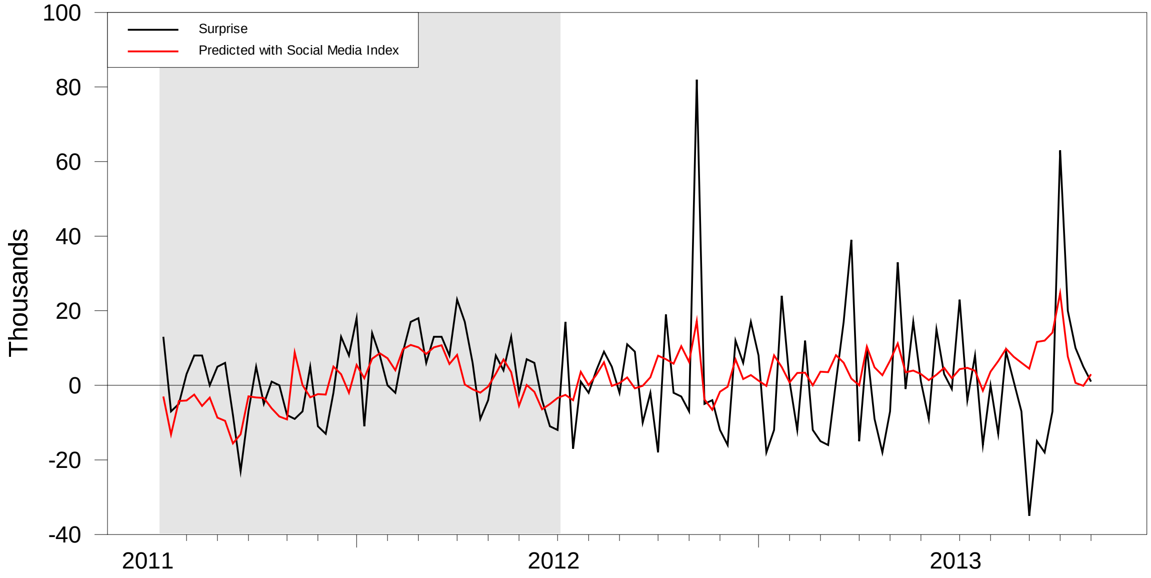
Note: Figure shows the Department of Labor's Initial Claims for Unemployment Insurance and the Social Medial Job Loss Index. The Social Media Job Loss Index is estimated in sample in the shaded area and recursively thereafter. See text for details.

Figure 4. Surprises Predicted by Social Media Job Loss Index

A. Preliminary Data

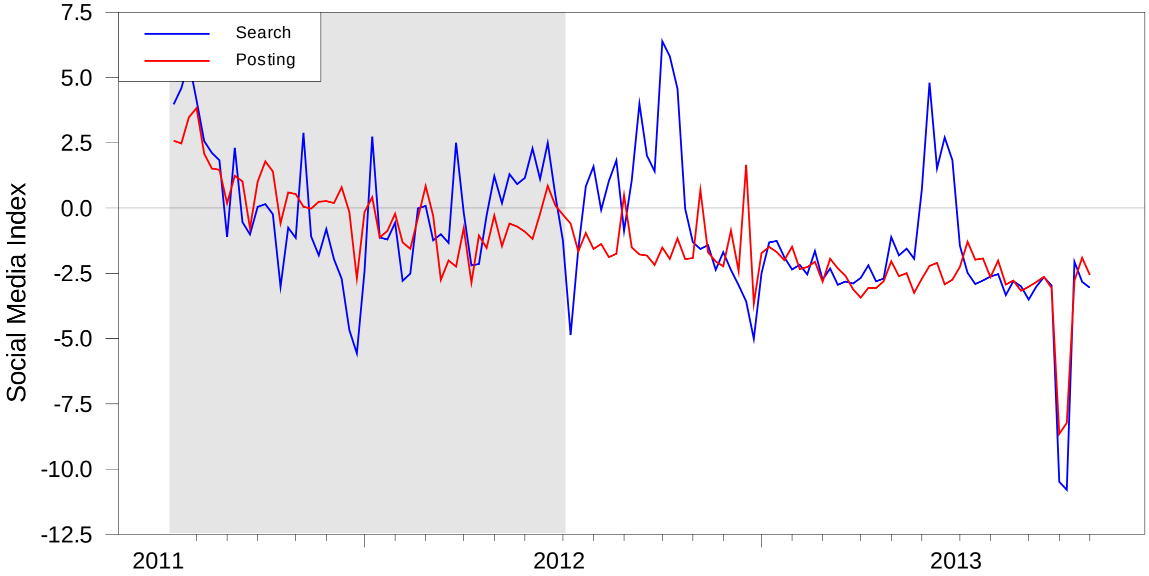


B. Revised Data



Note: Surprise is Department of Labor Initial Claims for Unemployment Insurance (preliminary or revised) minus the consensus forecast. Predicted with Social Media Job Loss Index constructed based on factor 1, as described in the text. The index is generated recursively except in the shaded area, where it is generated over the entire shaded sample.

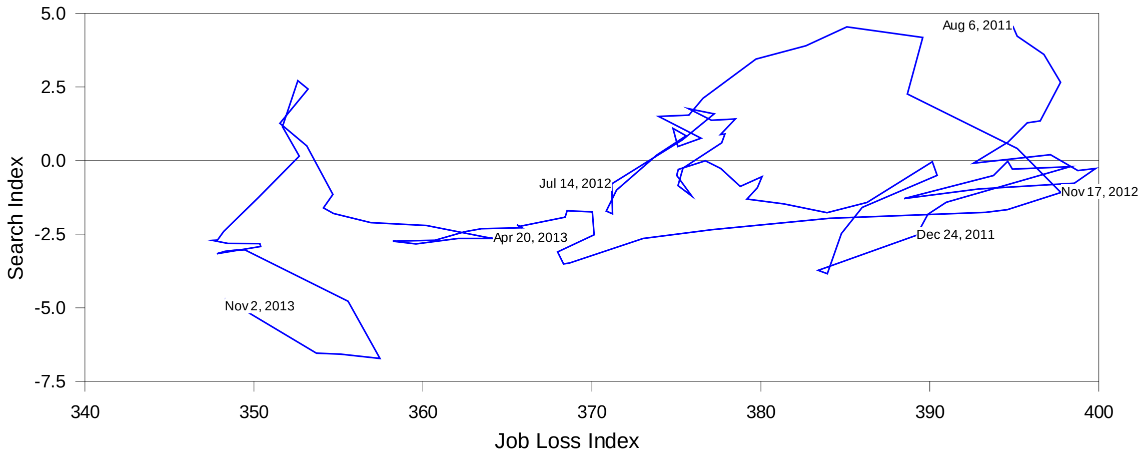
Figure 5. Social Media Indexes for Job Search and Job Posting



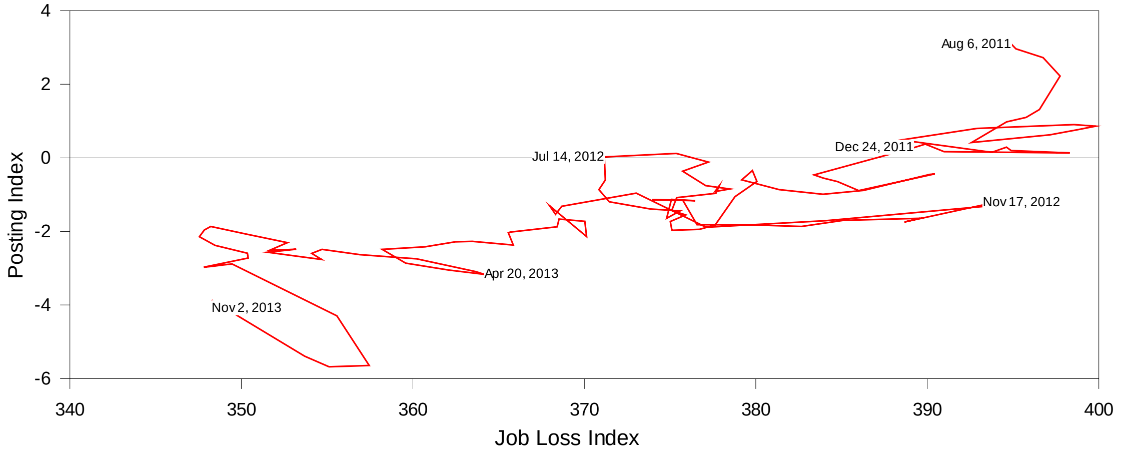
Note: Indexes are based on factor loadings in second two columns of Table 10. The social media indexes are estimated in sample in the shaded area and recursively thereafter.

Figure 6. Beveridge Curve

A. Search



B. Posting



Note: Figures show the four-week moving averages of the Social Media Job Loss Index versus the Search and Posting indexes.

Figure 7. Social Media Signal Related to Hurricane Sandy

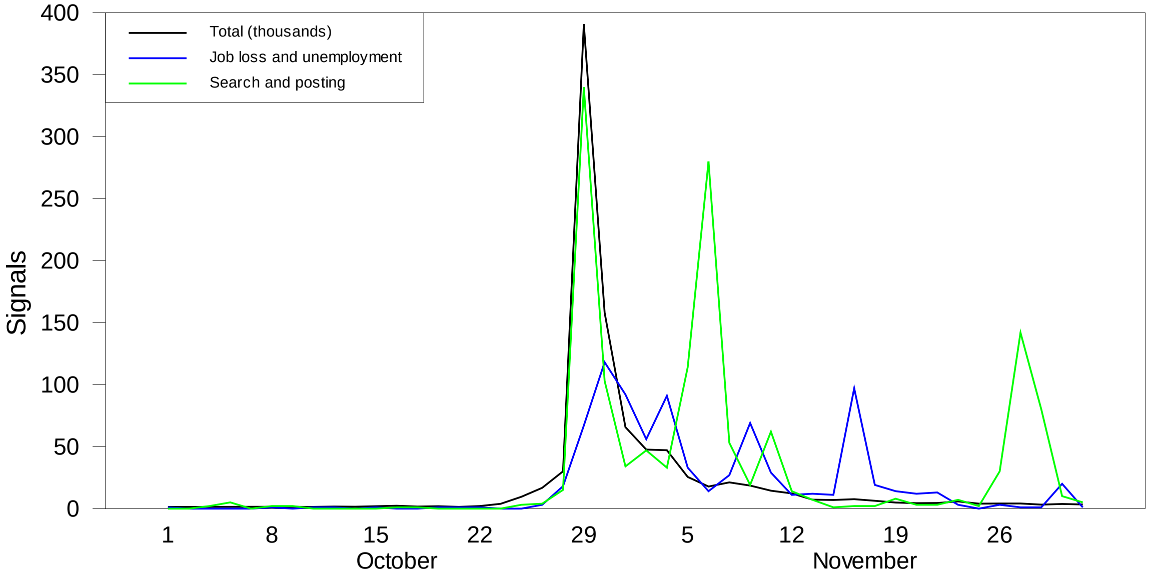
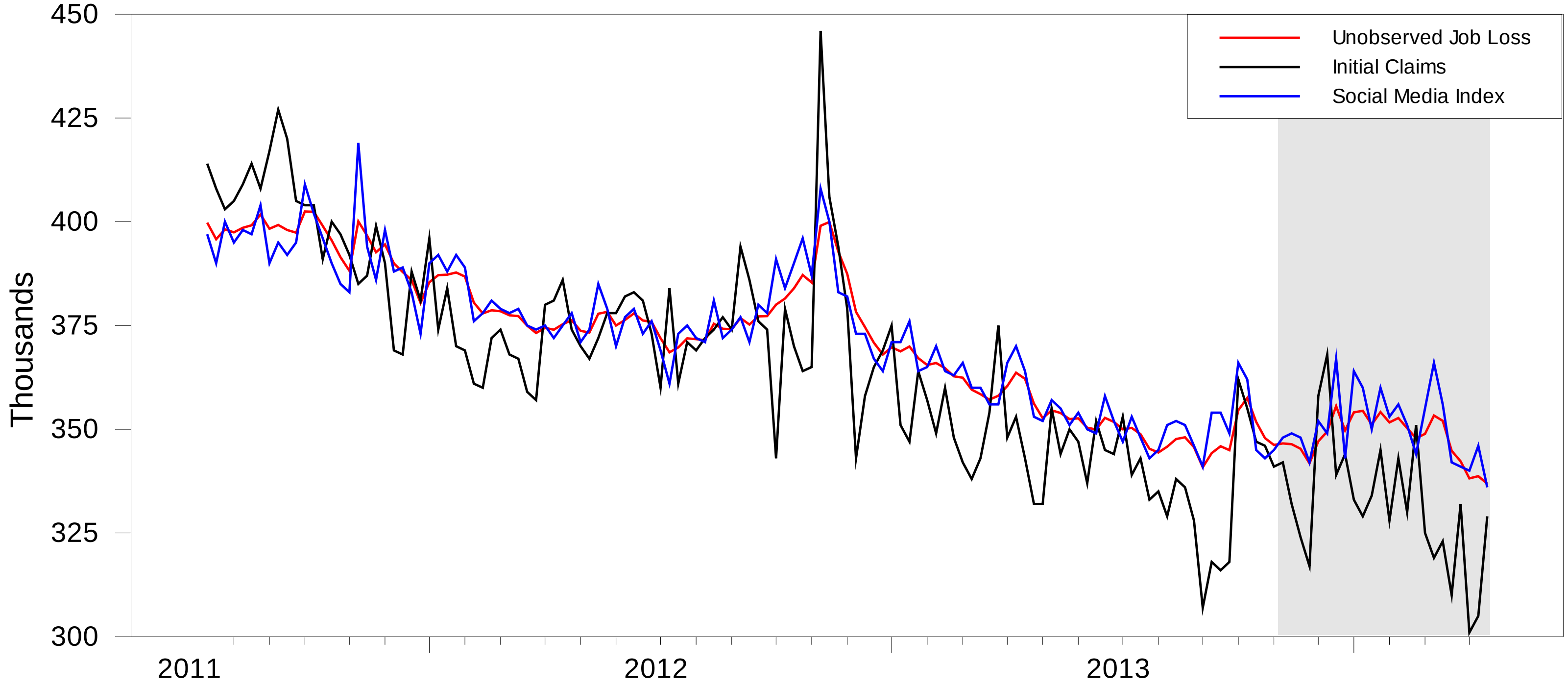


Figure 8. Unobserved Job Loss versus Indicators



Note: The red line is the estimated unobserved job lost calculated using the Kalman filter as described in the text. The black line is new claims and the blue line is the social media index. The shaded area (November 2, 2013 to April 19, 2014) indicates data that are not used to estimate the parameters for the Kalman filter.