Supplementary Information: Tracking Employment Shocks Using Mobile Phone Data

Jameson L. Toole,[1,*] Yu-Ru Lin,[2] Erich Muehlegger,[3]
Daniel Shoag,[4] Marta C. González,[5] and David Lazer[6]

[1]*Engineering Systems Division, MIT, Cambridge, MA, 02144*
[2]*School of Information Science, University of Pittsburgh, Pittsburgh, PA 15260*
[3]*Department of Economics, UC Davis, Davis, CA 95616*
[4]*Harvard Kennedy School, Harvard University, Cambridge, MA, 02144*
[5]*Department of Civil and Environmental Engineering, MIT, Cambridge, MA, 02144*
[6]*Lazer Laboratory, Northeastern University, Boston, MA 02115, USA*

[*] Correspondence to jltoole@mit.edu

I. SUPPLEMENTARY INFORMATION

A. Materials and Methods

CDR Data set 1 (D1): We analyze call detail records (CDRs) from two industrialized European countries. In the first country, we obtain data on 1.95 million users from a service provider with roughly 15% market share. The data run for 15 months across the years 2006 and 2007, with the exception of a gap between August and September 2006. Each call record includes a de-identified caller and recipient IDs, the locations of the caller's and recipient's cell towers and the length of the call. Caller or recipients on other network carriers are assigned random IDs. There are approximately 1.95 million individuals identified in the data, 453 million calls, and 16 million hours of call time. The median user makes or receives 90 calls per months.

CDR Data set 2 (D2): The second data set contains 10 million users (roughly 20% market share) within a single country over three years of activity. Like D1, each billing record for voice and text services, contains the unique identifiers of the caller placing the call and the callee receiving the call, an identifier for the cellular antenna (tower) that handled the call, and the date and time when the call was placed. Coupled with a data set describing the locations (latitude and longitude) of cellular towers, we have the approximate location of the caller when placing the call. For this work we do not distinguish between voice calls and text messages, and refer to either communication type as a "call." However, we also possess identification numbers for phones that are outside the service provider but that make or receive calls to users within the company. While we do not possess any other information about these lines, nor anything about their users or calls that are made to other numbers outside the service provider, we do have records pertaining to all calls placed to or from those ID numbers involving subscribers covered by our data set. Thus egocentric networks between users within the company and their immediate neighbors only are complete. This information was used to generate egocentric communication networks and to compute the features described in the main text. From this data set, we generate a random sample population of k users for each of the provinces, and track each user's call history during our 27-month tracking period (from December 2006 to March 2009). We discuss how the sample

size is chosen in a following subsection. Finally, we note that due to an error in data extraction from the provider, we are missing data for Q4 in 2007.

The use of CDR data to study mobility and social behaviors on a massive scale is becoming increasingly common. In addition to its large scale, its format is generally consistent between countries and mobile operators. In the context of this study, each mobile phone data set contains five columns: 1) an anonymized, unique identifier for the caller, 2) the ID of the tower through which the caller;s call was routed, 3) an anonymized, unique identifier of the receiver of the call, 4) the ID of the tower through which the receiver's call was routed, and 5) the timestamp down to the second which the call was initiated. In order to obtain the location of both caller and receiver, data is restricted to only calls between members of the same mobile operator. The tower IDs reflect the tower used upon starting the call and we have no information on changes in location that may be made during the call. We also obtain a list of latitudes and longitudes marking the coordinates of each tower. Although calls are generally believed to be routed through the tower that is geographically closest to the phone, this may not be the case if the signal is obscured by buildings or topology. For this reason, we consider a cluster of towers near the geographic area in question instead of a single tower.

To ensure privacy of mobile phone subscribers, all identifiers were anonymized before we received the data and no billing or demographic information was provided on individuals or in aggregate.

B. Filtering CDR Data

We limit our sample to mobile phone users who either make or receive at least ten calls connecting through one of the three cell towers closest to the manufacturing plant of interest. In addition, we require that users make at least one call in each month spanned by a given data set to ensure users are still active.

C. Manufacturing plant closure

We gather information on a large manufacturing plant closing that affected a small community within the service provider's territory from news articles and administrative sources

collected by the country's labor statistics bureau. The plant closure occurred in December 2006 and involved 1,100 employees at an auto-parts manufacturing plant in a town of roughly 15,000 people.

D. Town Level Structural Break Model

We model the pre-closure daily population of the town as consisting of three segments: a fraction of non-resident plant workers γ, a fraction of resident workers μ, and a fraction of non-workers normalized to $(1-\gamma-\mu)$. We postulate that each individual i has a flow probability of making or receiving a call at every moment p_i. Workers spend a fraction ψ of their day at their jobs and thus make, in expectation, $p_i\psi$ call on a given day during work hours. When losing their job in the town, both resident and non-resident workers are re-matched in national, not local, labor market.

Given this model, the expected daily number of cell phone subscribers making or receiving calls from the three towers serving the plant and town:

$$vol = \begin{cases} \gamma\psi\bar{p} + (1-\gamma)\bar{p} & \text{for } t < t_{layoff} \\ \mu(1-\psi)\bar{p} + (1-\gamma-\mu)\bar{p} & \text{for } t \geq t_{layoff} \end{cases}$$

This model predicts a discrete break in daily volume from the towers proximate to the plant of $(\gamma+\mu)\psi\bar{p}$ at the date t_{layoff}. For workers, the predicted percentage change in call volume from these towers is $\frac{(\gamma+\mu)\psi\bar{p}}{(\mu\bar{p}+\gamma\psi\bar{p}}$. Non-workers experience no change.

E. Individual Structural Break Model

We fit a model similar to the community structural break model to data for each individual user, i, based on the probability they made a call from the town on each day. For each individual, we use the non-linear estimator to select a break date t^i_{layoff} , and constant pre- and post- break daily probabilities $p_{i,t<t^i_{layoff}}$ and $p_{i,t>t^i_{layoff}}$ to minimize the squared deviation from each individuals' data. Figure 1 plots the distribution of break-dates for individuals. As expected, there is a statistically significant spike in the number of individuals experiencing a break in the probability of making a call from the town at the time of the closure and significantly fewer breaks on other, placebo dates. These two methods provide independent, yet complementary ways of detecting mass layoffs in mobile phone data.

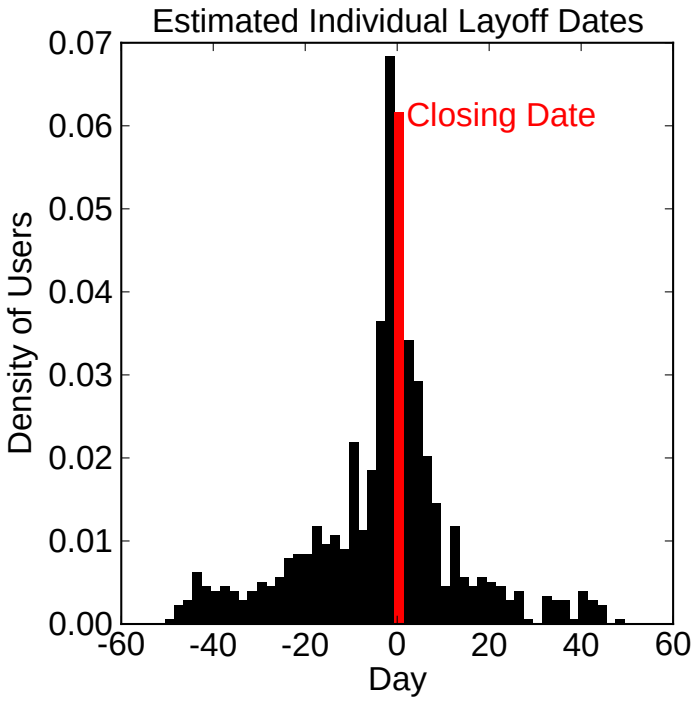

FIG. 1. We plot the distribution of break dates for the structural break model estimated for individuals. We find a strong, statistically significant peak centered on the reported closure date (red) with far fewer breaks on other, placebo dates. This is consistent with both our community wide model as well as the Bayesian model presented above.

F. Bayesian Estimation

On an individual level, Figure 2A shows days on which each user makes a call near the plant ranked from highest to lowest probability weight. Figure 2B provides greater detail for users probability weights between 50% and 100%. Users highly suspected of being laid off demonstrate a sharp decline in the number of days they make calls near the plant following the reported closure date. Figure 2C graphs the inverse cumulative distribution of probability weights. While we do not have ground-truth evidence that any of these mobile phone users was laid off, we find more support for our hypothesis by examining a two week period roughly 125 days prior to the plant closure.

Figures 2A and 2B illustrate that the call patterns of users assigned the highest probabilities change significantly after the plant closure. These users make calls from the town on a consistent basis before the layoff, but make significantly fewer calls from the town afterwards. In contrast, the call patterns of users assigned the lowest weights do not change following the plant closure. In aggregate, we assign 143 users probability weights between 50% and 100%. This represents 13% of the pre-closure plant workforce this fraction compares closely with the roughly 15% national market share of the service provider.

G. The European Labor Force Survey

Each quarter, many European countries are required to conduct labor force surveys to measure important economic indicators like the unemployment rates studied here. In person or telephone interviews concerning employment status are conducted on a sample size of less than 0.5% of the population. Moreover, participants are only asked to provide responses about their employment status during a 1 week period in the quarter.

These "microdata" surveys are then aggregated at the province and national levels. Confirmed labor force reports and statistics for a particular quarter are released roughly 14 weeks after the quarter has ended. For example, Q1 of 2012 begins January 1st, 2012 and ends March 31st, 2012. The survey data is analyzed and unemployment numbers are released between two and three weeks following the end of the quarter. These numbers, however, are unconfirmed and subject to revisions which can occur at any time in the following quarters.

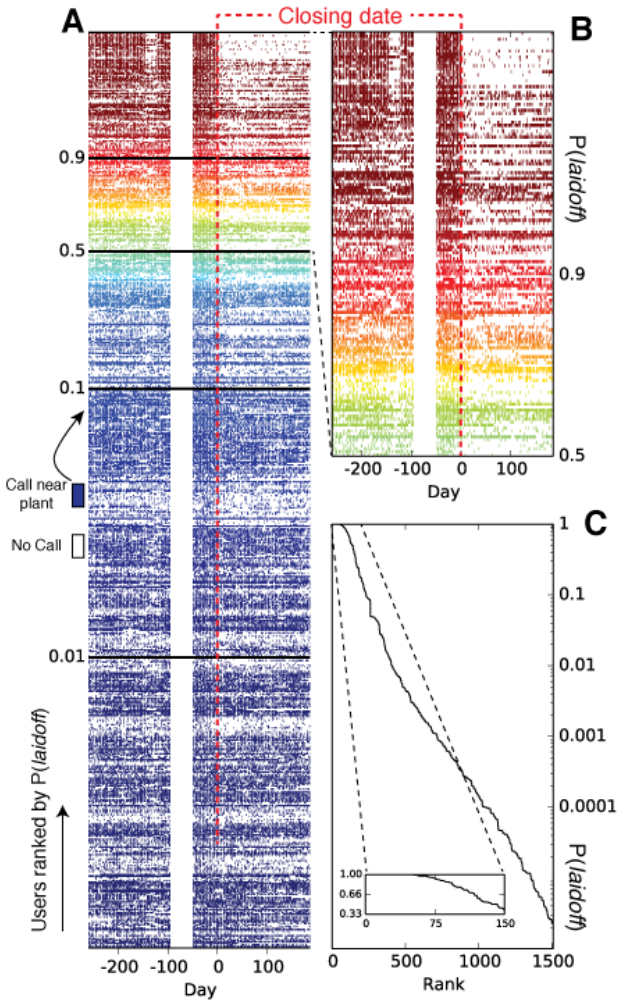

FIG. 2. Identifying affected individuals. A) Each user is represented by a row where we fill in a day as colored if a call was made near the plant on that day. White space marks the absence of calls. Rows are sorted by the assigned probability of that user being laid off. B) A closer view of the users identified as mostly to have been laid off reveals a sharp cut off in days on which calls were made from the plant. C) An inverse cumulative distribution of assigned probability weights. The insert shows an enlarged view at the probability distribution for the 150 individuals deemed most likely to have been laid off.

H. The Effect of Job Loss on Call Volumes

We measure the effect of job loss on six properties of an individual's social behavior and three mobility metrics.

1. CDR Metrics

Calls:: The total number of calls made and received by a user in a given month.

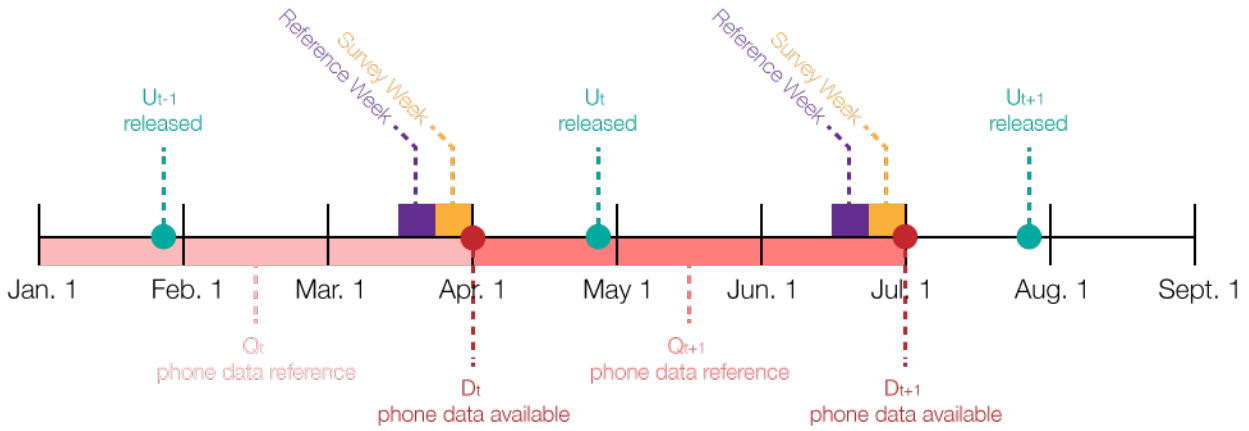

FIG. 3. A timeline showing the various data collection and reporting periods. Traditional survey method perform surveys over the course of a single week per quarter, asking participants about their employment status during a single reference week. Unofficial survey results, subject to revision are then released a few weeks following the end of the quarter. Mobile phone data, however, is continually collected throughout the quarter and is available for analysis at any time during the period. Analysis of a given quarter can be performed and made available immediately following the end of the month.

Incoming:: The number of calls received by a user in a given month.

Outgoing:: The number of calls made by a user in a given month

Contacts:: The number of unique individuals contacted by a user each month. Includes calls made and received.

To Town:: The fraction of a user's calls made each month to another user who is physically located in the town of the plant closure at the time the call was made.

Churn:: The fraction of a user's contacts called in the previous month that was not called in the current month. Let C_t be the set of users called in month t. Churn is then calculated as: $churn_t = 1 - \frac{|C_{t-1} - C_t|}{|C_t - 1|}$.

Towers:: The number of unique towers visited by a user each month.

Radius of Gyration, R_g:: The average displacement of a user from his or her center of mass: $R_g = \sqrt{\frac{1}{n}\sum_{j=1}^{n} |\vec{r}_j - \vec{r}_{cm}|^2}$, where n is the number of calls made by a user in the month and r_{cm} is the center of mass calculated by averaging the positions of all a users calls that month. To guard against outliers such as long trips for vacation or difficulty identifying important locations due to noise, we only consider months for

users where more than 5 calls were made and locations that where a user recorded more than three calls.

Average Distance from Top Tower, R_1:: The average displacement of a user from their most called location: $R_1 = \sqrt{\frac{1}{n}\sum_{j=1}^{n} |\vec{r}_j - \vec{r}_1|^2}$, where n is the number of calls made by a user in the month and r_1 is the coordinates of the location most visited by the user. To guard against outliers such as long trips for vacation or difficulty identifying important locations due to noise, we only consider months for users where more than 5 calls were made and locations that where a user recorded more than three calls.

I. Measuring Changes

For each user i, we compute these metrics monthly. Because individuals may have different baseline behaviors, we normalize a user's time series to the month immediately before the layoff denoted t^*. To assess differences in behavior as a result of the mass layoff, we construct three groups: (1) A group of laid off users from the town where the probability of being laid off is that calculated in the previous section, (2) a town control group consisting of the same users as group 1, but with inverse weights, and (3) a group of users selected at random from the country population. Each user in the final group is weighted equally.

For each month, we compute the weighted average of all metrics then plot the difference between the laid off group and both control groups in Figure 3 in the main text.

$$y_t = \sum_i w_i \frac{y_{i,t}}{y_{i,t^*}} \tag{1}$$

$$\Delta y_t = \bar{y}_t - \bar{y}_{t,control} \tag{2}$$

We estimate changes in monthly behavior using OLS regressions. We specify two models that provide similar results. For a metric :

$$y_i = \alpha_i + \beta_1 A_i + \beta_2 U_i + \beta_3 A_i U_i \tag{3}$$

where A_i is a dummy variable indicating if the observation was made in a month before or after the plant closure and U_i is a dummy variable that is 1 if the user was assigned a

greater than 50% probability of having been laid off and 0 otherwise. An alternate model substitutes the probability of layoff itself, for the unemployed dummy:

$$y_i = \alpha_i + \beta_1 A_i + \beta_2 w_i + \beta_3 A_i w_i \tag{4}$$

In many cases, we are more interested in relative changes in behavior rather than absolute levels. For this, we specify a log-level model of the form:

$$\log(y_i) = \alpha_i + \phi_1 A_i + \phi_2 w_i + \phi_3 A_i w_i \tag{5}$$

Now the coefficient ϕ_3 can be interpreted as the percentage change in feature $y_{i,n}$ experienced by a laid off individual in months following the plant closure. Changes to mobility metrics as well as changes to total, incoming, and outgoing calls were estimated using the log-level model. Churn and To Town metrics are percentages already and are thus estimated using a level-level model. The changes in the number of contacts each also estimated using a level-level model.

Models are estimated using data from users believed to be unemployed and data from the two control groups. Results are shown in Table I. Comparisons to each group produce consistent results.

J. Predicting Province Level Unemployment

To evaluate the predictive power of micro-level behavioral changes, we use data from a different undisclosed industrialized European country. As discussed in the main text, we use call detail records spanning nearly 3 years and the entire user base of a major mobile phone provider in the country. For each of the roughly 50 provinces within this country, we assemble quarterly unemployment rates during the period covered by the CDR data. At the national level, we collect a time series of GDP. We select a sample of users in each province and measure the average relative value of 7 of the variables identified to change following a layoff. To-town and distance from home variables are omitted as the former is only measured when we know the location of the layoff and the latter is strongly correlated with R_g.

First, we correlate each aggregate calling variable with unemployment at the regional

level. To control for differences in base levels of unemployment across the country, we first demean unemployment and each aggregate variable. Table II shows that each calling behavior is significantly correlated with unemployment and that these correlations are consistent with the directions found in the individual section of the paper. Moreover, we discover strong correlation between each of the calling behavior variables, suggesting that principal component analysis is appropriate.

1. Principal Component analysis

As shown in the individual section of the paper, changes in these variables following a mass layoff are correlated. This correlation is seen in province level changes as well (Table II). Given this correlation, we use principal component analysis (PCA) extract an independent mobile phone variable and guard against co-linearity when including all phone variables as regressors. The results from PCA and the loadings in each component can be found in Table III and Table IV, respectively. We find only the first principal component passes the Kaiser test with an eigenvalue significantly greater than 1, but that this component captures 59% of the variance in the calling data. The loadings in this component fall strongly on the social variables behavior. We then compute the scores for this component for each observation in the data and use these scores as regressors. The prominent elements of the first principal component are primarily related to the social behavior of callers.

2. Model Specification

We make predictions of present and future unemployment rates using three different models specifications of unemployment where each specification is run in two variants, one with the principal component score as an additional independent variable denoted as CDR_t and the other without. The sixteen models are described as follows:

1. **AR(1)**

$$U_t = \alpha_1 U_{t-1} \tag{6}$$

$$U_t = \beta_1 U_{t-1} + \gamma CDR_t \tag{7}$$

$$U_{t+1} = \alpha_1 U_{t-1} \tag{8}$$

$$U_{t+1} = \beta_1 U_{t-1} + \gamma CDR_t \tag{9}$$

2. **AR(1) + Quad**

$$U_t = \alpha_1 U_{t-1} + \alpha_2 U_{t-1}^2 \tag{10}$$

$$U_t = \beta_1 U_{t-1} + \beta_2 U_{t-1}^2 + \gamma CDR_t \tag{11}$$

$$U_{t+1} = \alpha_1 U_{t-1} + \alpha_2 U_{t-1}^2 \tag{12}$$

$$U_{t+1} = \beta_1 U_{t-1} + \beta_2 U_{t-1}^2 + \gamma CDR_t \tag{13}$$

3. **AR(1) + GDP**

$$U_t = \alpha_1 U_{t-1} + \alpha_2 gdp_{t-1} \tag{14}$$

$$U_t = \beta_1 U_{t-1} + \beta_2 gdp_{t-1} + \gamma CDR_t \tag{15}$$

$$U_{t+1} = \alpha_1 U_{t-1} + \alpha_2 gdp_{t-1} \tag{16}$$

$$U_{t+1} = \beta_1 U_{t-1} + \beta_2 gdp_{t-1} + \gamma CDR_t \tag{17}$$

To evaluate the ability of these models to predict unemployment, we use a cross-validation framework. Data from half of the provinces are used to train the model and these coefficients are used to predict unemployment rates given data for the other half of the provinces. We perform the same procedure switching the training and testing set and combine the out of sample predictions for each case. We evaluate the overall utility of these models by plotting predictions versus observations, finding strong correlation (see the main text). To evaluate the additional benefit gained from the inclusion of phone data, we compute the percentage difference between the same model specification with and without the mobile phone data, $\Delta RMSE\% = 1 - RMSE_{w/CDR}/RMSE_{w/out}$ .In each case, we find that the addition of mobile phone data reduces the RMSE by 5% to 20%.

3. Predictions using weekly CDR Data

Until now, we have have used data from the entire quarter to predict the results from the unemployment survey conducted in the same quarter. While these predictions would be available at the very end of the quarter, weeks before the survey data is released, we also make predictions using CDR data from half of each quarter to provide an additional 1.5 months lead time that may increase the utility of these predictions. We estimate the same models as described in the previous section and find similar results. Even without full access to a quarter's CDR data, we can improve predictions of that quarter's unemployment survey before the quarter is over by 3%-6%.

4. The Effect of Sample Size on Feature Estimation

It is important to consider the extent to which the sampling size is sufficient and does not affect much the feature estimation. We study the reliability of sample size (k) in terms of relative standard deviation (RSD). For each given sample size k, we sample T times (without replacement) from the population. The RSD with respect to sample size k for a particular feature, is given by $RSD(k) = \frac{s_k}{f_k}$ where s_k is the standard deviation of the feature estimates from the T samples, and f_k is the mean of the feature estimates from the T samples. We use $T = 10$ to study the feature reliability. In Figure 5, we plot the different features' %RSD by averaging the RSD values of all provinces. The plots show that the values of %RSD over sample size $k = 100, 200, ..., 2000$ decrease rapidly. When sample size $k = 2000$, the %RSD for all features, except for radius of gyration (R_g), is lower than 1%. The estimates of R_g exhibit the highest variation; however, we can still obtain reliable estimates with thousands of sampled individuals ($RSD(k) = 0.026$ for R_g, with $k = 2000$). For the results in the manuscript, a value of k=3000 was chosen with the confidence that sample size effects are small.

K. Mass Layoffs and General Unemployment

While mass layoffs provide a convenient and interesting natural experiment to deploy our methods, they are only one of many employment shocks that economy absorbs each month. We have measured changes in call behaviors due to mass layoffs, but these changes may

TABLE I. Regression Results - Social and Mobility Measures.

	(1) Log(calls)	(2) Log(inc)	(3) Log(out)	(4) contacts	(5) to town	(6) churn	(7) Log(towers)	(8) Log(R_g)	(9) Log(R_1)
Panel A: Compared to Random User									
Post-Layoff Dummy	0.0390**	0.0464***	0.0448**	0.213***	-0.00000264	0.0201***	0.0179	0.0657*	0.0588*
	(0.0155)	(0.0170)	(0.0193)	(0.0124)	(0.000135)	(0.00592)	(0.0721)	(0.0337)	(0.0352)
Laidoff Dummy * Post-Layoff Months	-0.415***	-0.335***	-0.446***	-0.785*	-0.0478***	0.0368**	-0.171***	-0.226**	-0.262**
	(0.0679)	(0.0738)	(0.0747)	(0.458)	(0.0136)	(0.0167)	(0.0550)	(0.106)	(0.116)
Observations	10011	9742	9456	10011	10011	10011	10011	6922	6908
R-Squared	0.828	0.805	0.803	0.923	0.892	0.338	0.812	0.655	0.657
Panel B: Comparison to Non-laidoff Town Users									
Post-Layoff Dummy	0.0511***	0.0497***	0.0574***	0.454***	-0.000999	0.0301***	0.00973	0.0345**	0.0371**
	(0.0106)	(0.0122)	(0.0118)	(0.110)	(0.00188)	(0.00368)	(0.00935)	(0.0166)	(0.0175)
Laidoff Dummy * Post-Layoff Months	-0.517***	-0.416***	-0.545***	-1.311***	-0.0499***	0.0312*	-0.207***	-0.199*	-0.258**
	(0.0679)	(0.0738)	(0.0747)	(0.458)	(0.0136)	(0.0167)	(0.0550)	(0.106)	(0.116)
Observations	17506	17342	17118	17506	17506	17506	17506	15474	15417
R-Squared	0.875	0.860	0.871	0.938	0.889	0.349	0.899	0.729	0.741

All specifications include user fixed effects. Robust standard errors clustered by user.
*, **, and *** denote significance at the 10%, 5% and 1% levels.
Note that differences in sample sizes are due to zero values in Log transformations.

be unhelpful if they do not result from other forms of unemployment like isolated layoffs of individual works. Though it is beyond the scope of this work to directly determine if individuals affected by mass layoffs experience the same behavioral changes as those experiencing unemployment due to other reasons, we do find strong correlations between the number of mass layoffs observed in a given time period and general unemployment rates.

Using monthly data provided by the United States Bureau of Labor Statistics (BLS), Figure 6 shows time series of the number of monthly initial claimants of unemployment benefits do to any change in employment status and due to mass layoffs directly. There is similarly high correlation between the number of distinct mass layoff events (irrespective of the number of claimants in each event). While the relationship between claimants due to mass layoffs and the overall unemployment rate is not as strong, there is still significant correlation (see Table IX). Moreover, these positive correlations hold true at a state level as shown by Table X.

TABLE II. Correlation coefficients between normalized, aggregated calling behaviors and unemployment rates the province level. Pairwise correlations between metrics are done quarterly for each province giving a total of 416 observations for each variable comparison

	(1)							
	Calls	Inc	Out	Contacts	Churn	Towers	R_g	Unemp
Calls	1							
Inc	0.746***	1						
Out	0.952***	0.724***	1					
Contacts	0.807***	0.667***	0.858***	1				
Churn	0.102*	-0.141**	0.132**	0.183***	1			
Towers	0.701***	0.522***	0.711***	0.584***	0.196***	1		
R_g	0.373***	0.193***	0.379***	0.269***	0.177***	0.696***	1	
Unemp	-0.428***	-0.356***	-0.396***	-0.169***	0.138**	-0.418***	-0.295***	1

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE III. PCA results for call variables.

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	4.10	2.93	0.59	0.59
Comp2	1.17	0.31	0.17	0.75
Comp3	0.86	0.53	0.12	0.89
Comp4	0.34	0.04	0.05	0.93
Comp5	0.29	0.11	0.04	0.97
Comp6	0.19	0.15	0.03	0.99
Comp7	0.04	.	0.01	1.00

TABLE IV. PCA Loadings. Significant elements are bolded.

	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
Calls	**0.47**	-0.117	0.098	0.013	-0.147	**-0.568**	**0.643**
Inc	**0.39**	**-0.376**	-0.020	0.232	**0.794**	0.131	-0.054
Out	**0.47**	-0.095	0.142	0.008	-0.268	**-0.346**	**-0.746**
Contacts	**0.431**	-0.096	0.284	0.167	**-0.406**	**0.710**	0.156
Churn	0.10	0.726	**0.605**	0.073	0.298	-0.046	-0.002
Towers	**0.388**	0.247	**-0.308**	**-0.802**	0.125	0.179	0.015
R_g	0.252	**0.487**	**-0.652**	**0.517**	-0.065	0.015	-0.006

TABLE V. Predicting Present Unemployment Rates - Cross Validation Model Coefficients

	Model Estimates with out CDR Data					
	(1)	(2)	(3)	(4)	(5)	(6)
	AR1	AR1	AR1 Quad	AR1 Quad	AR1 GDP	AR1 GDP
U_{t-1}	1.123***	1.061***	1.195***	1.599***	1.072***	1.032***
	(0.0325)	(0.0456)	(0.152)	(0.139)	(0.0316)	(0.0441)
U_{t-1}^2			-0.301	-2.001***		
			(0.615)	(0.562)		
gdp_{t-1}					0.00119***	0.00103***
					(0.000169)	(0.000149)
_cons	-0.00180	0.00304	-0.00541	-0.0256***	-0.0244***	-0.0183***
	(0.00343)	(0.00407)	(0.00788)	(0.00748)	(0.00423)	(0.00430)
Observations	144	168	144	168	144	168
R^2	0.871	0.891	0.872	0.903	0.898	0.910
Adjusted R^2	0.871	0.891	0.870	0.902	0.897	0.909
	Model Estimates with CDR Data					
	(1)	(2)	(3)	(4)	(5)	(6)
	AR1	AR1	AR1 Quad	AR1 Quad	AR1 GDP	AR1 GDP
U_{t-1}	1.064***	1.076***	1.180***	1.395***	1.056***	1.062***
	(0.0330)	(0.0389)	(0.146)	(0.139)	(0.0323)	(0.0403)
CDR_t	-0.00597***	-0.00605***	-0.00602***	-0.00534***	-0.00313**	-0.00481***
	(0.000928)	(0.000842)	(0.000925)	(0.000917)	(0.00120)	(0.00120)
U_{t-1}^2			-0.486	-1.192*		
			(0.586)	(0.566)		
gdp_{t-1}					0.000855***	0.000390
					(0.000225)	(0.000211)
_cons	0.00434	0.00290	-0.00144	-0.0141	-0.0148*	-0.00516
	(0.00338)	(0.00350)	(0.00767)	(0.00750)	(0.00591)	(0.00539)
Observations	144	168	144	168	144	168
R^2	0.893	0.918	0.894	0.922	0.902	0.920
Adjusted R^2	0.892	0.917	0.891	0.920	0.900	0.918

Standard errors in parentheses

Coefficients are reported for models trained on each half of the data.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE VI. Predicting Future Unemployment Rates - Cross Validation Model Coefficients

	Model Estimates with out CDR Data					
	(1)	(2)	(3)	(4)	(5)	(6)
	AR1	AR1	AR1 Quad	AR1 Quad	AR1 GDP	AR1 GDP
U_{t-1}	1.161***	1.040***	1.351***	2.140***	1.074***	0.991***
	(0.0669)	(0.0871)	(0.329)	(0.268)	(0.0601)	(0.0839)
U_{t-1}^2			-0.852	-4.051***		
			(1.421)	(0.983)		
gdp_{t-1}					0.00220***	0.00194***
					(0.000213)	(0.000174)
_cons	0.00906	0.0164*	-0.0000247	-0.0428**	-0.0349***	-0.0256**
	(0.00730)	(0.00776)	(0.0165)	(0.0150)	(0.00668)	(0.00801)
Observations	96	112	96	112	96	112
R^2	0.703	0.730	0.704	0.773	0.813	0.812
Adjusted R^2	0.700	0.728	0.697	0.769	0.809	0.808
	Model Estimates with CDR Data					
	(1)	(2)	(3)	(4)	(5)	(6)
	AR1	AR1	AR1 Quad	AR1 Quad	AR1 GDP	AR1 GDP
U_{t-1}	1.074***	1.135***	1.490***	1.877***	1.058***	1.070***
	(0.0608)	(0.103)	(0.272)	(0.300)	(0.0583)	(0.114)
CDR_t	-0.0137***	-0.0117***	-0.0140***	-0.0102***	-0.00716**	-0.00697*
	(0.00159)	(0.00180)	(0.00154)	(0.00211)	(0.00235)	(0.00309)
U_{t-1}^2			-1.870	-2.775*		
			(1.229)	(1.314)		
gdp_{t-1}					0.00147***	0.00108**
					(0.000324)	(0.000380)
_cons	0.0208***	0.0122	0.00110	-0.0278	-0.0141	-0.00932
	(0.00598)	(0.00819)	(0.0130)	(0.0155)	(0.00965)	(0.00655)
Observations	96	112	96	112	96	112
R^2	0.801	0.814	0.805	0.833	0.828	0.825
Adjusted R^2	0.797	0.811	0.798	0.828	0.822	0.821

Standard errors in parentheses

Coefficients are reported for models trained on each half of the data.

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE VII. RMSE with the addition of CDR data from the entire quarter with various fixed effects. The best performing model is bolded.

	No F.E.			Quarter F.E.			Province F.E.		
Model	RMSE	RMSE w/ CDR	ΔRMSE%	RMSE	RMSE w/ CDR	ΔRMSE%	RMSE	RMSE w/ CDR	ΔRMSE%
Present									
AR1	0.0192	0.0168	12.37	0.195	0.0179	8.18	0.0217	0.0171	21.34
AR1 Quad	0.0191	0.0168	11.88	0.0190	0.0178	6.23	0.0218	0.0171	21.18
AR1 GDP	0.0173	**0.0166**	3.93	0.0178	0.0176	0.89	0.0179	0.0168	6.24
Future									
AR1	0.0315	0.0257	18.45	0.0315	0.0281	10.76	0.0365	0.0318	12.92
AR1 Quad	0.0308	0.0262	14.89	0.0302	0.0285	5.55	0.0358	0.0314	12.21
AR1 GDP	0.0260	**0.0246**	5.37	0.0270	0.0269	0.61	0.0274	0.0313	-14.12

TABLE VIII. RMSE with the addition of CDR data from the half of the quarter with various fixed effects. The best performing model is bolded.

	No F.E.			Quarter F.E.			Province F.E.		
Model	RMSE	RMSE w/ CDR	ΔRMSE%	RMSE	RMSE w/ CDR	ΔRMSE%	RMSE	RMSE w/ CDR	ΔRMSE%
Present									
AR1	0.0192	0.0164	14.04	0.0194	0.0172	11.12	0.0217	0.0165	23.82
AR1 Quad	0.0190	**0.0163**	14.19	0.0189	0.0170	10.01	0.0217	0.0164	24.58
AR1 GDP	0.0173	0.0165	4.49	0.0177	0.0175	1.08	0.0179	0.0165	7.51
Future									
AR1	0.0314	0.0258	17.80	0.0315	0.0273	13.07	0.0365	0.0267	26.73936
AR1 Quad	0.0307	**0.0246**	19.77	0.0301	0.0262	13.02	0.0357	0.0256	28.36
AR1 GDP	0.0259	0.0258	0.50	0.0270	0.0270	-0.046	0.0274	0.0272	0.49

TABLE IX. Correlations between general unemployment and unemployment resulting from mass layoffs.

	Correlation Coefficient
Mass Layoff Events vs. Total Initial Unemployment Claimants	0.86
Initial Claimants due to Mass Layoffs vs. Total Initial Unemployment Claimants	0.73
Initial Claimants due to Mass Layoffs vs. Unemployment Rate	0.46

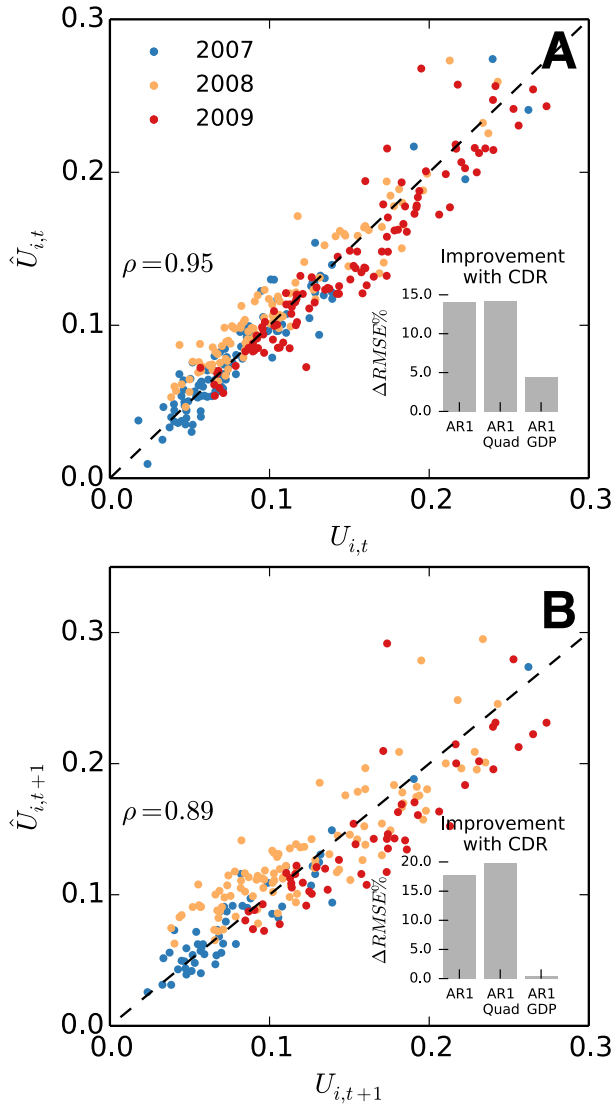

FIG. 4. Predicting unemployment rates using mobile phone data from only the first 6 weeks of each quarter. We follows the same procedure as the main text. Panel A compares predictions of present unemployment rates to observed rates and Panel B shows predictions of unemployment one quarter ahead using a simple AR1 model that includes co-variates of behaviors measured using mobile phones. Both predictions correlate strongly with actual values while changes in rates are more difficult to predict. The insets show the percent improvement to the RMSE of predictions when mobile phone co-variates are added to each of four traditional forecasting models. In general, mobile phone data reduces forecast errors by 3% to 6%.

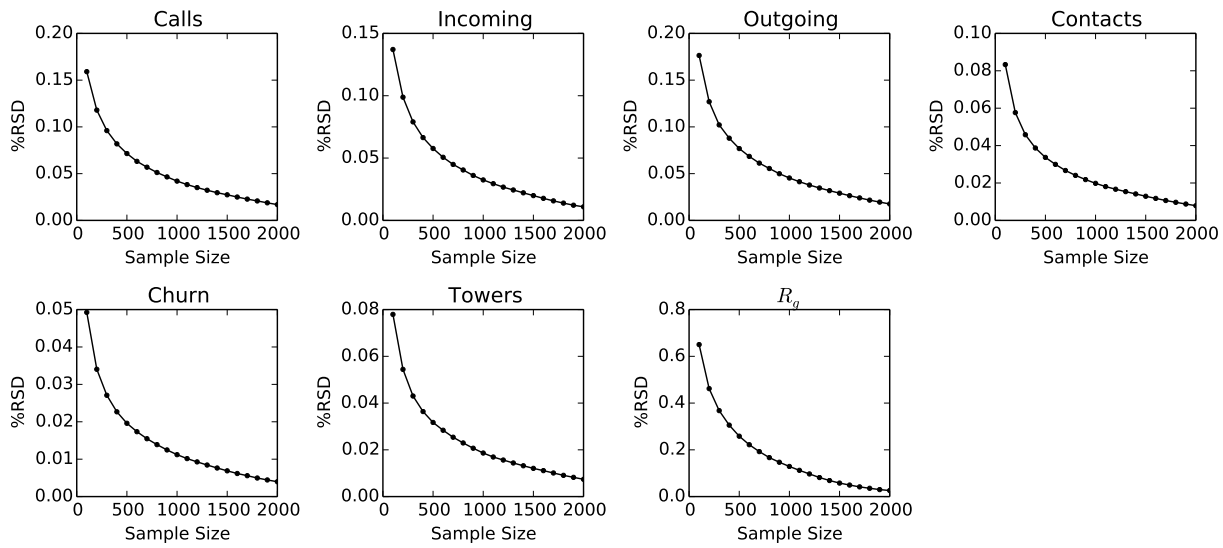

FIG. 5. The average values of %RSD against the number of samples per province for different features. For all features, the %RSD's decrease rapidly with sample size and stabilize to relative small values before $k = 2000$.

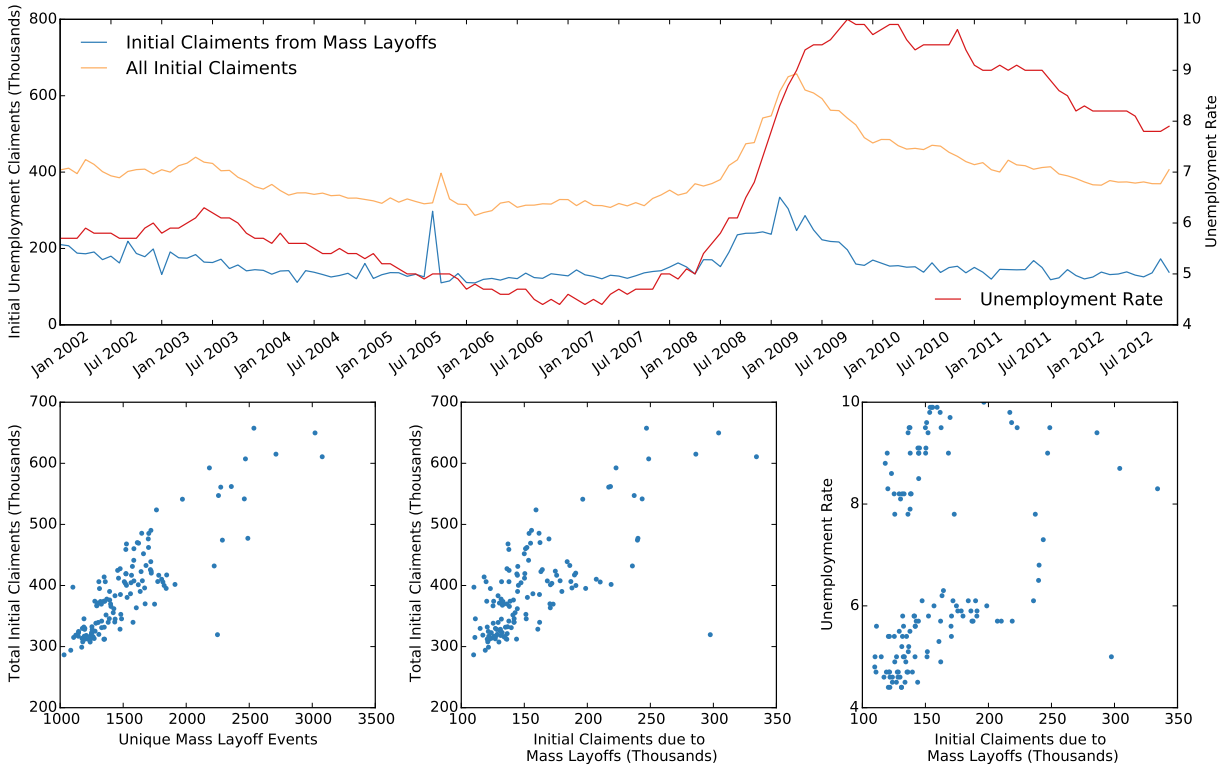

FIG. 6. Correlations between mass layoff events and general unemployment. Using BLS data, we plot various correlations between the number of mass layoff events, the number of initial unemployment claimants due to these events, general unemployment claims, and the unemployment rate. We find strong correlation between all of these variables suggesting that mass layoffs are a good proxy for general unemployment shocks, at least at the predictive level.

TABLE X. Correlations between mass layoff and unemployment and the state level.

State	Mass Layoff Events vs. Unemp. Rate	Mass Layoff Events vs. Unemp. Claims	Mass Layoff Claims vs. Unemp. Rate	Mass Layoff Claims vs. Unemp. Claims
Alabama	0.27	0.27	0.24	0.24
Alaska	0.13	0.23	0.04	0.17
Arizona	0.36	0.35	0.21	0.21
Arkansas	0.37	0.38	0.25	0.27
California	0.44	0.42	0.30	0.27
Colorado	0.40	0.39	0.35	0.33
Connecticut	0.27	0.28	0.15	0.16
Delaware	0.41	0.40	-0.22	-0.23
District of Columbia	-0.38	-0.36	-0.46	-0.50
Florida	0.69	0.70	0.67	0.67
Georgia	0.42	0.42	0.30	0.30
Hawaii	0.46	0.46	0.37	0.36
Idaho	0.31	0.31	0.32	0.28
Illinois	0.48	0.49	0.50	0.51
Indiana	0.33	0.34	0.19	0.20
Iowa	0.23	0.24	0.18	0.19
Kansas	0.33	0.34	0.25	0.25
Kentucky	0.48	0.48	0.31	0.32
Louisiana	0.44	0.45	0.43	0.43
Maine	0.08	0.08	0.06	0.05
Maryland	0.10	0.12	0.03	0.04
Massachusetts	0.12	0.12	0.09	0.09
Michigan	0.16	0.16	0.09	0.09
Minnesota	0.25	0.25	0.10	0.10
Mississippi	0.29	0.30	0.26	0.28
Missouri	0.31	0.31	0.12	0.11
Montana	0.34	0.38	0.26	0.27
Nebraska	0.15	0.13	0.11	0.10
Nevada	0.65	0.64	0.43	0.41
New Hampshire	0.30	0.28	0.11	0.08
New Jersey	0.28	0.28	0.17	0.17
New Mexico	0.25	0.32	0.08	0.16
New York	0.34	0.36	0.28	0.29
North Carolina	0.51	0.47	0.44	0.41
North Dakota	0.30	0.31	0.10	0.12
Ohio	0.25	0.25	0.15	0.15
Oklahoma	0.27	0.27	0.11	0.10
Oregon	0.39	0.40	0.33	0.33
Pennsylvania	0.40	0.41	0.33	0.34
Puerto Rico	0.17	0.22	0.01	0.05
Rhode Island	-0.07	-0.08	-0.19	-0.19
South Carolina	0.29	0.27	0.11	0.08
South Dakota	0.60	0.58	0.15	0.12
Tennessee	0.42	0.43	0.35	0.36
Texas	0.51	0.45	0.39	0.32
Utah	0.38	0.43	0.34	0.37
Vermont	0.31	0.32	0.22	0.23
Virginia	0.44	0.39	0.20	0.15
Washington	0.42	0.42	0.39	0.35
West Virginia	0.37	0.38	0.26	0.27
Wisconsin	0.34	0.34	0.27	0.26
Wyoming	-	-	-	-