Schaffner, Florian

Working Paper

Predicting US bank failures with internet search volume data

Working Paper, No. 214

Provided in Cooperation with:
Department of Economics, University of Zurich

Suggested Citation: Schaffner, Florian (2015) : Predicting US bank failures with internet search volume data, Working Paper, No. 214, University of Zurich, Department of Economics, Zurich, https://doi.org/10.5167/uzh-118268

This Version is available at:
https://hdl.handle.net/10419/126603

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

Working Paper No. 214

Predicting US Bank Failures with Internet Search Volume Data

Florian Schaffner

December 2015

Predicting US Bank Failures with Internet Search Volume Data*

Florian Schaffner
Department of Economics, University of Zurich

This version: December 2015

Abstract

This study investigates how well weekly Google search volumes track and predict bank failures in the United States between 2007 and 2012, contributing to the expanding literature that exploits internet data for the prediction of events. Different duration models with time-varying covariates are estimated. Higher Google search volumes go hand in hand with higher failure rates, and the coefficients for the Google volume growth index are highly significant. However, Google's predictive power quickly dissipates for future failure rates.

Keywords: *Bank failures, Internet, Financial crisis, Google, Survival analysis*
JEL: *G170, G180, G190, G210, G280*

*I thank Michel Habib, Steven Ongena, Rainer Winkelmann, Raphael Studer as well as the participants of the Zurich Workshop on Economics 2013 for helpful comments and suggestions.
Address for correspondence: University of Zurich, Department of Economics, Zürichbergstrasse 14, CH-8032 Zurich, Switzerland, Tel. +41 44 634 22 97, florian.schaffner@econ.uzh.ch.

1 Introduction

When US-Senator Chuck Schumer publicly questioned the financial health of the bank IndyMac in the summer of 2008, the bank's customers were quick to react. Within just three days, IndyMac lost USD 100 million in deposits (Los Angeles Times, 2008); after thirteen days and withdrawals amounting to USD 1.3 billion, the bank failed (Grind, 2012; Seabrook, 2008). These developments were closely tracked by the Google search volume index; on the day of Schumer's announcement, the index value for the search term "IndyMac" almost doubled, rising further up to 22-fold in the following days. During the same period Washington Mutual, a struggling competitor, lost a total of USD 9.4 billion due to a bank run, even surpassing IndyMac's deposit loss. Contrary to the much publicized IndyMac incident, the Washington Mutual run was largely unnoticed by the media or the analysts (Grind, 2012). The run did not escape the Google search volumes and Washington Mutuals share price, however: As in the IndyMac case, the search volume index for "Washington Mutual" more than doubled during the days of the run. The peak of the bank run on Tuesday, July 14, 2008 coincided with the high search volume on that same day. Two months later, Washington Mutual went through another bank run, with its peak on Thursday, September 18. Again, Google index values track a total outflow of approximately USD 16.7 billion between September 15 and September 24 (Office of Thrift Supervision, 2008).[1] Search volumes were surpassed only on September 25, 2008, the day when the Federal Deposit Insurance Company (FDIC) walked into Washington Mutual's offices and shut the bank down (see Figure 1).[2]

[Figure 1 about here.]

These two examples provide anecdotal evidence that Google search volumes can be a valuable proxy to reflect public attention, which is generally hard to capture. It has been shown in a variety of settings that Google can be instrumented to reflect such phenomena, from influenza epidemics to unemployment forecasting (Ginsberg et al., 2008; Breyer et al., 2011; McCarthy, 2010; Bollen et al., 2011). Whether it is useful for tracking developments in the banking industry is the subject of this paper.

Contrary to, say, a newspaper article, internet search volumes reflect a much more crowd-sourced and democratic approach. Users are actively looking for something rather than consuming

[1] Although there was extensive media coverage on the bank's health before its closure, the public was informed about the bank run only after the fact (Grind, 2012).

[2] The low index values occurring in regular intervals are typically weekend days - days when individuals spend less time on their computers and bank transactions cannot be executed.

information passively. While a newspaper article about a specific bank must be considered worthwile reporting in the first place and, in a second step, restricted to reflecting the facts, Google search queries can capture a much wider set of information: facts as well as speculations, banking experts as well as individual savers. Google queries also give an idea of how many individuals care about a specific topic. Each search query is an uptick in the volume, translated into a rising index value. In that sense, it is similar to trading volumes in financial markets (see also Mathiesen et al. 2013) - except that for the majority of the banks considered in this paper, such volumes do not exist because many of them are not listed on a stock exchange. In the absence of share prices, Google data may therefore prove to be a valuable source for understanding and predicting bank failures. Providing real-time data on the popularity of a bank's name on the internet on a weekly basis, Google can help modelling short term dynamics by incorporating information that is not fully captured in balance sheet positions or macro-level variables. Such information might be important: The deposit withdrawal at Washington Mutual in July 2008 is what Iyer et al. (2013) call a non-fundamental shock: a run that cannot be justified by the balance sheet fundamentals of the bank itself or that couldn't have already been justified at an earlier point in time. At the time such a shock would have become visible in Washington Mutual's balance sheet, the bank was already under the reign of JP Morgan. Google tracked that non-fundamental shock in a timely manner. Google data is also an interesting addition to fundamentals in light of the Iyer et al. (2013) finding that large depositors tend to orient themselves to and act on (possibly non-public) regulatory actions rather than fundamentals; Google might partly capture these movements.

Three main questions are tackled in this paper. Since Google data is not available for all banks (discussed in section 3.1), a first question seeks to answer whether the availability of Google data itself correlates with the survival of an individual bank. Second, given that Google data is available, the question of how well the Google query shares track bank failures is examined. Thirdly, I discuss the question of how indicative past changes in search volumes are when trying to predict future failures.

To answer these questions, this paper looks at 433 bank failures and 400 surviving banks in the United States in the period from January 2007 to March 2012, working with a dataset including both Google data as well as balance sheet and revenue data on the level of an individual bank. Using an exponential duration model with a piecewise-constant hazard rate and time-varying covariates, I analyze how well Google search volumes in the United States track and predict these bank failures. Results show that while the availabilty of Google data itself has no significant effect on a bank's survival, higher Google search volumes go hand in hand with higher hazard rates. As

one moves further away from the failure date, Google's predictive power dissipates.

The paper is structured as follows: in the following section, previous findings are discussed. In the third section, the data is presented. In the fourth section, I model the failure rates of individual banks, using weekly Google time series and balance sheet positions and revenue data from the FDIC as explanatory variables. Section 5 concludes.

2 Previous Literature

Working with Google data to model short-term developments has been successful before. From influenza epidemics (Ginsberg et al., 2008) to tracking kidney stone incidences (Breyer et al., 2011) and monitoring suicide risks (McCarthy, 2010), on to more economic applications in the field of unemployment (Askitas and Zimmermann, 2009; D'Amuri and Marcucci, 2010; Choi, 2009; Tefft, 2011), inflation (Guzman, 2011), consumer behavior (Choi and Varian, 2012; Goel et al., 2010), consumer sentiment (Radinsky et al., 2008; Della Penna and Huang, 2009; Preis et al., 2010) and housing prices (McLaren and Shanbhogue, 2011; Wu and Brynjolfsson, 2013). Financial markets have received some attention, too: Preis et al. (2013) quantified trading behavior using Google, Bollen et al. (2011) predict stock market movements using Twitter, Mathiesen et al. (2013) likened the statistical properties of Twitter data to the properties of trading volumes of stocks and Moat et al. (2013) studied the correlation of Wikipedia page views and stock market movements. To my knowledge, no paper to date has used Google search volumes to predict bank failures.

While there has been a variety of empirical work studying both wider banking panics as well as individual bank failures, this literature has focussed on balance sheet positions and revenue data, looking at issues of panics, contagion and information networks (for an overview, see Gorton and Winton, 2003). Calomiris and Mason (2003) analyze bank failures in the 1920's and 1930's using a duration model and data on individual banks as well as regional economic factors, disputing the Friedman-Schwartz argument that many bank failures resulted from unwarranted panic and finding evidence that most of the failures are justified by weak fundamentals. Saunders and Wilson (1996) look at the role of bank contagion and information in the same period, using data on deposit flows. Wheelock and Wilson (2000) make use of duration models to determine the effect of managerial inefficiency on the probability of failure and acquisition. Whalen (1991) assesses the usefulness of using proportional hazard models as early warning tool, concluding that "reasonably accurate early warning models can be built and maintained at relatively low cost."

Short term dynamics and irrational elements leading bank failures have proven difficult to

account for. Regarding bank runs on individual banks and micro-level withdrawal patterns, there exists only a small literature. A recent one is Iyer and Puria (2012), which looks at the dynamics of withdrawal patterns, deposit insurance and social networks in an Indian bank. A follow-up study (Iyer et al., 2013) looks at how depositors monitor banks, finding that regulatory agencies play an important role in the monitoring process. Other examples in the area of individual failures and information networks include Kelly and O Grada (2000) or O Grada and White (2003).

From a macro perspective, Donaldson (1992) finds that there are periods when banking panics are more likely to occur, but that exact starting dates of such panics are unpredictable. Gorton (1988) offers empirical evidence compatible with the idea that when depositors receive information forecasting a recession, they draw on their bank accounts, knowing that they will be dissaving and anticipating the higher bank failure rate during recessions. I try to take such factors into account by including macro-level variables.

3 Data Sources, Data Properties and Descriptive Statistics

As of December 2006, there were 8,681 active banks insured by the FDIC, compared to 7,357 at the end of 2011. Within these five years, 433 FDIC-insured banks failed.[3] In a first sample, I include the 433 failed banks in the period from January 1, 2007 to March 31, 2012. In addition, I randomly select a subset of 400 banks from the set of 7,357 active banks at the end of 2011 to include in the sample as control observations.[4] Focussing on a random sample of surviving banks instead of using the full sample is a result of the data collection procedure: As each query on Google needs to be executed manually, collecting data on the whole set of surviving banks is infeasible.

To restore adequate proportions between failures and survivors, I weight observations accordingly when estimating the models (discussed in Section 4). To have an equal entering date for all banks at risk of failure and to avoid complications when weighting observations (also discussed in Section 4), 18 banks founded after January 1st, 2007 were dropped, of which 5 were failed banks. For the resulting 815 banks, weekly Google search queries data and quarterly FDIC data was downloaded. The sample period covers 273 weeks or 21 quarters.

[3] Note that aside from failures, there also were mergers as well as newly founded banks.

[4] None of these randomly selected banks were merged into other banks or failed up to the first quarter of 2012. Sampling was done at the end of 2011 rather than at the end of the first quarter 2012 since data on the first quarter of 2012 was only added at a later stage.

3.1 Data Sources: Google Insights for Search

Weekly search query time series containing the bank's name have been executed and downloaded on "Google Insights for Search" (Google, 2012), Google's tool to analyze search volumes.[5] These time series reflect the query share of the bank's name in the overall search traffic categorized as "Finance" on a weekly basis. The structure and properties of Google data deserves some extra attention, as it has some non-standard restriction features.

The first restriction concerns the time horizon: Google time series go no further back than January 1, 2004 (Choi and Varian, 2012). There is no data available before that date.

Second, Google only publishes relative numbers, not absolute search volume numbers. The numbers are relative in two dimensions. First, the query share QS_{ijut} is the ratio of the number of queries n_{ijut} for a given search term i and the total number of queries N_{jut} in the selected category j in geographic area u at time t :

$$QS_{ijut} = \frac{n_{ijut}}{N_{jut}}, \quad 0 \leq QS_{ijut} \leq 1$$

The second dimension concerns the time series of the query share itself. All query shares are reported relative to the maximum query share M_{ijuS} in the selected period S multiplied by 100, which gives the Google index value GI_{ijut}:

$$GI_{ijut} = \frac{QS_{ijut}}{M_{ijuS}} \times 100,$$

$$\text{with} \quad M_{ijuS} = \max_{t \in S} QS_{ijut}, \quad 0 \leq GI_{ijut} \leq 100$$

GI_{ijut} is the number published by Google; all other numbers are not published. Under the assumption that internet usage is growing, a rising index value can always be interpreted as a rise in popularity for the search term. This is not true for falling values, as it is enough for the search term to be growing at a less than average rate in order for the index value to fall. Growth rates in query shares from t to $t+1$ are preserved in the published relative numbers, whereas percentage point differences are not. The levels of the index values are not comparable across banks. For these reasons, only Google growth rates are used in this paper.

Third, queries are "broad matched", meaning that queries such as "IndyMac bank run" are counted in the calculation of the query index for "IndyMac", but not vice-versa. Entering less and more general search terms increases the probability that unrelated searches are captured as well. For example, a query with the search term "forecast" may capture results related to forecasts of

[5] "Google Insights for Search" has been renamed to "Google Trends" in the meantime.

economic indicators, election results or weather, whereas a query for "weather forecast tomorrow Zurich" is much more specific and unlikely to include unrelated queries.

Fourth and linked to the third restriction, Google series for more restrictive queries are more likely not to be published at all. As mentioned above, Google publishes the index values only if the absolute number of search queries exceeds an unknow threshold (Choi and Varian, 2012). This has two consequences: First, it restricts the sample from 815 to 210 banks for which any Google data is available. Second, within the remaining 210 banks, the absolute search volume might temporarily fall under the threshold and a value of zero is published. Since the true index value is greater than or equal to zero, using these time series can bias estimation results. Focusing only on the complete cases with uncensored Google series, on the other hand, reduces the population to 25 failing banks and 23 surviving banks.[6]

[Table 1 about here.]

Google data is retrieved for the period of January 4, 2004 to March 31, 2012. Time series are on a weekly basis. Queries are restricted to the United States and to the "Finance" category to avoid counting unrelated queries in the index.[7] Queries outside the United States are unlikely to be related to the individual banks, while narrowing the geographic space to state levels would have resulted in more censored time series. A similar logic applies to the categoric restriction to "Finance": With a broader definition, unrelated queries might be captured in the index, while a narrower definition might exclude relevant queries or leads to censoring.

For each bank, there is a separate Google query containing the bank's name as a search term. For practical reasons, legal appendices such as *FSB*, *NA*, *National Association* or *Company* as well as "The" and "&" in bank names have been dropped, as it is unlikely that individuals search for their bank with legal appendices or include symbols such as "&".[8] Likewise, missing spaces (such as in *WashingtonFirst*) have been inserted. As for the case of popular bank names, there are three institutions named "First State Bank", two "The First State Bank", two "Premier Bank", two "Summit Bank", two "The Park Avenue Bank", two "Legacy Bank", two "First National Bank", two "Citizens National Bank" and two "Integrity Bank" in the sample.[9] In these cases,

[6] When calculating percentage changes for a Google series that was censored in the preceding period, the value was set to missing.

[7] Google classifies queries into about 30 categories at the top level and about 250 categories at the second level using a natural language classification engine (Choi and Varian, 2012).

[8] The exact names used for the queries are stored in the *Search_name* variable - a missing value means that the name has been used without any modification.

[9] Note that the "The" in bank names was dropped when Google data was downloaded, i.e. effectively there are five banks named "First State Bank".

identical Google query time series have to be used, as one cannot differentiate and assign unique series to each institution.

3.2 Additional Data Sources

The second major data souce for this paper is the FDIC database (Federal Deposit Insurance Corp., 2011). The FDIC provides a large set of balance sheet positions, revenue figures and other characteristics of individual banks, which a number of researchers have used for similar estimations. For the purpose of this paper, 11 variables were selected and downloaded on the bank level on a quarterly basis, the shortest time interval available. The variables can be broadly classified in the categories capital adequacy, asset quality, earnings, liquidity and other factors. The selection of the variables was guided by the selections in previous research papers estimating similar models (e.g. Cole and Gunther 1995; Calomiris and Mason 2003; Wheelock and Wilson 2000). In addition, an indicator variable for the FDIC insurance limit raise from USD 100'000 to USD 250'000 on October 3, 2008 was defined. The FDIC dataset comprises 836 institutions.

Bloomberg serves as an additional data source from which weekly LIBOR and overnight indexed swap (OIS) time series were downloaded. 2010 US Census data (United States Census Bureau, 2010) was used to define urban area dummy variables on the US county level.

3.3 Variables and Summary Statistics

Information on bank failures is taken from the FDIC, which lists failures in its failed banks list. The failure date is defined as the closing date that the FDIC lists on that same list. The FDIC has some discretion when it comes to the exact date of the closing, and therefore to exploit the weekend days to wind down a bank (i.e. when banks are closed), most of these closing dates are on a Friday. For my purposes, this means I can aggregate these closing dates to a weekly measure with little loss of information. The failure time is then defined as the week into which the closing date falls. Table 2 lists the number of failures in a given year. As one can see from the table, most of the failures occur in the years after 2006. With respect to survival analysis, there is little information in the years 2004 to 2006 since there are almost no failures. In addition, these failures are unlikely to be connected to the financial crises. I therefore dropped the years 2004 to 2006.[10]

[Table 2 about here.]

[10] I did run the analysis including these years as a robustness check, without any meaningful changes in the results.

Summary statistics are presented in Table 3. Aside from the Google variable, several balance sheet variables are listed, which can be roughly categorized into a capitalisation variable (capital), asset quality variables (troubled assets, commercial real estate, residential real estate), earnings (net income), liquidity (large CDs, insdep, securities) and miscellaneous factors (insider loans, holding company, entering age, urban). A description of the variables can be found in Table A1 in the appendix. The reported values in the table are averages over the period starting in January 2007 to March 2012 or the respective failure date where in a first step, the average over all periods is taken for each bank, and then the average is taken over all banks in the group (i.e. failures/survivors). The upper third includes all banks, the middle third only banks where Google data is available, and the bottom third only banks with uncensored Google series.

Even if averaged over time, survivors and failures differ in some of the variables, as can be seen by the stars indicating a difference in the Wilcoxon rank-sum test at the one percent significance level. Differences attenuate somewhat as one restricts the sample to banks having uncensored Google series, which is the sample main results will be based on. The differences in capital ratio or large cash deposits, for example, are not significant anymore. If you exclude the two largest failures, Washington Mutual and IndyMac, from the sample, the difference in gross assets is not significant anymore either. Surviving banks differ from their failing counterparts with respect to troubled assets, net income, securities and age.

[Table 3 about here.]

In terms of changes in Google query shares during the last weeks prior to failure, other bank failures resemble the pattern of Washington Mutual seen in the introduction. Figure 2 shows the weekly mean of the growth rate of Google query shares for the names of the 25 failed banks with uncensored Google series, compared to the corresponding means of the 23 surviving banks. To calculate the value for the control group, control group values were averaged in the corresponding week to failure for the failing bank. In a second step, these values were averaged over all failing banks. Values for failing banks remain on a low level up to five weeks before failure. From then on, there is a slight upward trend in rates, up to about one week before failure, when they spike and remain high in the weeks of and after the failure. Shortly after the failure, rates drop sharply. Meanwhile, changes in query shares for surviving banks stay constant.

[Figure 2 about here.]

Figure 3 shows the quarterly means of different key balance sheet positions of failed banks in the last quarters before failure, again contrasted by the same statistics for surviving banks (the values were calculated analogously to the Google values above). Note that the horizontal axis is measured in quarters as opposed to weeks in the graph before. One can see clear trends in capital ratios and troubled assets ratios that start out at least one year before failure. Ratios for large deposits and securities stay relatively constant over time, but show clear differences across the failing and the surviving group. Comparing these graphs suggests that fundamentals of failing banks deteriorate early, while surviving banks' advantageous securities and large deposit positions protect them when having to react to liquidity drains. Google search queries, on the other hand, react when failure is imminent, correlating with the timing of failure rather than with the probability of failure itself.

[Figure 3 about here.]

4 Model and Results

4.1 Model

I use a piecewise-constant exponential model to model bank failures, estimating the hazard rate semi-parametrically. Using a piecewise-constant hazard as opposed to a parametric model such as the exponential or Weibull has the advantage of modelling the baseline hazard semi-parametrically. This is important as the baseline hazard, i.e. the hazard common to all banks, is likely to change over time, especially during the financial crisis. To account for the changes in the hazard rate over time and work with time-varying covariates, the dataset is split into 273 weekly episodes.

The baseline hazard is modelled using time dummies as well as macro-variables (the LOIS spread). With respect to time dummies, three specifications will be used. The first involves splitting the 2007 to 2012 into just two subperiods: one before and one after the raise of the FDIC insurance limit from USD 100'000 to USD 250'000 in October 2008. This intervention is mainly an intervention to prevent potential bank runs from depositors; whether the hazard changes can be directly tested on the corresponding dummy variable for the intervention. A second specification uses yearly time dummies, changing the baseline hazard every year. A third specification uses quarterly dummies. As an alternative to the piecewise-constant hazard model and as a robustness check, I also estimate a Cox proportional hazard model. Note that in this case, the baseline hazard function is completely unspecified.

In the piecewise-constant hazard model, the hazard rate is a step function specified as

$$\begin{aligned}\theta(t, \mathbf{x}_{it}, \mathbf{z}_i, \mathbf{w}_t) &= \theta_0(t)\lambda(\mathbf{x}_{it}, \mathbf{z}_i, \mathbf{w}_t) \\ &= \bar{\theta}_t \exp(\beta' \mathbf{x}_{it} + \delta' \mathbf{z}_i + \gamma' \mathbf{w}_t) \\ &= \exp[\log(\bar{\theta}_t) + \beta' \mathbf{x}_{it} + \delta' \mathbf{z}_i + \gamma' \mathbf{w}_t] \\ &= \exp(\tilde{\lambda}_t)\end{aligned}$$

where $\bar{\theta}_t$ is the interval-specific baseline hazard common to all banks and $\lambda(\mathbf{x}_{it}, \mathbf{z}_i, \mathbf{w}_t)$ is the bank-specific hazard component in period t. $\mathbf{x}_{it}$ is a vector including individual time-varying covariates, $\mathbf{z}_i$ contains individual time-constant covariates and $\mathbf{w}_t$ contains common, time-varying elements at time t. The interval-specific baseline hazard is equivalent to including a period-specific dummy variables in the overall hazard.

In the case of two subperiods with $\mathbf{x}_{it} = \mathbf{x}_{i1}$ and $\mathbf{w}_t = \mathbf{w}_1$ if $t < s$ and $\mathbf{x}_{it} = \mathbf{x}_{i2}$ and $\mathbf{w}_t = \mathbf{w}_2$ if $t \geq s$, the corresponding survivor function is given by (see Jenkins, 2005)

$$\begin{aligned}S(t, \mathbf{x}_{it}, \mathbf{z}_i, \mathbf{w}_t) &= [S_0(s)]^{\tilde{\lambda}_1} \frac{[S_0(t)]^{\tilde{\lambda}_2}}{[S_0(s)]^{\tilde{\lambda}_2}} \\ &= \exp(-s\tilde{\lambda}_1) \exp\left[-(t-s)\tilde{\lambda}_2\right]\end{aligned}$$

Note that Google data varies on a weekly basis, while balance sheet data varies only quarterly.

As mentioned in the data section, the sample consists of all banks that failed in the period from January 2007 to March 2012, plus a random sample of surviving banks. While in the sample of 815 banks, more than half of them fail within the roughly eight years covered, these failures represent only about five percent of the whole bank population of 8,681 institutions in December 2006. This choice-based sampling therefore needs to be accounted for by weighting observations accordingly (Lancaster, 1992). Individual likelihood contributions are weighted by P/Q, where P represents the population fraction of failing institutions, and Q represents the sample fraction of failing institutions (correspondingly, $(1-P)$ and $(1-Q)$ are the weights for surviving institutions). Accordingly, failing banks get a lower weight than surviving banks. I reported both the absolute number of failures as well as the weighted failures in the result tables.

4.2 Results

In this section, I seek to answer the three questions raised in the introduction. The empirical hazard rate including all 815 banks is displayed in Figure 4. One can see that the hazard rate varies with time, peaking after week 170, or at the beginning of 2010. The result tables are split into five columns; the model in the first column is using Google as the only explanatory variable. The second column shows results for a piecewise-constant hazard rate model with two subperiods (pre- and post FDIC insurance limit raise), followed by the piecewise-constant hazard models with yearly and quarterly dummies. The last column shows the results of the Cox proportional hazards model. The bottom of the table specifies whether the Google data availability variable (Dummy) or the Google growth variable for the percentage change from the last period (Growth) was used. The table also lists the total number of banks, the absolute number of failures, and the weighted number of failures (which is around five percent of the total number of banks, as outlined in the previous section). Note that the majority of the explanatory variables are roughly bounded between 0 and 100, as they are percentages of gross assets. The tables report coefficients (as opposed to hazard ratios). A change in X_k changes the overall hazard by $\frac{\partial\theta(t,\mathbf{X_t},\mathbf{Z},\mathbf{W_t})}{\partial X_{kt}} = \theta(t,\mathbf{X_t},\mathbf{Z},\mathbf{W_t})\beta_k$ or increases the hazard by $100(\exp(\beta_k)-1)$ percent (approximatly $100\beta_k$ percent). A negative coefficient decreases the hazard accordingly.

[Figure 4 about here.]

Table 4 shows results using the Google dummy variable. The coefficient for the Google variable is positive, but remains statistically insignificant in all five models - whether Google data is available is not a significant predictor whether a bank fails or not. Capital has the anticipated, large negative effect on the hazard rate. Troubled assets positions increase the hazard rate as one would expect, while securities - which can serve as collateral when lending money - decrease the hazard, as do large cash deposits. Interestingly, the coefficient for the variable *insdep*, the interaction between the large deposits ratio and the FDIC insurance limit raise dummy, is positive, implying a relatively higher hazard for banks with large deposits after the FDIC intervention. Note that the coefficient on the FDIC intervention (the subperiod dummy) counteracts the effect with a negative coefficient of about the same magnitude (not shown in the table). Finally, coefficients on commercial real estate and *urban* are positive, while the coefficient for holding companies is negative. The remaining effects are not statistically significant.

[Table 4 about here.]

Table 5 presents the main results using only banks with uncensored Google series. Generally, effects increase the more flexible the baseline hazard is specified, with the exception of the macro-variable LOIS, whose effect is increasingly captured by the more flexible baseline hazard time dummies as one moves from the left to the right of the table. Coefficients are in line with expecations. In all specifications, the coefficient on the Google variable is significant, raising the hazard rate between approximately 2.4 and 4.8 percent, which is roughly comparable to the coefficient on troubled assets. Capital and securities have the largest effects, both reducing the hazard rate. The interaction variable between large deposits and the FDIC insurance limit raise still dampens the hazard-reducing effect of the introduction of the FDIC raise for banks with a high percentage of large deposits. A possible interpretation may be that depositors with accounts holding between USD 100'000 and USD 250'000 profited from the raise, but customers holding deposits in excess of USD 250'000 might have interpreted the intervention as a warning sign. Further, the more a bank was invested in residential real estate the lower its hazard, which may be counterintuitive given the financial crisis has its roots in the real estate sector. Lastly, it should be noted that the significant effect in assets is mainly driven by the failures of the three largest banks; excluding them from the analysis leads to statistically insignificant coefficients on assets (not shown in the table).

[Table 5 about here.]

The appendix further lists results including censored Google series in Table A9 as well as results ignoring weighting. The coefficients confirm the results shown above. The coefficient on the Google variable is attenuated towards zero when using censored Google series, which is expected as falls in the Google Index are overstated in the censored case.

Table 6 presents results for forecasting where the contemporaneous $Google_{it}$ growth variable is replaced with variables that are lagged by two to five weeks or with growth rates spanning two to five weeks. Again, the dataset containing only uncensored Google series is used. The control variables remain the same, but the output table is restricted to the coefficients for Google variables only. The top half presents specifications including lagged values of the Google variable from two up to five weeks, with the last column including all lags. The size of the Google coefficient goes toward zero and becomes statistically insignificant as one moves back in time. An exception is the last column in the upper half including all lags, showing significant effects for the three weeks before failure, confirming the pattern seen initially in Figure 2.

In the bottom half of Table 6, the Google variable covers the accumulated growth rate over a longer period, from a 2-week period up to a 5-week period. The results confirm the previous statements: Results are mainly driven by the Google growth values in the week of the failure; adding additional weeks and lengthening of the time period barely changes the estimated coefficients.

[Table 6 about here.]

5 Conclusion

Washington Mutual was still considered well capitalized shortly before its closure (Grind, 2012), but the situation changed rapidly in mid-September. Within a few weeks, a well-capitalized bank - which admittedly did have problems with its mortgages - had to be shut down, as closing the bank was apparently the only option to stop the ongoing run on deposits. Once a bank run is kicked off, a vicious feedback-loop is started and withdrawals spread like a virus. As the Diamond and Dybvig (1983) model shows, one ends up in an equilibrium where it becomes rational for every agent to pull out their funds; even deposit insurance may prove ineffective at this point (Iyer and Puria, 2012; Grind, 2012).

Such bank failures are hard to predict. Empirical research analyzing the survival and survival time of banks by making use of their balance sheets provides insights, but these studies have their limits when it comes to the timing of the failure. Other research focussing on single banks helps understanding the dynamics during a bank run, but cannot explain when or why the bank run occurred in the first place. Google data can provide additional insights and accuracy in this field. As this study demonstrates, Google search volumes start rising up to two weeks before failure, indicating increased attention on the internet for an individual bank. By capturing short term dynamics that cannot be reflected in quarterly balance sheet and revenue data, Google queries can be a valuable improvement to more traditional predictions, especially when it comes to the timing of the failure. Compared to other instruments used to capture publicly available information, Google has the advantage of being "democratically" weighted rather than binary or influenced by other variables, reflecting the spread of information more accurately. While it is hard to know how many readers read a newspaper article, a rising Google index can always be translated into more people being concerned.

Nevertheless, it should be pointed out that Google search queries have their limits, too. First and foremost, one does not know what drives the spike in search queries or what actions follow after the Google search, making any causal claims hard to defend. Whether a news article leads to

the rising search volume or customers looking for their e-banking accounts is unknown. It would be rash to equate a rising Google Index with a bank run. What this study shows is that it can serve as a warning signal that failure is imminent. Still, the timing of spikes in search volumes remains hard to predict. As one moves further away from the bank's failure date by more than three weeks, Google loses its predictive power.

One should also keep in mind Google data's technical limitations. First, the data are censored. Particularly small banks fail to pass the search volume threshold, which means that there is no data available at all. Second, Google publishes only relative numbers, which allows for the use of growth rates only. Third, Google is not the internet. Google may be a popular search engine, but it does not track all activity on the web. Instead of a substitute, Google data should therefore be seen as a supplement to balance sheet and revenue analysis.

References

Askitas, N. and K. F. Zimmermann (2009). Google econometrics and unemployment forecasting. Technical report, German Institute for Economic Research.

Bollen, J., H. Mao, and X. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science*.

Breyer, B. N., S. Sen, D. S. Aaronson, M. Stoller, B. A. Erickson, and M. L. Eisenberg (2011). Use of Google Insights for Search to track seasonal and geographic kidney stone incidence in the United States. *Urology*.

Calomiris, C. W. and J. R. Mason (2003). Fundamentals, panics, and bank distress during the depression. *American Economic Review*, 1615–1647.

Choi, H. (2009). Predicting initial claims for unemployment benefits. *SSRN 1659307*.

Choi, H. and H. Varian (2012). Predicting the present with Google Trends. *Economic Record 88*(s1), 2–9.

Cole, R. A. and J. W. Gunther (1995). Separating the likelihood and timing of bank failure. *Journal of Banking & Finance 19*(6), 1073–1089.

D'Amuri, F. and J. Marcucci (2010). Google it! Forecasting the US unemployment rate with a Google job search index. *FEEM Working Paper No. 31.2010*.

Della Penna, N. and H. Huang (2009). Constructing consumer sentiment index for US using Google searches. *Working Paper*.

Diamond, D. W. and P. H. Dybvig (1983). Bank runs, deposit insurance, and liquidity. *The Journal of Political Economy*, 401–419.

Donaldson, R. G. (1992). Costly liquidation, interbank trade, bank runs and panics. *Journal of Financial Intermediation 2*(1), 59–82.

Federal Deposit Insurance Corp. Bank data and statistics. `http://www.fdic.gov/bank/statistical/`.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant (2008). Detecting influenza epidemics using search engine query data. *Nature 457*(7232), 1012–1014.

Goel, S., J. M. Hofman, S. Lahaie, D. M. Pennock, and D. J. Watts (2010). Predicting consumer behavior with web search. *Proceedings of the National Academy of Sciences 107*(41), 17486–17490.

Google. Google Insights for Search. `http://www.google.com/insights/search/`.

Gorton, G. (1988). Banking panics and business cycles. *Oxford economic papers 40*(4), 751–781.

Gorton, G. and A. Winton (2003). Financial intermediation. *Handbook of the Economics of Finance 1*, 431–552.

Grind, K. (2012). *The Lost Bank: The Story of Washington Mutual-The Biggest Bank Failure in American History*. Simon & Schuster.

Guzman, G. (2011). Internet search behavior as an economic forecasting tool: The case of inflation expectations. *Journal of Economic and Social Measurement 36*(3), 119–167.

Iyer, R., M. Puri, and N. Ryan (2013). Do depositors monitor banks? Technical report, National Bureau of Economic Research.

Iyer, R. and M. Puria (2012). Understanding bank runs: the importance of depositor-bank relationships and networks. *The American Economic Review 102*(4), 1414–1445.

Jenkins, S. P. (2005). Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*.

Kelly, M. and C. O Grada (2000). Market contagion: Evidence from the panics of 1854 and 1857. *American Economic Review*, 1110–1124.

Lancaster, T. (1992). *The econometric analysis of transition data*. Number 17. Cambridge University Press.

Los Angeles Times (2008, June 28). Senator asks regulators to probe the financial health of Indymac.

Mathiesen, J., L. Angheluta, P. T. H. Ahlgren, and M. H. Jensen (2013). Excitable human dynamics driven by extrinsic events in massive communities. *Proceedings of the National Academy of Sciences*.

McCarthy, M. J. (2010). Internet monitoring of suicide risk in the population. *Journal of affective disorders 122*(3), 277–279.

McLaren, N. and R. Shanbhogue (2011). Using internet search data as economic indicators. *Bank of England Quarterly Bulletin No. 2011 Q2*.

Moat, H. S., C. Curme, A. Avakian, D. Y. Kenett, E. H. Stanley, and T. Preis (2013). Quantifying Wikipedia usage patterns before stock market moves. *Scientific reports 3*.

O Grada, C. and E. N. White (2003). The panics of 1854 and 1857: A view from the emigrant industrial savings bank. *The Journal of Economic History 63*(1), 213–240.

Office of Thrift Supervision (2008, September 25). OTS fact sheet on Washington Mutual Bank.

Preis, T., H. S. Moat, and E. H. Stanley (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific reports 3*.

Preis, T., D. Reith, and H. Stanley (2010). Complex dynamics of our economic life on different scales: insights from search engine query data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 368*(1933), 5707–5719.

Radinsky, K., S. Davidovich, and S. Markovitch (2008). Predicting the news of tomorrow using patterns in web search queries. In *Web Intelligence and Intelligent Agent Technology*, Volume 1, pp. 363–367. IEEE.

Rigobon, R. and T. M. Stoker (2007). Estimation with censored regressors: Basic issues. *International Economic Review 48*(4), 1441–1467.

Saunders, A. and B. Wilson (1996). Contagious bank runs: Evidence from the 1929–1933 period. *Journal of Financial Intermediation 5*(4), 409–423.

Seabrook, A. (2008, July 12). Interview with Burt Ely. *National Public Radio*.

Shin, H. S. (2009). Reflections on Northern Rock: the bank run that heralded the global financial crisis. *The Journal of Economic Perspectives*, 101–120.

Tefft, N. (2011). Insights on unemployment, unemployment insurance, and mental health. *Journal of Health Economics*.

United States Census Bureau. 2010 Census urban and rural classification and urban area criteria. `http://www.census.gov/geo/www/ua/2010urbanruralclass.html`.

Whalen, G. (1991). A proportional hazards model of bank failure: an examination of its usefulness as an early warning tool. *Federal Reserve Bank of Cleveland Economic Review 27*(1), 21–31.

Wheelock, D. and P. Wilson (2000). Why do banks disappear? The determinants of US bank failures and acquisitions. *Review of Economics and Statistics 82*(1), 127–138.

Wu, L. and E. Brynjolfsson (2013). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economics of Digitization*. University of Chicago Press.

Figures

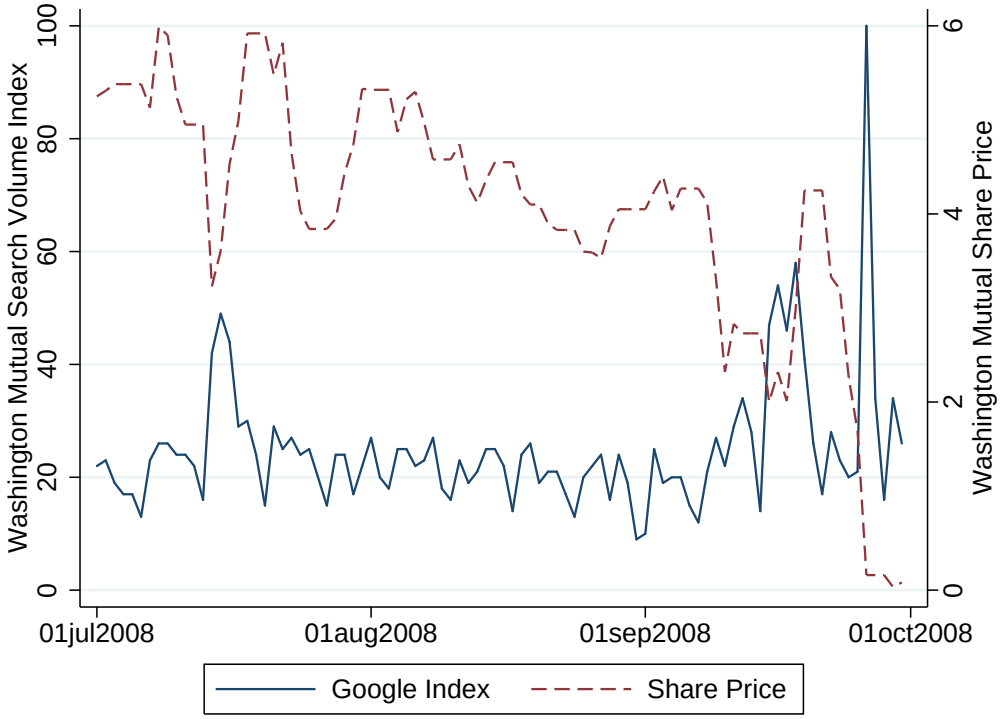


Figure 1: Google Search Volume Index and Share Price for "Washington Mutual"

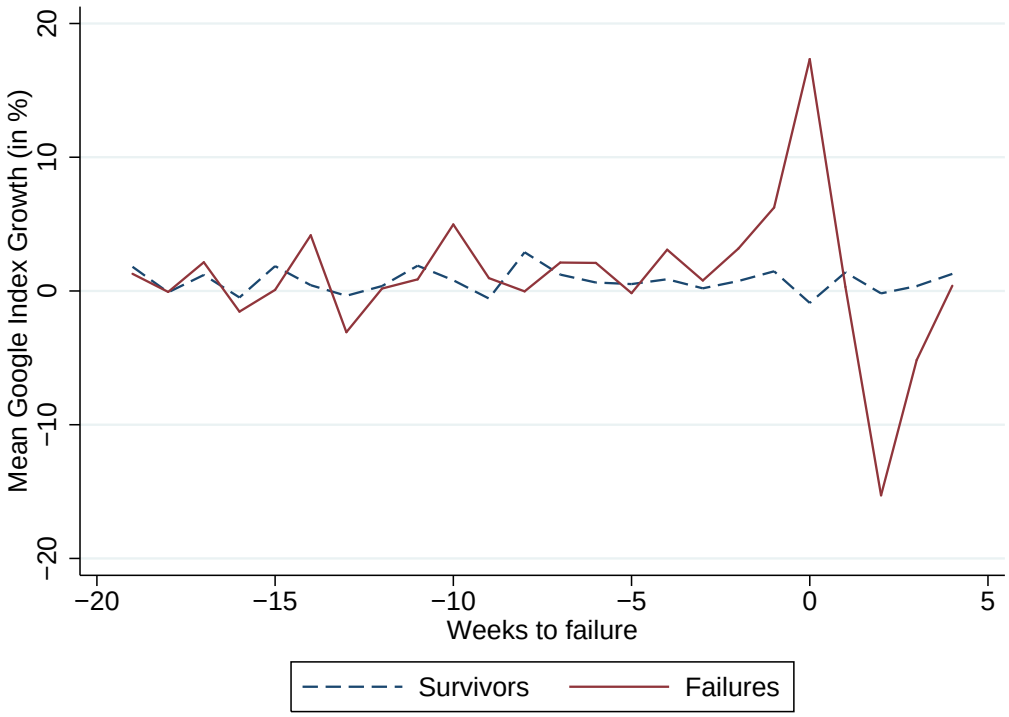
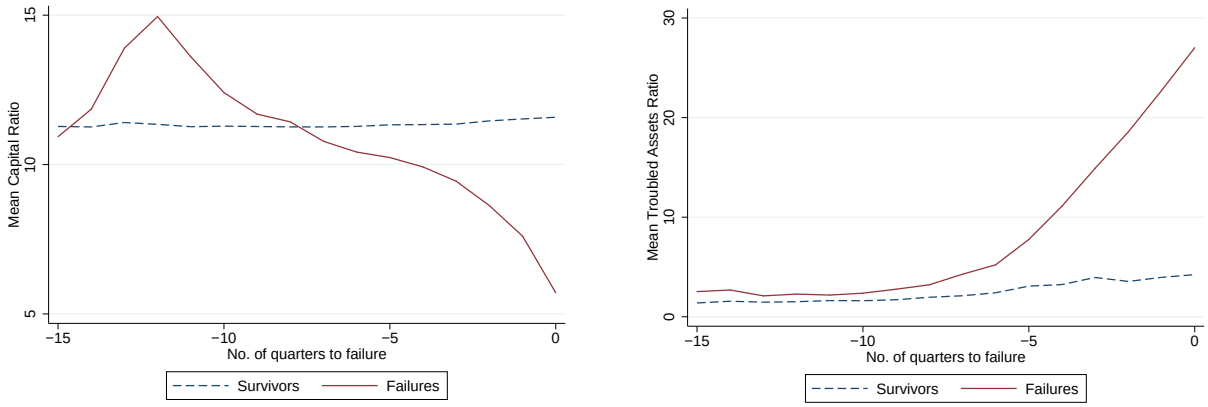
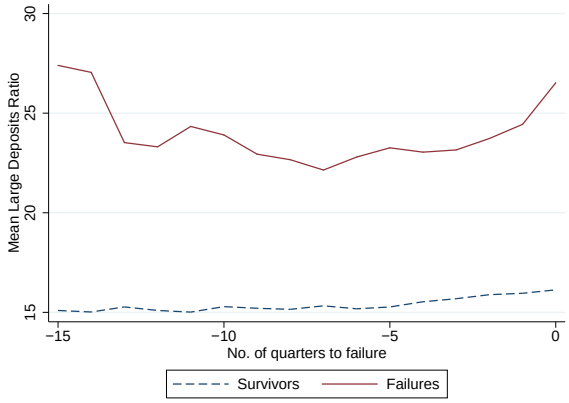


Figure 2: Google Growth Rates in the Weeks Prior to Failure

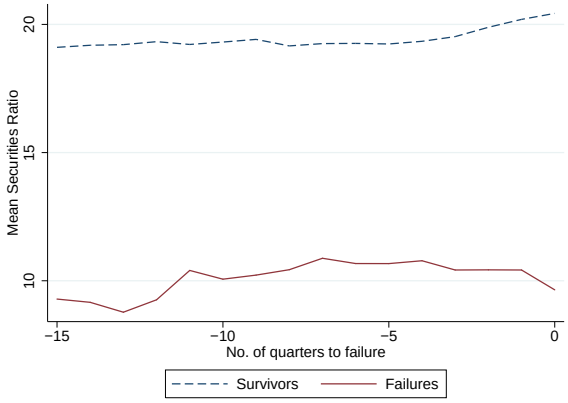


Capital

Troubled Assets



Large CDs



Securities

Figure 3:
Key Balance Sheet Positions Before Failure, conditioned on observing uncensored Google series

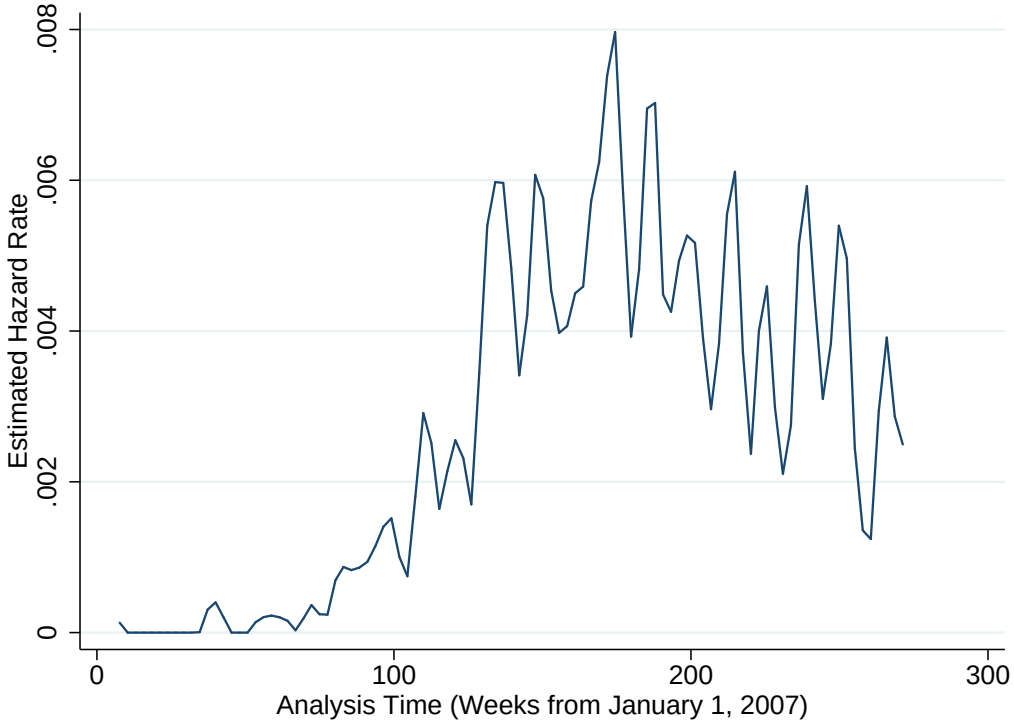


Figure 4: Smoothed Hazard Rate Estimate

Tables

Table 1: Overview of Samples Used

Sample	**Observations**	**Failing banks**	**Surviving banks**
Original sample	180'291	428	387
With Google series	45'835	115	95
With uncensored Google series	10'296	25	48

Table 2: Bank failures over time

	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
No. of failures	3	0	0	3	25	140	157	92	16	436

The year 2012 covers only the first quarter of the year.

The years 2004, 2005, 2006 are excluded from the analysis.

Source: FDIC

Table 3: Summary Statistics

Full sample

	Failures				*Survivors*			
Variable	**Mean**	**St. Dev.**	**Min**	**Max**	**Mean**	**St. Dev.**	**Min**	**Max**
Google data dummy	.269	-	0	1	0.245	-	0	1
Gross assets *(in USD millions)*	1'518.531*	15'727.860	0	322'059.800	394.069*	1'088.604	3.220	12'585.320
Capital	7.502*	3.213	0	22.828	12.411*	6.076	0	77.759
Troubled assets	8.252*	5.293	0	42.943	3.364*	2.928	0.290	17.944
Net income	-0.498*	0.333	-2.729	0.312	0.130*	0.286	-1.073	2.317
Securities	8.191*	6.634	0	40.711	21.743*	14.452	0	76.510
Large CDs	16.209	8.164	0	53.467	16.173	7.440	0	39.799
Insider	1.093	1.389	0	10.956	1.345	1.488	0	10.716
Holding Co.	0.702	-	0	1	.692	-	0	1
Entering age	35.665*	38.406	0.071	156.493	68.375*	43.438	0.186	170.012
Urban	0.341*	-	0	1	0.437*	-	0	1
Observations	428				387			

With Google data

	Failures				*Survivors*			
Variable	**Mean**	**St. Dev.**	**Min**	**Max**	**Mean**	**St. Dev.**	**Min**	**Max**
Google growth rate	2.699*	2.331	-8.673	6.960	*1.632	2.515	-11.384	5.146
Google data dummy	1	0	1	1	1	0	1	1
Gross assets *(in USD millions)*	4'618.891*	30'184.530	40.085	322'059.800	727.244*	1'803.306	20.207	12'585.320
Capital	9.996*	2.502	5.522	22.828	11.976*	7.248	6.540	77.759
Troubled assets	10.265*	6.488	1.191	42.943	3.849*	3.417	0.420	17.944
Net income	-0.619*	0.426	-2.729	0.057	0.155*	0.344	-0.613	2.317
Securities	11.212*	7.978	0	40.711	20.825*	14.152	0	73.761
Large CDs	20.869*	7.916	5.282	53.467	17.598*	7.744	0	39.799
Insider	1.124	1.639	0	10.956	1.321	1.214	0	4.966
Holding Co.	0.687	-	0	1	0.726	-	0	1
Entering age	38.656*	38.511	0.624	156.493	60.966*	46.139	1.572	144.471
Urban	0.304	-	0	1	0.474	-	0	1
Observations	115				95			

Uncensored Google series only

	Failures				*Survivors*			
Variable	**Mean**	**St. Dev.**	**Min**	**Max**	**Mean**	**St. Dev.**	**Min**	**Max**
Google growth rate	1.313	1.396	-0.024	6.857	0.791	1.010	0.127	5.146
Google data dummy	1	0	1	1	1	0	1	1
Gross assets *(in USD millions)*	15'400.920*	64'138.350	60.974	322'059.800	405.449*	574.211	36.744	2'316.473
Capital	10.119	2.769	5.522	15.224	11.684	2.164	8.894	18.799
Troubled assets	9.868*	6.794	2.388	30.363	3.971*	2.852	0.878	11.329
Net income	-0.718*	0.512	-2.729	-0.152	0.109*	0.225	-0.467	0.526
Securities	10.194*	6.784	1.575	28.671	20.281*	12.676	1.964	47.787
Large CDs	22.849	10.780	5.701	53.467	15.693	5.719	7.828	27.899
Insider	1.514	2.707	0	10.956	1.418	1.134	0	3.969
Holding Co.	.560	-	0	1	.696	-	0	1
Entering age	28.636*	27.770	0.953	99.806	69.443*	47.741	3.773	143.411
Urban	0.360	-	0	1	0.609	-	0	1
Observations	25				23			

Sources: Google Insights for Search, FDIC, Bloomberg, 2010 US Census.

1 Reported values are averaged by institution and cover the period from Jan 2007 to Mar 2012 or up to failure, respectively.

2 An $*$ indicates that the Wilcoxon rank-sum test statistic for a shift in the location parameter between the two groups is significant at the one percent level.

3 The difference in gross assets in the lower third of the table (uncensored Google series only) is not signifcant anymore if the two largest banks Washington Mutual and Indymac are excluded.

4 ComRE, ResRE and Insdep variables have been omitted.

Table 4: Results on Survival and Google data availability

	Google only	**PWC-FDIC**	**PWC-yearly**	**PWC-quarterly**	**Cox PH**
Variable	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
Google	0.120	0.315	0.289	0.273	0.234
	(0.157)	(0.232)	(0.234)	(0.232)	(0.236)
Capital	-	-0.470***	-0.475***	-0.474***	-0.489***
		(0.055)	(0.028)	(0.029)	(0.030)
Troubledassets	-	0.055***	0.055***	0.057***	0.058***
		(0.007)	(0.007)	(0.007)	(0.007)
Netincome	-	0.067	0.054	0.049	0.033
		(0.047)	(0.043)	(0.041)	(0.038)
Securities	-	-0.064***	-0.061***	-0.060***	-0.058***
		(0.014)	(0.014)	(0.014)	(0.014)
LargeCDs	-	-0.094**	-0.093**	-0.101**	-0.091*
		(0.036)	(0.030)	(0.038)	(0.041)
Insdep	-	0.090*	0.088**	0.097*	0.089*
		(0.037)	(0.030)	(0.039)	(0.041)
ComRE	-	0.034**	0.035**	0.035**	0.036**
		(0.011)	(0.011)	(0.011)	(0.0011)
ResRE	-	-0.000	0.001	-0.000	-0.001
		(0.010)	(0.010)	(0.010)	(0.010)
Insider	-	-0.047	-0.054	-0.055	-0.075
		(0.060)	(0.060)	(0.060)	(0.074)
Assets	-	0.078	0.076	0.074	0.072
		(0.074)	(0.074)	(0.074)	(0.074)
Age	-	-0.001	-0.001	-0.001	-0.002
		(0.003)	(0.003)	(0.003)	(0.003)
Holding	-	-0.705**	-0.703**	-0.690**	-0.664**
		(0.237)	(0.241)	(0.243)	(0.248)
Urban	-	0.436*	0.419*	0.427*	0.444*
		(0.191)	(0.192)	(0.190)	(0.191)
LOIS	-	0.180	0.054	-0.360	-
		(0.497)	(0.194)	(0.380)	
Piecewise constant haz.	quarterly	2-period	yearly	quarterly	-
Google Variable	Dummy	Dummy	Dummy	Dummy	Dummy
Observations	181'113	181'113	181'113	181'113	181'113
Subjects	818	818	818	818	818
Failures	428	428	428	428	428
Weighted failures	40.330	40.330	40.330	40.330	40.330
Log-pseudolikelihood	-169.566	-33.191	-32.762	-31.821	-104.610

Significance levels : † : 10% * : 5% ** : 1% *** : 0.1%

1 Clustered standard errors reported in parentheses (clustered on subject).

2 Balance sheet and revenue variables are all expressed in percentages, i.e. are roughly in a range from 0 to 100.

3 Google changes are expressed in percentages, i.e. one percent is 1, onehundred percent are 100.

4 Episodes are split on a weekly basis.

5 The Cox PH model uses the Breslow method for ties.

Table 5: Main Results, uncensored Google series only

Variable	**Google only** Coefficient	**PWC-FDIC** Coefficient	**PWC-yearly** Coefficient	**PWC-quarterly** Coefficient	**Cox PH** Coefficient
Google	0.042***	0.024*	0.024***	0.042*	0.048***
	(0.012)	(0.010)	(0.007)	(0.016)	(0.010)
Capital	-	-0.506***	-0.539***	-0.579***	-0.625***
		(0.094)	(0.090)	(0.107)	(0.160)
Troubledassets	-	0.049†	0.059*	0.070**	0.077**
		(0.028)	(0.028)	(0.024)	(0.026)
Netincome	-	0.083	0.072	-0.076	-0.050
		(0.066)	(0.069)	(0.130)	(0.130)
Securities	-	-0.248**	-0.261**	-0.317**	-0.348**
		(0.086)	(0.092)	(0.119)	(0.134)
LargeCDs	-	-0.342	-0.237	-1.453**	-5.083***
		(0.222)	(0.217)	(0.517)	(1.212)
Insdep	-	0.413†	0.304	1.515**	5.155***
		(0.231)	(0.219)	(0.502)	(1.204)
ComRE	-	-0.046	-0.055	-0.026	-0.027
		(0.044)	(0.051)	(0.053)	(0.051)
ResRE	-	-0.068***	-0.073***	-0.094*	-0.113*
		(0.016)	(0.022)	(0.037)	(0.047)
Insider	-	0.105	0.166	0.176	0.153
		(0.140)	(0.153)	(0.133)	(0.138)
Assets	-	1.021*	1.102*	0.826***	0.737**
		(0.436)	(0.508)	(0.240)	(0.268)
Age	-	-0.025	-0.029†	-0.030*	-0.023†
		(0.016)	(0.017)	(0.014)	(0.012)
Holding	-	2.007*	2.147*	1.056	0.635
		(0.940)	(1.019)	(0.787)	(0.935)
Urban	-	0.881	0.996	2.823**	2.976**
		(0.879)	(1.018)	(1.002)	(1.028)
LOIS	-	0.571*	0.322	0.126	-
		(0.269)	(0.807)	(1.102)	
Piecewise constant haz.	quarterly	2-period	yearly	quarterly	-
Google Variable	%-change	%-change	%-change	%-change	%-change
Observations	10'296	10'296	10'296	10'296	10'296
Subjects	48	48	48	48	48
Failures	25	25	25	25	25
Weighted failures	2.367	2.367	2.367	2.367	2.367
Log-pseudolikelihood	-8.249	1.517	1.653	2.863	2.366

Significance levels : † : 10% * : 5% ** : 1% *** : 0.1%

1 Clustered standard errors reported in parentheses (clustered on subject).

2 Balance sheet and revenue variables are all expressed in percentages, i.e. are roughly in a range from 0 to 100.

3 Google changes are expressed in percentages, i.e. one percent is 1, onehundred percent are 100.

4 Episodes are split on a weekly basis.

5 The Cox PH model uses the Breslow method for ties.

Table 6: Forecasting

Variable	**PWC**	**PWC**	**PWC**	**PWC**	**PWC**
Variable	Coefficient	Coefficient	Coefficient	Coefficient	Coefficient
1-week lag	-	-	-	-	0.039***
					(0.011)
2-week lag	.027	-	-	-	0.036**
	(0.019)				(0.012)
3-week lag	-	0.013	-	-	0.038*
		(0.014)			(0.019)
4-week lag	-	-	-.006	-	0.023
			(0.017)		(0.021)
5-week lag	-	-	-	0.007	0.018
				(0.011)	(0.013)
Piecewise constant haz.	quarterly	quarterly	quarterly	quarterly	quarterly
Google Variable	%-change	%-change	%-change	%-change	%-change
Observations	10'296	10'296	10'296	10'296	10'296
Subjects	48	48	48	48	48
Failures	25	25	25	25	25
Weighted failures	2.367	2.367	2.367	2.367	2.367
Log-pseudolikelihood	2.353	2.219	2.191	2.195	3.243
2-week period	0.025***	-	-	-	
	(0.004)				
3-week period	-	0.021***	-	-	
		(0.003)			
4-week period	-	-	0.022***	-	
			(0.003)		
5-week period	-	-	-	0.024***	
				(0.003)	
Piecewise constant haz.	quarterly	quarterly	quarterly	quarterly	
Google Variable	%-change	%-change	%-change	%-change	
Observations	10'296	10'296	10'296	10'296	
Subjects	48	48	48	48	
Failures	25	25	25	25	
Weighted failures	2.367	2.367	2.367	2.367	
Log-pseudolikelihood	3.067	3.185	3.105	3.181	

Significance levels : † : 10% * : 5% ** : 1% *** : 0.1%

1 Clustered standard errors reported in parentheses (clustered on subject).

2 Balance sheet and revenue variables are all expressed in percentages, i.e. are roughly in a range from 0 to 100.

3 Google changes are expressed in percentages, i.e. one percent is 1, onehundred percent are 100.

4 Episodes are split on a weekly basis.

F Appendix

Table A7: Description of Variables

Variable	Variable Name	Definition
Googledata	Google Data	Dummy variable indicating whether Google data is available
Google growth	Google	Google growth rate (weekly; rate covering 1 to 5 weeks)
Capital	Capital	Ratio of equity capital and loan loss reserves to gross assets.
Troubled Assets	Troubledassets	Ratio of loans past due 90 days or more, nonaccrual loans, and other real estate owned to gross assets
Net Income	Netincome	Ratio of net income to gross assets
Securities	Securities	Ratio of investment securities to gross assets
Large CDs	LargeCDs	Ratio of time deposits of USD 100'000 or more to gross assets
Commercial Real Estate Loans	ComRE	Ratio of construction loans and loans secured by multifamily, nonresidential, or farm real estate to gross assets
Residential Real Estate Loans	ResRE	Ratio of loans secured by 1-4 family real estate to gross assets
Insider Loans	Insider	Ratio of insider loans to gross assets
Gross assets	Assets	Logarithm of gross assets (USD thousands)
Entering age	Age	Age of the institution (years) when first entering the dataset
Holding Company	Holding	Dummy variable to indicate whether the institution belongs to a holding company
Urban	Urban	One for urban counties, zero otherwise
Insdep	Insdep	Interaction of insurance dummy and large deposits ratio
LIBOR	LIBOR	3 month London Interbank Offered Rate
OIS	OIS	3 month Overnight Indexed Swap (OIS)
LOIS	LOIS	Difference between LIBOR and OIS as a measure of health of the banking system
Insurance	Insurance	Dummy variable to indicate the raise of the FDIC insurance limit in October 2008
Quarter	1-21	Dummies indicating quarter
Year	2008-2012	Dummies indicating year

Sources: Google Insights for Search, FDIC, Bloomberg, 2010 US Census.

Table A8: Google Query Index Value Growth Rates

	Google Growth Rate				
Statistic	**1 week**	**2 weeks**	**3 weeks**	**4 weeks**	**5 weeks**
Observations	13'152	13'152	13'152	13'152	13'152
Mean	0.834	0.947	1.041	1.111	1.213
Median	0.000	0.000	0.000	0.000	0.000
Standard Deviation	13.635	15.164	16.035	16.569	16.938
Minimum	-73.077	-76.000	-81.00	-82.000	-81.000
Maximum	200.000	316.667	455.556	455.556	455.556

Includes only uncensored observations.

Source: Google Insights for Search

Table A9: Results including censored Google series

Variable	**Google only** Coefficient	**PWC-FDIC** Coefficient	**PWC-yearly** Coefficient	**PWC-quarterly** Coefficient	**Cox PH** Coefficient
Google	0.019***	0.016***	0.016***	0.017***	0.030***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.005)
Capital	-	-0.661***	-0.665***	-0.687***	-0.693***
		(0.041)	(0.049)	(0.055)	(0.065)
Troubledassets	-	0.041***	0.042***	0.045***	0.043***
		(0.009)	(0.009)	(0.010)	(0.011)
Netincome	-	-0.010	-0.007	0.010	-0.020
		(0.046)	(0.047)	(0.052)	(0.062)
Securities	-	-0.055**	-0.060**	-0.062**	-0.071**
		(0.021)	(0.021)	(0.022)	(0.023)
LargeCDs	-	-0.167	-0.143*	-0.182	-0.123
		(0.142)	(0.072)	(0.145)	(0.161)
Insdep	-	0.135	0.111	0.152	0.098
		(0.144)	(0.073)	(0.147)	(0.162)
ComRE	-	0.040*	0.037*	0.034*	0.038*
		(0.018)	(0.017)	(0.017)	(0.016)
ResRE	-	-0.017	-0.021	-0.025	-0.027
		(0.018)	(0.018)	(0.019)	(0.018)
Insider	-	-0.163	-0.183	-0.206	-0.318†
		(0.167)	(0.171)	(0.174)	(0.181)
Assets	-	-0.040	-0.027*	-0.045	-0.177
		(0.135)	(0.137)	(0.145)	(0.148)
Age	-	-0.007	-0.007	-0.007	-0.005
		(0.006)	(0.006)	(0.006)	(0.006)
Holding	-	0.471	0.431	0.480	0.506
		(0.322)	(0.333)	(0.350)	(0.353)
Urban	-	0.963**	0.956**	1.012**	1.074***
		(0.325)	(0.332)	(0.335)	(0.330)
LOIS	-	0.282	0.259	-0.358	-
		(0.202)	(0.332)	(0.662)	
Piecewise constant haz.	quarterly	2-period	yearly	quarterly	-
Google Variable	%-change	%-change	%-change	%-change	%-change
Observations	42'659	42'659	42'659	42'659	42'659
Subjects	210	210	210	210	210
Failures	115	115	115	115	115
Weighted failures	10.354	10.354	10.354	10.354	10.354
Log-pseudolikelihood	-33.154	-0.891	-0.702	-0.007	-7.816

Significance levels : † : 10% ∗ : 5% ∗∗ : 1% ∗ ∗ ∗ : 0.1%

1 Clustered standard errors reported in parentheses (clustered on subject).

2 Balance sheet and revenue variables are all expressed in percentages, i.e. are roughly in a range from 0 to 100.

3 Google changes are expressed in percentages, i.e. one percent is 1, onehundred percent are 100.

4 Episodes are split on a weekly basis.

5 The Cox PH model uses the Breslow method for ties.

Table A10: Unweighted Results, including uncensored Google series only

	Google only	**PWC-FDIC**	**PWC-yearly**	**PWC-quarterly**
Variable	Coefficient	Coefficient	Coefficient	Coefficient
Google	0.038***	0.023***	0.022***	0.039***
	(0.006)	(0.007)	(0.006)	(0.012)
Capital	-	-0.423***	-0.461***	-0.499***
		(0.087)	(0.095)	(0.105)
Troubledassets	-	0.043**	0.050***	0.064***
		(0.016)	(0.015)	(0.018)
Netincome	-	0.121†	0.119	0.050
		(0.071)	(0.073)	(0.095)
Securities	-	-0.179**	-0.191**	-0.270**
		(0.063)	(0.066)	(0.087)
LargeCDs	-	-0.232	-0.218	-1.266*
		(0.365)	(0.244)	(0.622)
Insdep	-	0.277	0.263	1.314*
		(0.366)	(0.242)	(0.624)
ComRE	-	-0.049†	-0.055*	-0.047
		(0.026)	(0.027)	(0.032)
ResRE	-	-0.054*	-0.057*	-0.080**
		(0.021)	(0.023)	(0.031)
Insider	-	0.052	0.118	0.097
		(0.178)	(0.177)	(0.180)
Assets	-	0.622*	0.706*	0.477†
		(0.251)	(0.277)	(0.270)
Age	-	-0.011	-0.012	-0.011
		(0.011)	(0.011)	(0.011)
Holding	-	1.405*	1.458*	0.713
		(0.701)	(0.729)	(0.713)
Urban	-	0.770	0.885c	2.366*
		(0.728)	(0.787)	(0.928)
LOIS	-	0.466	0.389	0.205
		(0.368)	(0.668)	(0.883)
Piecewise constant haz.	quarterly	2-period	yearly	quarterly
Google Variable	%-change	%-change	%-change	%-change
Observations	10'296	10'296	10'296	10'296
Subjects	48	48	48	48
Failures	25	25	25	25
Weighted failures	25	25	25	25
Log-pseudolikelihood	-21.653	23.007	24.997	36.494

Significance levels : † : 10% * : 5% ** : 1% *** : 0.1%

1 Clustered standard errors reported in parentheses (clustered on subject).

2 Balance sheet and revenue variables are all expressed in percentages, i.e. are roughly in a range from 0 to 100.

3 Google changes are expressed in percentages, i.e. one percent is 1, onehundred percent are 100.

4 Episodes are split on a weekly basis.