International Journal of Economics and Finance; Vol. 6, No. 8; 2014
ISSN 1916-971X E-ISSN 1916-9728
Published by Canadian Center of Science and Education

Big Data and the Dot Com Bubble

James Cicon[1]

[1] School of Management, New Jersey Institute of Technology, New Jersey, USA

Correspondence: James Cicon, School of Management, New Jersey Institute of Technology, University Heights, Newark, New Jersey 07102, USA. Tel: 573-823-0555. E-mail: cicon@njit.edu

Received: April 8, 2014 Accepted: May 22, 2014 Online Published: July 25, 2014

doi:10.5539/ijef.v6n8p15 URL: http://dx.doi.org/10.5539/ijef.v6n8p15

Abstract

I develop a big-data model which predicts dot-com market behavior. My model also predicts the dot-com collapse three months prior to its occurrence. My model differs from others that fail to explain the dot-com market in three ways. First it uses an objective machine driven methodology to analyze media news stories. Second, it treats news articles as complex multi-thematic constructs. Third it requires that news stories mention the firm in its headline. I submit that these three factors enable my model to explain dot-com market behavior where other models fail to do so.

Keywords: financial crisis, asset pricing, market efficiency, international financial markets, financial forecasting and simulation

1. Introduction

The literature provides mixed evidence as to the relationship between media coverage and bubble formation and collapse. Research immediately following the dot-com bubble collapse suggests that media-coverage exacerbates bubble formation and collapse. Shiller (2000) proposes that positive media feedback leads investors to overvalue internet stocks prior to the bubble, and after the bubble, leads investors to undervalue stocks. Thus media forces drive prices too high, prior to the bubble, and too low after the bubble. Subsequent research supports Shiller's finding that media coverage influences investor behavior. Tetlock (2007) published a series of papers showing that negative media drives stock prices down. Loughran and McDonald (2011) and Hanley and Hoberg (2009) show that investors respond to how the firm presents information in SEC filings, which are mined by media for information on the firm.

Recently, several papers have appeared which draw into question the results of the earlier research stream outlined above. Bhattacharya et al. (2009) reads and classifies media news items appearing between the years 1996 and 2000. They subjectively place the news articles into 'positive', 'neutral' and 'bad' news categories. They then test if the news articles explain bubble stock overpricing and/or underpricing. They fail to find such evidence and conclude that “media hype is unable to explain the stock bubble”. Campbell, Turner and Walker (2012) use Bhattacharya's methodology to analyze the Railway bubble which occurred in England in the mid 1840s. They conclude that the media did not exacerbate bubble development. Rather, their evidence leads them to find that the media provides a “fair and unbiased” information resource to investors, which helps investors make rational buying/selling decisions. Thus, inexplicably, the latter literature on media's impact on bubbles is inconsistent with the earlier literature. This inconsistency provides me with an opportunity to contribute to the literature.

In this study I provide evidence that the disparity of the findings between researchers, such as Shiller (2000) and Bhattacharya et al. (2009), are rooted in the methodologies used to analyze media hype. Specifically, I propose that the Bhattacharya study has several shortcomings. The first shortcoming is that the classification of the articles may have involved too much subjectivity. He and a coauthor spent two years reading their sample of 171,488 media releases, judging which category to place them in. I propose that for a human reading, that two years is too long and that their sample is too large, for the process to be consistent and repeatable.

Second, each news story is categorized as *prima facia*, positive, neutral or negative. Unfortunately news is not so simple. Loughran and McDonald (2011), Hanley and Hoberg (2009) and Brockman and Cicon (2013) all show that news is more complex, and that news stories contain many themes which add independent, incremental,

information to the reader. Thus, when Bhattacharya et al. forces a single subjective label onto a news story, they confound the subtle content that the media is communicating to the investing public.

Lastly, Bhattacharya et al. (2009) use all new articles about a firm in which the firm is mentioned anywhere within the news article. I propose that most readers focus on the headline of a news article and perhaps the first paragraph. Thus, using news articles that discuss the firm only in the latter part of the article body, but not in the headline or first paragraph, introduces considerable noise into the study which may lead to the weakening of results.

I avoid these three shortcomings. First, I do not read the news stories; instead I use a computer to read them. To do this, each news story is saved as a text file to the computer's hard drive. These text files are then read into computer memory where a matrix is created for each file. The rows of this matrix represent each unique word found in the text file, and the columns of this matrix represent the counts of the words. Thus, a vector of matrices is created, with each matrix representing one news story and the entire vector representing the entire corpus being investigated. This approach has been used by many researchers in the finance literature (Loughran & McDonald, 2011; Hanley & Hoberg, 2009; Brockman & Cicon, 2013; Cicon, Clarke, Ferris, & Jayaraman, 2013). The primary advantage of this approach is that it removes the subjective nature of trying to classify documents read at different periods (days or years apart) and in different circumstances (home, office or campus). The computer reads all of the documents within a few minutes time and categorizes them in a way that is 100% replicable with regards to time and location.

I solve the second shortcoming by using methodologies similar to that used in Cicon, Clarke, Ferris and Jayaraman (2013). In this paper the authors break each press release into composite themes, treating each theme as incrementally informative. I use the six themes defined by Loughran and McDonald, (2011): positivity, negativity, litigiousness, uncertainty, modal strong and modal weak. Each of these themes is defined by a list of words that capture that theme's semantic context in the English language. For example, some of the words in the positivity dictionary are able, abundance, achieve, beautiful, charitable, etc. The positivity dictionary consists of over 350 words. The more of these words that the document contains, the higher the 'score' the document receives for positivity. I create scores for each document for each of the six L&M dictionaries. This differs from Bhattacharya's approach which treats each news article as a single token of information, an approach which weakens their study and may explain why they conclude that "media hype is unable to explain the stock bubble".

I resolve the last shortcoming by limiting the breath of news coverage in my study. I require that the firm's name and/or ticker appear in the headline of the article in order for the news article to relevant to the firms in my study. Researchers provide evidence that readers searching for information about a firm do not read all available news articles about all available firms in order to divine information relevant to a particular firm. Rather, readers search headlines only for information about specific firms and read only those articles which mention the firm they are interested in the headline of the article. This interesting phenomenon is substantiated by findings from the Media Insight Project, an initiative of the American Press Institute (Note 1).

2. Data and Methodology

I begin building my sample of firms by searching SDC for all IPOs that issued over the period beginning in 1996 and ending in 2001. I exclude those IPOs which are unit offerings, rights offerings, closed end mutual funds, REITs and American depository receipts (ADRs). This search yields 2,706 firms. I download these firms and match them to the list of internet firms provided by Loughran and Ritter (2004). This step leaves me with a sample of 442 firms. I next create a matching sample of non-internet firms based on offer size and date. I match without replacement. Throughout this process I cross check each match with CRSP and COMPUSTAT. If the 'match' does not exist in both of these databases, I drop it and seek a match which does. At the end of this procedure I have a sample of 884 firms, 442 of which are 'internet' firms and 442 of which are 'non-internet' firms.

I next create my database of news articles for these firms. I begin by searching LexisNexis Academic for all news articles concerning my sample of 882 firms. I identify and download a total of 134,990 articles. The average news article length is 577 words. The total number of words in my article database is 77,883,240. I write C++ code to serialize each text file and extract company name, ticker symbol, the company's industry, the news headline and the news body. I then implement the methodology of Loughran and McDonald (2009) to compute softscores for each of their semantic constructs: litigiousness, modal strong, modal weak, negative tone, positive tone, and uncertainty.

3. Empirical Results

I begin this section of my paper by replicating the methodology of Bhattacharya et al. (2009). First I compute the mean one-day accrual for the internet stocks. I cumulate over time from 1997 through 2000. I repeat this procedure for non-internet stocks. Then, for each day in my sample, I take the difference between the return on the internet firms and the non-internet firms and call this 'actual cumulative returns'. I plot the actual returns in Figure 1. This figure is virtually identical to Figure 3 in Bhattacharya et al. (2009). This provides evidence that my sample and methodology are materially similar Bhattacharya's.

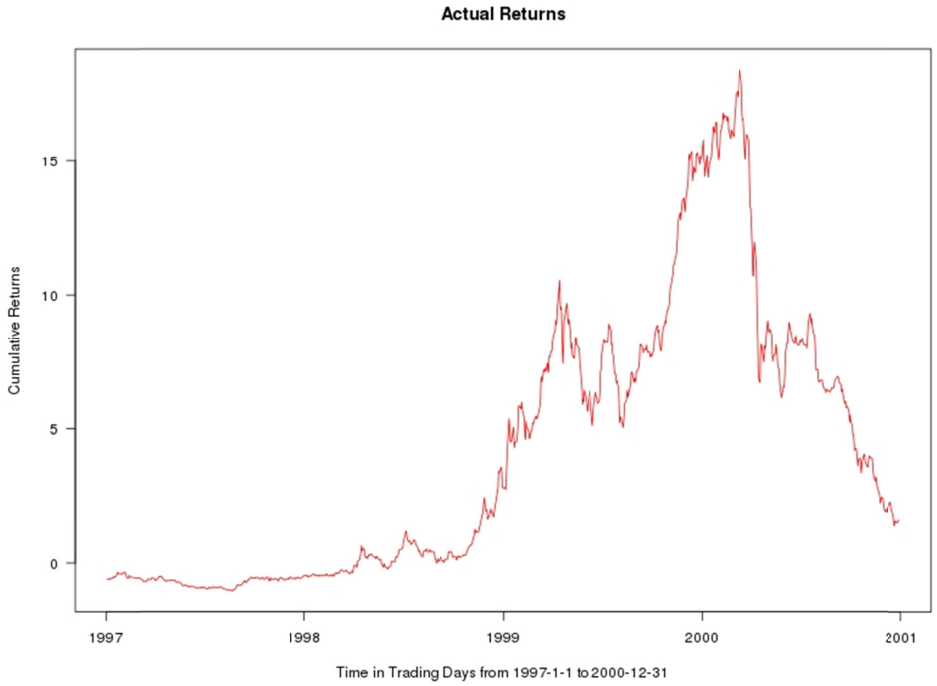

Figure 1. Actual cumulative returns

Note. In this figure I replicate the 'internet sample–non-internet sample' plot in Figure 3 of Bhattacharya et al. (2009). I generate data for this plot by, first, computing the mean one-day accrual for the internet stocks. I cumulate them over time from 1997 through 2000. I repeat this procedure for non-internet stocks. Then, for each day in my sample, I take the difference between the return on the internet firms and the non-internet firms and call this 'actual returns'.

Next I compute the Fama French abnormal returns. I begin by evaluating the following model.

$$RET_{i,t} = \alpha_0 + \Phi_0^T \cdot FF_{i,t} + \quad _{i,t} \tag{1}$$

where $RET_{i,t}$ is the stock return for firm i on day t and $FF_{i,t}$ are the three contemporary Fama French factors. After fitting this model, I extract the predicted returns. I repeat this procedure for the non-internet stocks. I then take the difference between the two sets of predicted returns and subtract this difference from the actual cumulative returns (computed above) to arrive at the Fama French abnormal returns.

I use the Fama French abnormal returns to determine if my methodology explains dot-com internet stock performance. I do so by estimating the following model:

$$abn_{ff} = \alpha_0 + \beta_1 \cdot lmLit + \beta_2 \cdot lmNeg + \beta_3 \cdot lmPos + \beta_4 \cdot ModStrng + \beta_5 \cdot ModWeak + \beta_6 \cdot lmUncrt + \varepsilon \tag{2}$$

where all variables are defined in Table 1. I report estimation results in Table 2. In Panel A, I report the results for only the internet firms. In Panel B, I report the results for only the non-internet firms. Panel A reports that the 'positive' and the 'modal strong' (commitment) word dictionaries significantly explain the internet firms with coefficients (and t-stats) of -0.00398 (-3.373) and 0.00405 (3.649) respectively. This means that a 0.4% decrease

in 'positive' words leads to a 1% lower abnormal returns, and 0.4% increase in 'commitment' words leads to a 1% increase in abnormal returns.

Table 1. Variable definitions

Variable Name	Description
L&M Soft Variables	The Loughran and McDonald (2009) soft variables
lmLitigious (lmLit)	Words denoting legal risk and/or legal propensity
lmNegative (lmNeg)	Negative finance words
lmPositive (lmPos)	Positive finance words
lmModalStrong (lmMS)	Words that denote commitment
lmModalWeak (lmMW)	Words implying lack of commitment
lmUncertainty (lmUnc)	Uncertain finance words
Finance Variables	
Abnff	The abnormal Fama French returns computed by taking the difference of actual abnormal returns from predicted abnormal returns.
RETi,t	Predicted returns
FFi,t	The three Fama French factors, obtained from Gene Famas website
Other Variables	
bigDataFactor	I create the factor by following Loughran and McDonald (2011): bidDataFactor = (lmUncertainty + lmModalWeak + lmNegative + lmPositive) – (lmLitigious + lmModalstrong)

Table 2. Fama french abnormal returns

Panel A. Internet firms

Coefficient	Estimate	Std. Error	T-Stat	P-Value	
lmLitigious	0.0017312	0.0012142	1.426	0.154128	
lmNegative	-0.000931	0.0011059	-0.842	0.399994	
lmPositive	-0.003983	0.0011809	-3.373	0.000761	***
lmModalStrong	0.0040461	0.0011088	3.649	0.000271	***
lmModalWeak	-0.0008978	0.0014494	-0.619	0.535696	
lmUncertainty	-0.0012745	0.0012773	-0.998	0.318513	

Panel B. Non-internet firms

Coefficient	Estimate	Std. Error	T-Stat	P-Value	
lmLitigious	0.0094554	0.0023961	3.946	8.64E-05	***
lmNegative	-0.0026467	0.0030215	-0.876	0.38132	
lmPositive	-0.0023895	0.0024182	-0.988	0.32339	
lmModalStrong	-0.0006781	0.002327	-0.291	0.77081	
lmModalWeak	-0.007273	0.0028137	-2.585	0.00992	**
lmUncertainty	0.0104135	0.0023065	4.515	7.29E-06	***

Note. This table tests the explanatory power of the Loughran and McDonald word dictionaries against the Fama French abnormal returns for the dot-com internet stocks. In Panel A, I report results for only the Internet Firms, and in Panel B, I report results for only the Non-internet Firms. Results are based on the model below:

$$abn_{ff} = \alpha_0 + \beta_1 \cdot lmLit + \beta_2 \cdot lmNeg + \beta_3 \cdot lmPos + \beta_4 \cdot ModStrng + \beta_5 \cdot ModWeak + \beta_6 \cdot lmUncrt + \varepsilon$$

where all variables are defined in Table 1. Statistical significance is denoted as follows: '***' denotes 0% significance, '**' denotes 0.1% significance, '*' denotes 1% significance, '.' denotes 5% significance, and ' ' denotes no significance.

Panel B reports results for the non-internet firms. I draw attention to my observation that the results in Panels A and B are orthogonal. Whereas lmPositive and lmModalStrong are significant in Panel A, they are not significant in Panel B. On the other hand, the three factors significant in Panel B are not significant in Panel A. Thus the internet stocks, and the non-internet stocks, are being driven by different soft factors. Those factors which are significant in Panel B are lmLitigious, lmModal Weak and lmUncertainty, at 0.009554, -0.007273 and 0.010414 respectively.

I next create and test a more parsimonious model (Brau, Cicon, & Ferris, 2014). In their regressions of first day returns, Loughran and McDonald (2011) report that their word lists take the following signs: uncertainty (+), weak modal (+), negative (+), positive (+), legal (-) and strong modal (-). Based on these results, I conflate all six of the Loughran and McDonald scores into a single big data factor as shown below:

$$bigDataFactor = (uncertainty + modalweak + negative + positive) - (litigious + modalstrong) \quad (3)$$

I use this factor to then create my parsimonious model:

$$abn_{ff} = \alpha_0 + \beta_1 \cdot bigDataFactor + \varepsilon \quad (4)$$

where all variables are defined in Table 1. I report results in Table 3. In Panel A, I report results for the internet firms. The intercept and the bigDataFactor have about equal magnitude, thus my big data factor explains half of the total variance in the Fama French returns (0.011590) and the intercept explains the other half (0.011590) . Both variables are significant at better than 1%. Panel B, on the other hand, report that the non-internet firms are not explained by the bigDataFactor, and all variance is captured by the intercept. This finding suggests that over this period of time that it is only the internet firms which are being driven by media forces, not the non-internet firms.

Table 3. Fama french abnormal returns–the parsimonious model

Panel A. Internet firms

Coefficient	Estimate	Std. Error	t-stat	p-value	
Intercept	0.011418	0.001077	10.6	< 2e-16	***
Big Data Factor	0.011590	0.002547	-4.55	5.77E-006	***
R^2	1.12%				

Panel B. Non-internet firms

Coefficient	Estimate	Std. Error	t-stat	p-value	
Intercept	0.011941	0.002231	5.353	1.13E-007	***
Big Data Factor	0.005165	0.005333	0.969	3.33E-001	
R^2	0.17%				

Note. This table repeats the analysis in Table 2, but it replaces the six Loughran and McDonald (2009) parameters with a more parsimonious mode. In Panel A, I report the results for only the Internet Firms, and in Panel B, I report the results for only the Non-internet Firms. I accomplish this by fitting the model below: $abn_{ff} = \alpha_0 + \beta_1 \cdot bigDataFactor + \varepsilon$

where all variables are defined in Table 1. Statistical significance is denoted as follows: '***' denotes 0% significance, '**' denotes 0.1% significance, '*' denotes 1% significance, '.' denotes 5% significance, and ' ' denotes no significance.

The difference between the Bhattacharya et al. (2009) study and my study is best appreciated by analyzing Figure 2. This figure first plots the abnormal returns that are not explained by the Fama French three factor model as a red line. The blue line plots the results of Equation 4 (scaled to the same magnitude as the red line). Unlike in Bhattacharya, my model (blue line) closely follows the Fama French abnormal returns, with the exception of a deviation about three months prior to market collapse. This implies that my model may have the power to predict a collapse.

The power of my model to potentially predict an imminent market collapse is the most interesting part of this paper. It is not a finding that I expected. However, close inspection of Figure 3 shows that my model closely follows the dot-com market for most of the dot-com period. It is only about three months prior to the collapse of the dot-com market that my model deviates. I propose that up until this point, that media hype was driving the market. However, on or about December 7, 1999, the media started to draw back. At this point the market itself continued to overvalue internet stock, ignoring media that recommended prudence, and carried on the inertia of its previous exuberance.

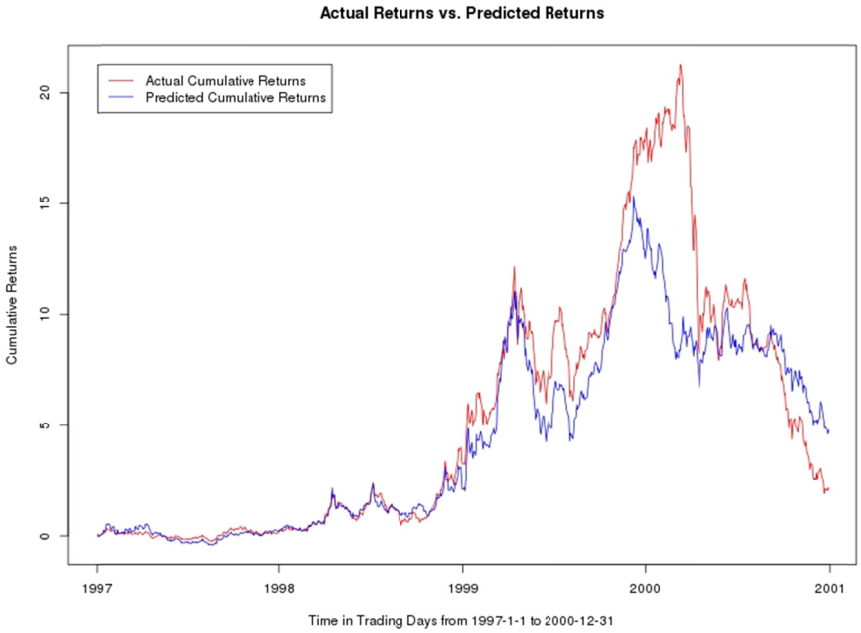

Figure 2. The explanatory power of big data over the dot-com abnormal returns

Note. In this figure I repeat the 'internet sample – non-internet sample' plot in Figure 3 of Bhattacharya et al. (2009), repeated in Figure 1 of this paper. I then add a control variable for the big data model developed in this paper and plot the result as the blue line (scaled to the same magnitude as the red line). This plot has two remarkable characteristics. First, the blue line closely tracks the red line for most of the plot. Second the blue line predicts a market collapse about three months before the red line actually collapses.

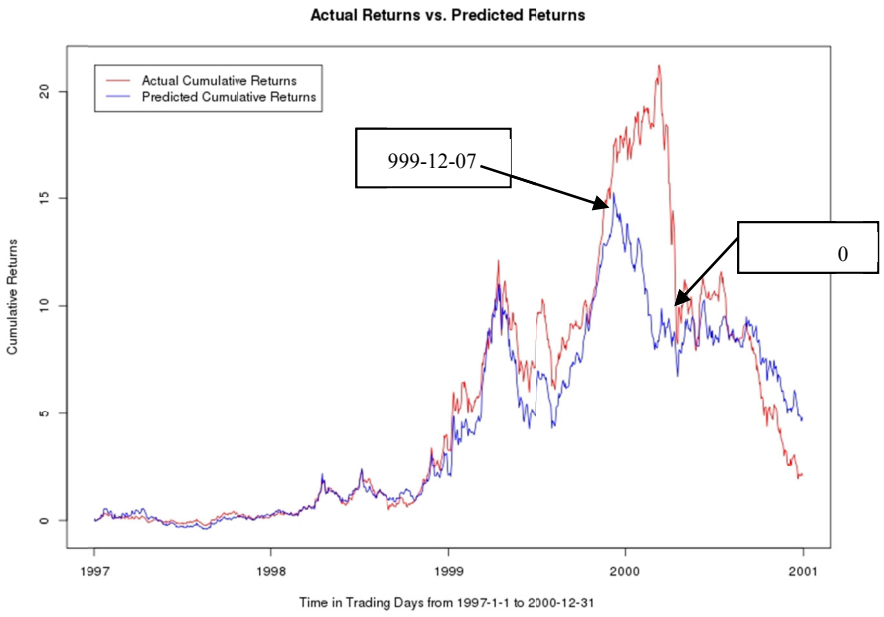

Figure 3. The predictive power of the big data model

Note. The Big Data model developed in this paper predicts the collapse of the dot-com bubble. The big-data model closely follows the market up until about December 7, 1999. At this point the big data model and market behavior deviates. My big data model predicts that the market should be collapsing, but the market continues onward for three months. This may be due to the inertia of market exuberance.

4. Conclusion

In this paper I develop a big data model of the dot-com market. I show that my model explains dot-com market performance, despite the fact that Bhattacharya et al. (2009) asserts that "media hype is unable to explain the stock bubble". I propose that my model succeeds where Bhattacharya et al. (2009) fails, based on three factors. First, I remove subjectivity in the research methodology by using a machine to read the news articles, not relying on humans to read them over years of time. Second, I do not treat news as merely 'good' or 'bad'. Instead I extract measures of positivity, negativity, uncertainty, commitment and litigiousness. Lastly, I do not accept all news articles that mention the firm anywhere within the article. Instead, I limit my sample to news articles in which the firm is mentioned in the article headline.

Surprisingly, my model also predicts the dot-com collapse: my model provides evidence that the market should have started falling on or about December 7, 1999. However the market continued to expand beyond that point for another three months. I propose that media was indeed driving dot-com market sentiment, but only until December 7, 1999. At that point the media began souring on dot-com internet stocks, however, market momentum continued to carry stock prices for another three months, well past the point where a prudent media observer would have pulled out.

Additional work needs to be performed on this study. First the control variables used by Bhattacharya et al. (2009) should be introduced. I do not expect that this will change my results materially because I add my big data variable at the same point where Bhattacharya adds his controls. Bhattacharya finds that the controls do not change his model or his results. Likewise, I expect that they will neither change mine. Secondly, my model should be applied to other bubbles as well to test if it has predictive power of them too, or if it merely reports an artifact of the dot-com era.

Acknowledgements

I am grateful to seminar participants at the Leir Charitable Foundation and at the New Jersey Institute of Technology. I thank the Leir Charitable Foundation for financial support for this paper.

References

Bhattacharya, U., Neal, G., Rina, R., & Yu, X. (2009). The Role of Media in the Internet IPO Bubble. *Journal of Financial and Quantitative Analysis, 44*(3), 657–682. http://dx.doi.org/10.1017/S0022109009990056

Brau, J., James, C., & Steve, F. (2014). *Fooled From the Start: Camouflageed Corporate Goverance and IPOs.* Working paper, Brigham Young University, New Jersey Institute of Technology, University of Missouri.

Brockman, P., & James, C. (2013). The Information Content of Managment Earnings Forecasts: An Analysis of Hard Versus Soft Information. *Journal of Financial Research, 36*(3), 147–174. http://dx.doi.org/10.1111/j.1475-6803.2013.12006.x

Campbell, G., John, T., & Clive, W. (2012). *The Role of Media in a Bubble* (pp. 461–481). Explorations in Economic History. http://dx.doi.org/10.1016/j.eeh.2012.07.002

Cicon, J., Jonathon, C., & Steve, F. (2013). Narayanan Jayaramen, Managerial Expectations of Synery and the Performance of Acquiring Firms: The Contribution of Soft Data. *Journal of Behavioral Finance.*

Cicon, J., Steve, F., Armin, K., & Gregory, N. (2012). European Corporate Governance: a Thematic Analysis of National Codes of Governance. *European Financial Management, 18*, 620–648. http://dx.doi.org/10.1111/j.1468-036X.2010.00542.x

Hanley, K. W., & Hoberg, G. (2009). The Information Content of IPO Prospectuses. *Review of Financial Studies, 23*, 2821–2864. http://dx.doi.org/ 10.1093/rfs/hhq024

Loughran, T., & McDonald, B. (2011). When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-K. *Journal of Finance, 66*, 35–65. http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x

Shiller, R. (2000). *Irrational Exuberance.* Princeton, NJ: Princeton University Press.

Tetlock, P. C. (2007). Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *Journal of Finance, LXII*(3), 1139–1168. http://dx.doi.org/10.1111/j.1540-6261.2007.01232.x

Note

Note 1. Tom Rosenstiel, Jeff Sonderman, Kevin Loker, Millie Tran (2014, March 17). The Personal News Cycle. *American Press Institute*. Retrieved from

http://www.americanpressinstitute.org/wp-content/uploads/2014/03/The_Media_Insight_Project_The_Personal_News_Cycle_Final.pdf