Fake News in Financial Markets

Shimon Kogan*
MIT Sloan School of Management
Interdisciplinary Center Herzliya

Tobias J. Moskowitz
Yale University
School of Management
NBER
AQR

Marina Niessner
Yale University
School of Management

March 2018

Preliminary and incomplete. Do not cite without permission.

* We thank Diego Garcia as well as conference and seminar participants at UCLA (Anderson), Rice University (Jones), University of Miami Business School, ASU Sonoran Winter Finance Conference, 3rd Annual News & Finance Conference, University of Colorado at Boulder, for their helpful comments and suggestions Contact emails: skogan@mit.edu, tobias.moskowitz@yale.edu, and marina.niessner@yale.edu

Abstract

While social media platforms, blogs, and other unmonitored media outlets are becoming a main source of news for many, they also offer scope for providing misleading or false information. We use two unique datasets and a linguistic algorithm developed to detect deception in expression, to examine the impact of fake news in financial markets. The first dataset is a set of paid-for articles obtained from an SEC investigation that are known to be false, that allow us to validate the linguistic algorithm. The second dataset applies the linguistic algorithm to quantify the probability of an article being fake on a much larger set of articles. We find a strong temporary price impact and subsequent reversals from the fake news articles for small firms, permanent negative price impact for mid-size firms, and no impact for large firms. In addition, for small and mid-size firms we find that around the release of fake articles, managers are more likely to issue press releases, file 8-K forms with the SEC, and buy stock in their own firm, hinting that such firms are possibly engaging in stock price manipulation. No such patterns are found for large firms.

1. Introduction

The potential role of new shared information platforms for information transmission is receiving ever increasing attention. Social media platforms, blogs, and other unmonitored media outlets are quickly becoming a main source of news for many people.[1] While such platforms can enhance the speed with which information is disseminated and lower the cost of obtaining information, they also offer scope for providing biased, misleading, or false information. One prominent example is the proliferation of "fake news," defined as hoaxes, frauds, or deceptions designed to mislead consumers of information.[2] As Alcott and Gentzkow (2017) argue, social media platforms enable content to be disseminated with no significant third party filtering or monitoring, allowing false information to be spread quickly through a vast social network. Vosoughi, Roy, and Aral (2018) find that fake news diffuses faster, deeper, and more broadly than actual news, in part because the fake news is often more extreme and exaggerated and designed to increase diffusion. Hence, the potential for fake news to infiltrate users' information sets quickly and extensively is becoming a major concern. One prominent example includes the potential influence fake news had on the 2016 U.S. Presidential election (Alcott and Gentzkow (2017)). Another example is a recent study by ReviewMeta (2016) that fake reviews on Amazon are misleading consumers toward various products (often paid for by the producers of the products) and becoming an increasing problem. The issue of fake content has become important enough that Amazon, Google, Facebook, Wikipedia, and many others are currently investing heavily to curb its dissemination.

The economics of fake news is an interesting area of study that is very recently receiving attention. Alcott and Gentzkow (2017) for instance, argue that fake news can arise in equilibrium where it is cheaper to provide than precise signals when consumers cannot costlessly infer accuracy. Extending the model of Gentzkow and Shapiro (2006) on media bias,

[1] According to a survey from the Pew Research Center (Gottfried and Shearer (2016)), 62% of American adults get news from a social media site.

[2] Facebook defines fake news as "hoaxes shared by spammers" for personal or monetary reasons.

they argue that fake news may generate utility for some consumers (who may enjoy fake stories or derive utility from content slanted toward their priors or biases as in Mullainathan and Shleifer (2005)), and where news is deemed higher quality if closer to the consumer's priors. When verifying accuracy is costly, this will lead to biased reporting in equilibrium (Gentzkow, Shapiro, and Stone (2006)), of which fake news can be one aspect. Alcott and Gentzkow (2017) argue that fake news may maximize short-run gains over long-term reputation. Consumption of fake news occurs in equilibrium because agents cannot verify the truth costlessly and the news is tailored to match the agent's priors. Aymanns, Foerster, and Georg (2017) provide an equilibrium model of an adversary using fake news to target agents with a biased private signal. The model shows that transmission of fake news is more successful when the adversary knows the agent's signal and information network. Conversely, when agents are aware of the adversary they discount news appropriately and make fake content less effective.

There is ongoing debate over the potential costs and importance of fake news. False content can impose private and public costs by making it more difficult for consumers to infer the truth. Alcott and Gentzkow (2017) point to the cost of consumers having less-accurate beliefs as the result of fake news, which could not only make them worse off but also possibly reduce positive social externalities from these media platforms. As a byproduct, consumers may also become more skeptical of legitimate news producers if they are hard to distinguish from fake news producers. Finally, these distortions could cause resource misallocation, such as misleading or misdirecting consumers toward worse products in the Amazon example or diverting votes in elections. These potential costs must be balanced against any welfare benefits from consumers deriving utility from fake news. However, there is very little evidence documenting these costs, their significance and where and when they arise.

In this paper, we examine the impact of fake news in financial markets – a setting where we can readily assess the costs of such news and where arguably there is little benefit to consumers of such news. Fake financial news, unlike political or social news, should not provide any utility from an entertainment or bias perspective as in Mullainathan and Shleifer (2005). The goal is to accurately assess the state of the world for financial gain. On the other hand, the costs of fake content here are clear in that if fake news causes less accuracy or erroneous financial decisions, we can directly measure the consequences and financial costs of those actions to the individual investor, as well as perhaps in aggregate from a resource misallocation perspective. For example, if fake financial news moves prices, then this would provide direct evidence of resource misallocation.

Of course, if financial markets are informationally efficient (Fama (1970)), where the cost of information is close to zero, then fake news should not matter at all and would have no bearing on prices. Hence, finding a significant price response to fake news suggests three things: 1) that markets may not be fully efficient, 2) that the cost of information (at least for that security) may be significant,[3] and 3) that fake news may be even more important in other settings, where competition for information and the ability to exploit it is less fierce. Our study, therefore, provides a unique test of market efficiency that circumvents the joint hypothesis problem – fake news under any equilibrium model of pricing should have no effect. Our experiment is the flip side of the classic event study test (Fama, Fisher, Jensen, and Roll (1969)): we conduct a "fake-event" study where the price response should be precisely zero and can measure the magnitude and duration of any price response. Second, while the marginal cost of information is at the heart of determining how informationally efficient financial markets are (Grossman and Stiglitz (1980)), attempts to estimate such costs are empirically elusive. Our study may perhaps provide a glimpse into the cost and speed of price discovery across firms by examining the impact of fake news events. Finally, given the competition for information and the ability to trade on that information in financial markets, if we find significant effects from fake news here, it stands to reason that its impact may

[3]The cost of information can be both a direct cost of gathering, processing, and analyzing information, as well as the indirect costs of misperceiving or misreacting to information stemming from psychological or behavioral biases (e.g., a mental processing cost that can include psychological barriers to interpreting information).

be even greater in settings where competition and arbitrage activity are more limited (e.g., consumer retail, political elections). Thus, the effects of fake news on financial markets may provide a lower bound in its potential influence more broadly.

Our study begins with a unique dataset of fake news articles. Rather than try to estimate information directly and its impact on prices, which is a product of both the cost of information and its interpretation and value placed by investors, we examine the impact of fake or false information on prices, where the price response should be zero under any equilibrium asset pricing model. Our sample consists of two datasets to estimate the impact of fake news. The first is a unique dataset of paid-for articles obtained from an industry "whistle-blower", Rick Pearson, a regular contributor on Seeking Alpha, a crowd-sourced content service for financial markets. Mr. Pearson went undercover to investigate other authors on the site and uncovered fake paid-for articles now being investigated by the SEC. The sample size is small, but the identity of fake news is clean – 171 articles by 12 authors covering 46 companies. We compare these to all other articles written by the same authors (334 in total) that were published on the same platform that may not have been fake – covering 171 stocks in total.

While the first sample represents our cleanest experiment, where there is no ambiguity in identifying fake articles, it is a small and narrow sample, which may make it more difficult to draw general conclusions. To address this issue, we also use a second dataset of hand-collected articles that were published on two (and eventually three) of the most prominent financial crowd-sourced platforms: Seeking Alpha and Motley Fool covering 203,545 articles from 2005 to 2015 for Seeking Alpha, and 147,916 articles from 2009 to 2014 for Motley Fool. Using a linguistic algorithm that scores the authenticity of an author's text, together with the first dataset with taged fake news, we probabilistically identify "fake" news. This creates a second and a much larger set of false news events. Importantly, we use our first and smaller dataset of *known* fake articles to validate the algorithm's ability to identify fake news stories. Having an unambiguous sample of fake articles from the undercover sting operation by the SEC is a key advantage, because without it the authenticity algorithm cannot be validated.

Echoing the importance of this statement are the challenges Google, Facebook, and Twitter are currently facing trying to identify fake news on their own platforms. For example, Google and Facebook are currently using human editors to evaluate content in the hopes of training an algorithm to identify false content systematically (Leong (2017)). Absent a set of known and identifiable fake articles, such endeavors have yielded little success. For the same reason, an investor at the time of the fake article's publication could not have constructed such an algorithm either, since the fake articles from our dataset were not yet known or identified.

Using the set of identifiable fake articles from the SEC to train our algorithm and cross-validate it, we find our type II error to be very low – less than 5% of articles are identified as false positives. Hence, our method for identifying fake news in the second dataset is quite conservative, where we classify only 2.8% of articles as being fake in our sample, with the frequency peaking in 2008 at 4.8%, but where we have high confidence that these articles contain false content. Hence, our method is designed to minimize type II errors at the expense of increasing type I errors, where we are likely missing many other fake articles.

Using the samples of truly fake articles and probabilistically fake articles, we investigate whether they impact the market and test various theories of fake news production. We find that the incidence of fake articles is slightly higher for small firms, and that the price response in markets is much larger for small and mid-size firms and negligible (precisely zero) for large firms. Small firm prices rise on fake news, which is predominantly positive, by 8% upon its release, which subsequently gets fully reversed over the course of a year. Hence, for small firms the market appears fooled by these articles initially, overpricing small firms with fake news by 8% on average, but then eventually corrects the mispricng. For mid-size firms, the price impact is negative immediately, and there is a permanent 4% discount associated with fake articles written about the firm, suggesting that mid-size firms having fake articles is a bad signal about the firm. For large firms, there is no price impact – initially or long-term – from fake articles written about it. These results suggest that the market is efficient with respect to large firms and appears inefficient for small firms, consistent with intuition suggesting that the cost of information (direct and indirect/psychological) is larger for small firms. Where the cost of information is lowest and competition is highest, we see no impact from fake news, and we show that production of fake news in equilibrium is consistent with these results, as paid-for fake content is done by small firms, but not large firms.

To better understand these results, it is useful to consider what motivates the production of fake news about firms. One motivation for the fake articles, which is related to how Rick Pearson went undercover and why the SEC is involved, is that the firms themselves may be orchestrating a promotional pump-and-dump campaign to manipulate the stock price. Another possibility, of course, is that rogue authors wish to create a false narrative about a firm for their own intentions, having no direct connection to the firm itself. To investigate the first possibility we look at a set of firm actions the firm may be pursuing at the time of the fake articles' release. For example, if these articles are part of paid campaigns by firms orchestrated by a public relations agency, then other actions taken by the firms around these events are likely to be present. We find that the fake news articles are often accompanied with press-releases by the firm among small firms, but not among large cap firms. We also find that 8-k filings are more likely to accompany the news of small firms, but not large firms. We further find that insider trading in the direction of the fake news (to take advantage of the price impact) is also more likely for small and mid-cap firms, but not large stocks. These results are consistent with a deliberate campaign by the firm to manipulate the stock price and take advantage of the price impact among small firms. For large firms, however, we do not find the same patterns, consistent with fake news about large firms being driven by authors outside of or unassociated with the firm.

We explore what characteristics of firms and articles are associated with the propensity of fake news as well as the magnitude of temporary price impact from the fake news, to test theories of fake news and to better understand the variation in information environments across firms and articles. We find that for small and mid-cap firms, high past volatility and volume are associated with higher propensity of fake news, consistent with more retail investor attention (see Barber and Odean (2007)). This finding could be consistent with fake news production catering toward less sophisticated investors who may derive psychological utility from fake news correlated with their priors (Alcott and Gentkzkow (2017)). Likewise, article email circulation, a proxy for the popularity of the firm among retail investors, is positively related to fake news, consistent with fake news being more effective with a broader network (Aymanns, Foerster, and Georg (2017)).[4] In addition, other proxies for attention and strength of network, such as larger analyst coverage, larger number of stock tweets mentioning the firm, more media coverage, and more readers' comments, are all associated with higher propensity of fake news. We also examine other actions taken by the firm and its insiders prior to the fake news event date, such as share purchases, insider sales, and initiation of press releases. These actions are more prevalent and coincide with the fake news for small and mid-cap firms, but not for large firms, consistent with smaller firms engaging in short-term profit maximizing schemes from fake news.

We also find that these variables are associated with the magnitude of the positive temporary price impact found for small firms, and the permanent negative price impact found for mid-cap firms. The price impact for large firms is non-existent and does not vary with any of these variables. Specifically, for small firms, we find that the price reaction is stronger for firms with higher turnover and volatility, less media coverage, more retail ownership, and more frequently talked about on StockTweets. These characteristics predict both the magnitude of the initial price rise as well as the size of the subsequent reversal. We also find that firms whose articles usually get a lot of comments have a quicker return reaction than firms with fewer comments. Similarly, if the author of the article has many followers, or if many readers subscribe to that firm's articles, the return reaction is much stronger for those firms. Finally, the amount of trading by insiders and the number of press releases issued by the firm affect the differential return reaction, where firms whose management engages in insider buying and press releases around the time of the article have a stronger return

[4] We have data on email circulation for Seeking Alpha articles only.

reaction.

For mid-size firms, these same characteristics also predict the magnitude of the price reaction. However, the same characteristics associated with small firms having a more positive return reaction are associated with mid-size firm's more negative permanent price reaction. These results suggest that both small and mid-cap firms may be engaging in stock manipulation strategies to pump up the share price and take advantage of the higher price by issuing shares or buying shares before the run-up. In the case of small cap stocks, the market appears to be fooled by this scheme, causing a temporary price increase that subsequently gets reversed. For mid-cap stocks, however, one interpretation is that the market does not get fooled because markets are more efficient for these larger stocks (e.g., the cost of information for these stocks is lower). Instead, the market identifies the news as fake and reacts immediately to the fake news in a negative way by permanently discounting the firms's share price. This evidence is consistent with the reputational story of Alcott and Gentzkow (2017) where fake news producers may sacrifice longer-term reputational capital in lieu of short-term capital gain. In the case of mid-cap stocks, the market deciphers the fake news and punishes the firm with a permanent price discount.[5]

For large cap firms, they neither seem to be initiating or taking actions in conjunction with the fake news articles and there appears to be no market price reaction of any kind to fake articles written about large firms. These results are consistent with the market being efficient with respect to fake news about large firms and as a result, large firms not attempting to engage in a futile effort to manipulate the share price. Rather, it appears that the fake articles about large firms are being written by rogue authors unaffiliated with the firms in an effort to increase their own utility.

Our study provides evidence on the prevalence and effect of promotional articles from

[5]Of course, it's also possible that mid-cap firms pumping scheme actually works if the returns would have been even worse had they not initiated the promotional articles. Hence, another interpretation is that the mid-cap firms fool the market, too, but only do so when other bad news about the firm is present. This narrative is less consistent with the data, however, since we find no evidence of other bad news associated with mid-size firms around the time of the articles.

crowd-sourced financial platforms that continue to grow and gain attention. How important are these platforms and what impact are they having, and specifically, how pervasive and important are fake articles on these platforms? While these social news networks may simply be a side-show for financial markets, our results indicate that there is significant price impact from these promotional fake articles. The results are consistent with other findings suggesting that crowd-sourced services can impact markets (Hu, Chen, De, and Hwang (2014), who show that information content and sentiment from these services predicts returns) and suggests that if fake content can impact financial markets in the U.S., it is likely to have an even greater influence in settings where the cost of information is higher and the ability to exploit it is more limited, such as on-line consumer retail reviews, crowd-funding, elections, news cites, etc.).

The rest of the paper is organized as follows. Section 2 details our sample on fake news articles and presents our methodology for identifying fake news. Section 3 examines the market's response to fake news, including both temporary and permanent price responses, as well as the drivers of differential market responses across different types of firms. Section 4 then examines whether managers are more likely to issue press releases and file 8-K forms around publications of fake articles, and also whether they engage in insider trading. Section 5 concludes the paper.

2. Data and Identifying Fake News

We describe the data we obtain on fake articles, our algorithm and its validation, and provide an example of a specific fake article and its consequences.

2.1. Obtaining Fake Articles

The popularity of financial crowd-sourced platforms such as Seeking Alpha, Motley Fool, TheStreet, etc., has grown exponentially over the last fifteen years. For example, Seeking Alpha went from having two million unique monthly visitors in 2011 to over nine million by 2014. While this innovation has allowed for unprecedented levels of 'democratization' of financial information production, some concerns have been raised about these platforms being susceptible to pump-and-dump schemes, as they are frequented by retail investors, and since many authors on these platforms use pseudonyms instead of writing under their real name.[6] Whereas it is theoretically legal for an author to talk up or down a stock that she is either long or short, she has to disclose any positions she has in the stock in a disclaimer that accompanies the article. While many authors add such disclaimers to their articles, that is not something the platforms actually verify. What is *illegal*, according to Section 17b of the securities code, is to fail to disclose any direct or indirect compensation that the author received from the company, a broker dealer, or from an underwriter[7].

Even though it does not come as a surprise that such promotions and pump-and-dump schemes exist, they can be hard to identify. Even harder is to prove that the authors of such articles received payment for writing them. Our analysis starts out with a unique dataset of paid-for articles obtained from an industry “insider," Rick Pearson. Rick, who is a regular contributor to Seeking Alpha, was approached by a PR firm that helps promote stocks. The PR firm asked him to write articles for a fee without disclosing the payment. Rick went undercover to investigate other authors and has uncovered many fake, paid-for articles where the authors did not disclose their compensation. These fake articles were subsequently taken down by the platforms where they originally appeared and the firms are now being investigated by the SEC. The SEC had since filed two lawsuits: first in 2014 and subsequently in 2017 against authors of promotional articles and the PR firms who were paying the authors to generate those articles.[8] Rick has kindly shared with us the articles that he has determined to be fake, providing us with 111 fake articles by 12 authors covering 46 companies. We also were able to obtain a second set of promotional articles. During the investigation, the SEC lawyers were able to identify further articles that were paid for by

[6] Even though the platforms claim that they always know the true identity of the author, in case that information is subpoenaed by the SEC.

[7] In June 2012, Seeking Alpha announced it would no longer permit publication of articles for which compensation had been paid.

[8] http://securities.stanford.edu/filings-documents/1051/GBI00_01/20141031_r01c_14CV00367.pdf and https://www.sec.gov/litigation/complaints/2017/comp23802-lidingo.pdf

the stock promotion firms.[9]. We contacted Seeking Alpha, and they kindly shared 147 of those articles with us. Of those, we were able to match 60 with CRSP dataset (the rest of the articles were about firms that traded over the counter). So our final dataset of for-sure fake articles consists of 171 articles written by 20 authors about 47 firms.

We furthermore were able to obtain all other articles written by the same authors (334 in total) that were published on Seeking Alpha, many of which presumably not paid-for promotional articles, as a baseline comparison for the same authors. These other articles were often written about large firms (171 stocks in total), which as we will show, are less likely to engage in this sort of stock promotion. Furthermore, authors need to establish a reputation writing non-promotional articles before they can write (and get away with) pump-and-dump type articles.

2.2. Example of a Fake Article

To illustrate the process and the impact of fake articles, we provide an example of Galena Biopharma Inc., one of the companies that hired a PR firm to order paid-for promotional articles about its stock. Several very positive articles about Galena appeared on Seeking Alpha and other platforms from 2012 to 2014, which coincided with a substantial run-up and a subsequent drop in Galena's stock price. Figure 1 shows the price of Galena's stock from 2012 to 2015 in light blue, and the appearances of promotional articles in red. Over that time period six identified promotional articles appeared about Galena, with four of them published towards the end of 2013 and early 2014. The four fake articles were all written by the same author, John Mylant. John Mylant had been an active contributor to Seeking Alpha since 2009, even though since this incident all of his articles have been taken down by Seeking Alpha. As the graph shows, Galena's stock price started to increase drastically, when several fake articles were published, more than tripling in 4 months, before it plummeted back down, once the promotional articles stopped.

[9]The full list can be found here: https://ftalphaville-cdn.ft.com/wp-content/uploads/2017/04/10231526/Stock-promoters.pdf

One natural question that follows is why the companies pay for these promotional articles. In Figure 1 in dark blue we plot all instances of insider trading between 2012 and 2015 (an indicator variable for whether any insider buys/sells were reported to the SEC through Form 4). The graph shows that insiders executed trades after the stock price almost tripled and right before the stock price crashed again. The SEC brought charges against Galena and its former CEO Mark Ahn "regarding the commissioning of internet publications by outside promotional firms." Mr. Ahn was fired in August 2014 over the controversy, and in December 2016, the SEC, Galena, and Mr. Ahn reached a settlement. The example of a promotional article about Galena is shown in Appendix A, and the 8-K form documenting the settlement in presented in Appendix B. Interestingly, if one were to search for this promotional article now, Seeking Alpha just displays a message saying "This author's articles have been removed from Seeking Alpha due to a Terms of Use violation."

2.3. Further Identifying Fake Articles – LIWC and the Authenticity Score

While the unique data of fake articles is illustrative of the phenomenon, the sample size is small and it is difficult to draw more general conclusions based on it. The goal of our paper is to estimate the prevalence and the effects of these fake articles on financial markets. In order to do so, we manually download all articles that were published on two of the more prominent financial crowd-sourced platforms: Seeking Alpha and Motley Fool. For Seeking Alpha we obtained 203,545 articles dating from 2005 to 2015 and for Motley Fool we have 147,916 articles dating from 2009 to 2014.

To understand how pervasive the phenomenon is in general and what effect fake articles have on financial markets we develop an objective and scalable measure that captures the authenticity of the article. To that end, we use LIWC2015 (Pennebaker et al. (2015)). LIWC is a linguistic tool that focuses on individuals' writing or speech style, rather than content, and thus appears to be uniquely adept at measuring individuals' cognitive and emotional states across domains. Specific to authenticity, Newman et al. (2003), use an experimental setting to develop an authenticity score based on expression style components. While the exact formula for the authenticity score is proprietary, James Pennebaker describes which linguistic traits are associated with honesty in his book "The Secret Life of Pronouns" Pennebaker (2011). In particular he finds that truth-tellers tend to use, for example, more self-reference words and communicate through longer sentences compared to liars. Intuitively, when people lie, they tend to distance themselves from the story by using fewer “I" or “me"-words. Furthermore, people use fewer insight words such as *realize*, *understand*, and *think*, and include less information about time and space. On the other hand, liars tend to use more discrepancy verbs, like *could*, that assert that an event might have occurred, but possibly did not. James Pennebaker and co-authors then use a combination of these linguistic traits to generate the authenticity measure in LIWC.

2.4. Validation

Given that the LIWC authenticity score was not developed in the context of financial media, one may be skeptical about its ability to distinguish fake from non-fake articles in Finance. After all, financial blogs and articles tend to point to facts, trends, and figures, which are different from narratives. To address this concern, we start out by validating the LIWC authenticity score using the small sample of 171 fake articles and 334 non-fake articles, all written by the same set of authors. That is, we compare the LIWC authenticity score, which is normalized between 0 and 100, for the two samples. The difference in the LIWC authenticity score across the two samples is both economically and statistically large. Relative to an average authenticity score of 33 for non-fake articles, fake articles had a much lower average score of 19 (statistically significant at 1% level). The density plots in Figure 2, Panel A illustrate how different the two distributions are. It is important to note that we control for any differences the authors' writing style may have on the authenticity score, as the sample consists of both fake and non-fake articles written by the *same set of authors*. To provide more specific examples, in Panel B, we provide the density plots of authenticity scores for two specific authors: *John Mylant*, and *Equity Options Guru*. We can see that while some of the non-fake articles also had a low authenticity score, most of the fake articles had a very low authenticity score.

While the exact composition of the authenticity score is proprietary, several language characteristics are associated with being more or less authentic (described in Pennebaker (2011)). In Table 1 we provide a summary of those characteristics for the promotional and non-promotional articles (written by the same authors). *For-sure Fake* articles are articles that have been shared with us by Rick Pearson and that were subpoenaed by the SEC. *Non-fake* articles, are articles that were written by the same authors, but about larger firms, that are unlikely to be promotional. We display the number of articles in each category as well as the mean of the *Authenticity* measure from LIWC. From the table, we see that the promotional articles' authenticity score is about half the size of non-promotional articles. We also report the means of several other variables provided by LIWC to help better understand the authenticity score. In particular we display the means of the average of the *1st person singular* measure (examples: I, me, mine), *Insight* measure (examples: think, know), *Relativity* measure (examples: area, bend, exit), *Time* measure (examples: end, until, season), *Discrepancy* measure (examples: should, would), and the average number of words per sentence. According to research by James Pennebacker and co-authors, when people lie they tend to use fewer self-referencing words, fewer words per sentence, fewer insight and relativity words, and more discrepancy verbs. The results in the table line up well with those findings: fake articles' self-referencing score is about half of non-promotional articles, and fake articles have a lower insight, and relativity scores, and higher discrepancy score. It's important to note that the promotional articles that we obtain from Rick Pearson and from Seeking Alpha (that have been subpoenaed by the SEC) are crucial to being able to use LIWC to identify fake articles in Finance, as they provide an out-of-sample test for a methodology that was developed outside of finance.

2.5. Probability of Being Fake

The above validation demonstrates that the LIWC authenticity score has the ability to distinguish between fake and non-fake articles. At the same time, it is not clear how to interpret the cardinal nature of the score – what does a 14 point difference in authenticity score mean? Ideally, we would measure the *probability* of an article being fake, but the LIWC authenticity score is not a probability. To provide a more direct interpretation of the results and their economic meaning, we develop a mapping of the authenticity score into the probability space. Starting with the validation sample and applying Bayes rule to the overall sample of Seeking Alpha articles, we create a function that maps the authenticity score into a conditional probability of an article being fake.

Specifically, let S be the authenticity score and F (T) denote a fake (true) article. The key to our ability to tranform the ordinal authenticity score into a cardinal probability measure is the fact that in the validation sample, we know which articles are F and which ones are T. Thus, we can compute $Prob(S|F)$ and $Prob(S|T)$. From Bayes rule, we know that:

$$Prob(F|S) = \frac{Prob(S|F)Prob(F)}{Prob(S|F)Prob(F) + Prob(S|T)Prob(T)}.$$

If we integrate $Prob(F|S)$ over the empirical distribution of scores, we get $Prob(F)$. The issue, of course, is that $Prob(F)$ is also an input in the calculation. The solution to the fixed point problem can be found assuming that $Prob(F)$ in the sample is representative of $Prob(F)$ in the overall population.

We apply this approach to the entire sample of Seeking Alpha articles published between 2005 and 2015, over 203,000 in total, covering over 7,700 firms. There are a number of findings that arise. To start with, we observe the resulting mapping of LIWC authenticity scores (S) into the conditional probability of being fake ($Prob(F|S)$). As Figure 3 depicts, the relation between the two is highly non-linear. As the figure shows, an authenticity score of 31 – the average for the non-fake articles – corresponds to a conditional probability of being fake of close to zero, while an authenticity score of 17 – the average for the fake articles – corresponds to a significant probability of being fake of 3.6%.

This has two important implications. The first is that using the LIWC authenticity score is not equivalent to using the probability. The second is that the sharp increase in probability in the very low authenticity range suggests that articles can be well classified into fake and non-fake ones. Put differently, using a probability cutoff can be an efficient way of separating articles into various types.

Next, we use these results to answer a key question: how pervasive are fake articles on financial crowd-sourced platforms? We find that the unconditional probability of a Seeking Alpha article being fake is 2.8%, peaking at 4.8% in 2008 and dropping to 1.6% in 2013.

We next examine how accurate our method is at identifying fake news. We take the 505 articles (171 promotional and 334 non-promotional articles written by the same authors), generate an authenticity score for them, and calculate their probability of being fake. We then use the cutoff of $Prob(Fake) > 20\%$ to classify articles as being fake[10]. Our algorithm classifies 18 out of the 505 articles as being fake. Out of those 18 articles 17 are actual promotional articles. This suggests that our Type II error rate is very low - we have very few false positives, and our method is very conservative. In other words, while we most likely miss some promotional articles, the ones our algorithm identifies as fake are highly likely to be truly promotional articles.

We classify articles with $Prob(Fake) < 1\%$ as being non-fake. Our algorithm identifies 165 articles (out of 505) are being non-fake. Of those 14 are actually promotional articles, and the rest are not. Therefore, our Type I error is about 5%, which is quite low. So when we look at articles that our algorithm identifies as being "non-fake," most of them happen to be non-promotional. We exclude articles with $1\% \leq Prob(Fake) \leq 20\%$ from our analysis, as for those articles Type I and Type II errors will be large, and would make our analysis noisy.

Table 2 shows, for different types of articles on Seeking Alpha and Motley Fool, summary statistics of various LIWC textual measures, the probabilities of being fake, and firm characteristics of the covered firms. *For-sure Fake Articles* are the articles that have been

[10] Our results are not sensitive to the specific cutoffs

shared with us by Rick Pearson, who went undercover to expose authors who were being paid to write promotional articles for companies, without disclosing the payments, and also articles that were subpoenaed by the SEC and shared with us by Seeking Alpha. *Seeking Alpha Articles* and *Motley Fool Articles* are regular articles that we downloaded from Seeking Alpha and Motley Fool. Of those articles, *Fake* articles are articles whose probability of being fake (according to our measure) is higher than 20%, *Non Fake* articles are articles with probability of being fake less than 1%, and the rest are classified as *Other*.
In Panel A, we display the number of articles in each category as well as the mean of the *Authenticity* measure that we use to construct the probabilities of being fake. In all instances, the authenticity score is much lower for fake articles, than for non-fake articles in both Seeking Alpha and Motley Fool datasets. The differences are statistically significant. The authenticity scores for *Fake* Seeking Alpha and Motley Fool articles are especially low, which is by construction.

We also report the means of several other variables provided by LIWC to help better understand the authenticity score. In particular we display the means of the average of the *1st person singular* measure (examples: I, me, mine), *Insight* measure (examples: think, know), *Relativity* measure (examples: area, bend, exit), *Time* measure (examples: end, until, season), *Discrepancy* measure (examples: should, would), and the average number of words per sentence. While LIWC doesn't disclose details about their authenticity measure, Pennebaker (2011) suggests that these variables are important linguistic differentiators between honest and deceptive reports. Fake articles, as identified by our algorithm, have a much lower fraction of *1st person singular* then the non-fake articles, suggesting that the authors try to distance themselves from the article. The fake articles also have a lower insight and relativity scores relative to non-fake articles, suggesting that the authors seem to draw fewer insights in the articles, and also reference time and space less.

In Panel B, we display the average probability of being fake for each of the article categories. For Seeking Alpha and Motley Fool, the difference in magnitudes of the probability of being fake are by construction.

In Panel C, we display the average fraction of retail investors, the average number of analysts covering the firm, and the average firm size (in Millions of dollars) for the respective article groups. For-sure fake articles tend to cover firms with a higher fraction of retail investors, whereas the fake versus non-fake articles seem to target firms with similar fraction of retail investors. Similarly, for-sure fake articles tend to concentrate on smaller firms with low analyst coverage, which is not the case for fake and non-fake articles. Finally, we examine whether fake articles tend to cluster in specific industries. In particular, in Table 3, we separate articles by the 12 Fama-French industries that the firms the articles are about belong to. Some interesting patterns emerge. For articles provided to us by Rick Pearson, and articles that were subpoenaed by the SEC and shared with us by Seeking Alpha, 81% of for-sure fake articles are about firms in the *Healthcare* industry. This finding is not too surprising as these articles came from authors who were hired by two PR firms, and concentrated on the healthcare industry. For the non-fake articles, the majority of firms belong to *Business Equipment*, *Healthcare*, *Finance* , and *Manufacturing* industries. The distribution of Fake versus Non-Fake articles on Seeking Alpha and Motley Fool is similar to the Non-Fake articles in column (2), with majority of them coming from *Business Equipment*, *Finance*, and *Healthcare*.

2.6. Other Datasets

We investigate the motivation behind these fake articles, where one hypothesis is that these campaigns are ordered by firms and orchestrated by a PR agency. To test this hypothesis, we obtain a dataset of press releases from RavenPack from 2001 to 2015 and collect message volume for a given firm from a Twitter-like platform called StockTwits. If firms are coordinating these articles, they may issue press releases simultaneously to provide material for the promotional articles and may also start Twitter campaigns to reinforce the messages in the fake articles.

We obtain stock price data from CRSP and firms' financial information from COMPU-

STAT. We also obtain data on insider trades from Form 4 from Thomson Reuters to see if insiders are trading around these events. Finally, we obtain dates of SEOs from the SDC Platinum database to look at firm equity issuance around these events.

3. Market Impact

Using the sample of promotional articles as well as the set of probabilistically fake articles from the broader sample, we investigate the market's response to fake news.

3.1. Return Reaction

First we examine the return reaction to the promotional articles that were provided to us by Rick Pearson. We separate the firms into small and mid-size firms and examine the firms' return response to the promotional articles. We classify a firm as small, if its market cap is in the bottom 10th percentile of NYSE stocks, and as medium if it's in the 20th-90th percentile by market cap among NYSE stocks. The cumulative abnormal returns, measured as equal-weighted 4-factor residuals, are constructed starting the day after the article was published until 251 trading days after the article was published. We generate equal-weighted *Mom*, *SMB*, *HML*, and *Mkt* factors, and estimate betas for a given stock i for day t using the window $t - 252$ to $t - 1$. We then use those betas to calculate the residual (abnormal) cumulative returns for stock i for days $t + 1$ to $t + 251$.

Figure 4 plots the cumulative abnormal returns for the promotional articles for small (blue line) and mid-size (red line) firms. Out of the 111 articles that we have, 69 are about small firms, 35 are about mid-size firms, and 7 are about large firms. Returns for small firms increase after the article is published, reaching as much as 20%, cumulatively, after about 60 days, before giving up all the gain, and ending up with a 10% loss towards the end of the year. The permanent price impact of -10% for small firms indicates that once the market figures out the news is fake, investors view this as a bad signal about the firm. Interestingly, for mid-size firms, there is no gain followed by reversal - the price starts dropping after the fake article comes out, and continues to decrease throughout the year. These results suggest that for both small and mid-size firms, the fact that management is trying to prop up the stock price with promotional articles is a signal for deteriorating underlying performance. However, either due to larger limits to arbitrage or to less sophisticated investor base, the promotional articles are successful at temporarily propping up the stock price of small firms, but not of mid-size firms.

The articles we obtained from Rick Pearson are a small sample. We next examine the market response to articles that we classify as fake using LIWC. Since our analysis is at the company-day level, we need to define whether a company had a promotional article on a given day. In order to do so, we calculate the average authenticity score among all articles for a given company/day, and define that a company had fake articles on a given day, if the average authenticity score translates into the probability of being fake of 20% or more, and that a company had no fake news, if the average authenticity score translates into the probability of being fake of less than 1%.

Figure 5 plots the difference between abnormal cumulative returns following days with fake articles, relative to days with no fake articles separately for small, mid-size, and large firms in our sample (that have at least one fake article). As the blue line shows, the returns for small firms increase for 6 months by about 8% following a fake article, relative to a non-fake article, and then revert back to their original level. Whereas the returns for mid-size firms (red line) start dropping almost immediately, and come to a steady state of -5% after about 10 months. It's important to note that small firms experience temporary positive returns following fake articles, whereas mid-size firms see a decrease in returns, which is very similar to return patters following promotional articles that Rick Pearson shared with us (shown in Figure 5), which helps to corroborate the patterns. For large firms (green line) the returns appear to first go up and then decrease, but the magnitudes are quite small – 50 to 100 basis points and are not statistically different from zero. Suggesting that the markets for those firms are quite efficient, and also that the articles that we identify as fake are probably one-off rogue investors, rather than those companies launching promotional campaigns. The promotional articles that we obtained from Rick Pearson only included a few firms in this size category.

Next, we examine whether the patters in cumulative abnormal returns for different-sized firms we observe in Figure 5 are statistically significant. In order to do so, we estimate the following model:

$$Ret_{i,(t+1,t+T)} = \alpha + \beta Fake_{i,t} + \varepsilon_{i,t}$$

where $Ret_{i,(t+1,t+T)}$ are cumulative abnormal 4-factor returns for firm i, from 1 day after the fake article was published until T days, where T is either 51, 101, 151, 201, or 251. The results are presented in Table 4. As we saw in Figure 5, for small firms, the returns in the first 6 to 7 months following fake articles, are more positive than following non-fake articles, and the difference is statistically significant. This gain disappears after about 10 months and is basically 0 after a year. For mid-size firms, the returns start decreasing following the publication of fake articles, relative to days with non-fake articles, and continue to decrease for about 10 months, before coming to a steady state at around -4%. Finally, for large firms, the difference is very small (60 basis points after 3 months) and is barely statistically significant.

3.2. Determinants of Market Reaction

S o far, we have examined differential return reactions to fake articles relative to non-fake articles for small, mid-size, and large firms. Next, we examine whether the proxy for attention that the firms' articles usually get has an effect on the differential market reaction. An alternative approach to measuring return reaction to the release of news, in this case fake and non-fake articles, is to observe their impact on contemporaneous and subsequent volatility. This test is potentially more powerful as it eliminates the need to sign articles directional impact on returns. Specifically, we use the 4-factor residual returns over the 3 day period following the release of articles, squared, as a measure of market reaction to the release of the article (skipping a day is a conservative way of getting around the exact time of article release).

In table 5 we regress this measure on various article and author characteristics. In all cases, we include firm and year fixed effects as these are prime drivers of idiosyncratic volatility. The first column reports how characteristics of the article relate to its impact on volatility. We find that the number of comments is negatively associated with impact on volatility, potentially consistent this being a proxy for disagreement, while both the number of users that received the article and title length are positively associated with article impact. The latter is inconsistent with Umar (2017), which argues that article title length is negatively associated with the attention it receives.

The second column focuses on authors' observed characteristics. Consistent with the idea that article impact is related to author's reputation, we find that articles written by authors with high lagged average article volatility impact and with a larger number of previous articles have a bigger impact; tenure does not appear to be important. Column three puts together all these variables and the results are qualitatively unchanged.

Next, we focus on the effect fake news have on volatility. In column four we add a fake news dummy and observe that fake news have a bigger impact on volatility compared to non-fake news, controlling for article characteristics and author reputation. The economic magnitude of the coefficient is quite large – fake news increase idiosyncratic variance by around 10.5% compared to non-fake news. Interestingly, the fake news dummy drives out the relation between title length and volatility, suggesting that a key driver of article length in our sample is driven by the release of fake news. We further explore the market impact of fake news by interacting them with author reputation (columns 5-7) and find that reputation interacts positively with fake news and negatively with article length. Finally, we find that the propensity of the author to publish fake news has a reputational cost as it lowers the effect of subsequent fake articles.

4. Insider Trading and Press Releases

So far we have provided evidence that fake articles can influence asset prices. Next we examine whether firms take any actions to facilitate the promotion of the articles, and what incentives managers have to try to pump up their stock price. In particular, we examine whether companies are more likely to issue press releases around the time of fake articles, to give the authors of the fake articles some material to write about. We further look at whether managers engage in what's called "pump-and-dump" schemes, where one acquires shares at a low price, then inflates the price through fake articles, and then sells the stock. While we cannot observe trades for regular investors, this is something we can observe for managers, as they have to report their insider trades to the SEC using Form 4.

4.1. Press Releases

When authors write fake articles about the firms, it would make the articles much more believable if the firm issued a press-release or filed an SEC filing at the same time. The press release and/or an SEC filing can also provide some material facts for the author to write about, and the author can then potentially exaggerate what those facts mean for the future of the firm. In order to examine this question we regress whether there were fake articles in a given week, on whether there was a press-release or an SEC (8-K filling[11]) in the prior week, the week of and a week after the fake articles are published. We define an indicator variable, *Fake Article*, to be 1 if the probability of being fake associated with the average authenticity score for articles written about the firm in the given week is great than 20%. We perform our analysis separately for small, mid-size, and for large firms. Small firms are defined as firms in the bottom 10th percentile of NYSE firms, mid-size firms are defined as firms in the 20th-90th percentile of NYSE firms, and large firms are defined as firms above the 90th percentile of NYSE firms.

We find that small firms are substantially more likely to have fake articles written about

[11] An 8K-filling is a form a firm has to file with the SEC if a material event has occurred at the firm within the past 5 business days.

them in the week before, the week of, and the after they issue a press release or file an 8-K form with the SEC. The coefficients become insignificant if we go out further than those weeks. Mid-size firms have an increased probability of having fake articles written about them in the week of the press release or an 8-K filing, and the is no effect for large firms. These results are consistent with the anecdotal evidence that companies often issue press releases to provide some material for the fake articles. Large firms are unlikely and do not seem to engage in this behavior. These results are consistent with a deliberate campaign by the firm to manipulate the stock price and take advantage of the price impact.

4.2. Insider Trading

First, at the monthly level, we regress an indicator variable for whether a firm had predominantly fake articles in a given month on whether the firm was a net insider buyer or a net insider seller in the previous month and in the contemporaneous month. A firm is a net buyer (seller) if insiders bought more shares, in dollar value, than they sold in a given month (sold more shares than they bought). We define an indicator variable, *Fake Article*, to be 1 if the probability of being fake associated with the average authenticity score for articles written about the firm in the given month is great than 20%. We perform our analysis separately for small, mid-size, and for large firms. Small firms are defined as firms in the bottom 10th percentile of NYSE firms, mid-size firms are defined as firms in the 20th-90th percentile of NYSE firms, and large firms are defined as firms above the 90th percentile of NYSE firms.

We do not tabulate the results for brevity. We find that for small and mid-size firms, insider buying and issuing of press releases is strongly associated with the prevalence of fake articles in the same month. Whereas there is no association between insider trading and fake articles in the previous and the following months.

The above analysis shows that insiders buy stock and issue press releases in the same month as fake articles come out. Next, we zoom in on the weeks around fake articles and examine the timing in more detail. We run similar regressions as we did at the monthly level, except now everything is defined at the weekly level. Therefore, we regress an indicator variable for whether a firm had predominantly fake articles in a given week, on whether insiders were net buyers or net sellers in the week before, the week of, and the week after the fake news came out. *Net Buyer* (*Net Seller*) is an indicator for whether insiders bought more shares, in dollar value, than they sold in a given week (sold more shares than they bought). We define a dummy variable for whether a firm had predominantly fake articles in a given week as 1 if the probability of being fake associated with the average authenticity score for articles written about the firm in the given month is great than 20%. We perform our analysis separately for small, mid-size and for large firms.

The results are presented in Table 7. For small firms and mid-size firms, insiders don't trade in the weeks leading up to fake articles, trade a little bit in the week before the fake article, and then start actively buying the week of and the week after the fake articles come out. These findings are consistent with Figure 1, where insiders start buying Galena's stock around/after fake articles come out. We do not find a similar result for large firms.

5. Conclusion

Examining the impact of false information on prices using our unique datasets of fake articles, our novel test of market efficiency finds that markets respond to erroneous information in small stocks, possibly leading to potential price manipulation. The "non-event" studies we conduct find strong temporary price impact and subsequent reversals from fake news for small firms that coincide with insider trading and firm press releases, that predict the magnitude of the price reaction and reversal. We find similar results for mid-size firms except there is no temporary price increase and only a permanent price decrease associated with fake news, especially when coordinated with insider trades and press releases by the firm. Large cap stocks exhibit none of these patterns nor any price impact from the fake articles.

The evidence suggests that markets are efficient with respect to fake news for large and possibly mid-cap firms, but is inefficient for small cap stocks, consistent with information costs being greater for smaller firms. Small firms therefore engage in possible price manipulation that temporarily props up the share price and eventually reverses over the course of the year. Mid-size firms seem to engage in similar behavior, but the market isn't fooled and applies an immediate permanent price discount on those firms. Large firms do not engage in this behavior, consistent with its share prices being immune from fake news and the cost of information low enough in large firms that prices remain efficient.

Further research seeks to understand why firms may or may not engage in price manipulation, how these social media platforms can amplify or hinder the price discovery process, and potentially provide a way to measure the cost of information across firms.

References

Grossman, S. J. and J. E. Stiglitz (1980). On the impossibility of informationally efficient markets. *The American economic review 70*(3), 393–408.

Newman, M. L., J. W. Pennebaker, D. S. Berry, and J. M. Richards (2003). Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin 29*(5), 665–675.

Pennebaker, J., R. Booth, R. Boyd, and M. Francis (2015). Linguistic inquiry and word count: Liwc 2015 operator's manual. retrieved april 28, 2016.

Pennebaker, J. W. (2011). The secret life of pronouns: what our words say about us. Chapter 6. New York: Bloomsbury Press.

APPENDIX

Appendix A: An example of a promotional article about Galena Biopharma Inc.

Appendix B: 8-K form documenting the settlement between the SEC, Galena, and Mr. Ahn.

Figure 1. Example of a pump-and-dump scheme

This figure depicts the stock price of Galena Biopharma Inc. from 2012 - 2015, as well as occurrences of fake articles being published on Seeking Alpha and instances of trading by insiders at Galena.

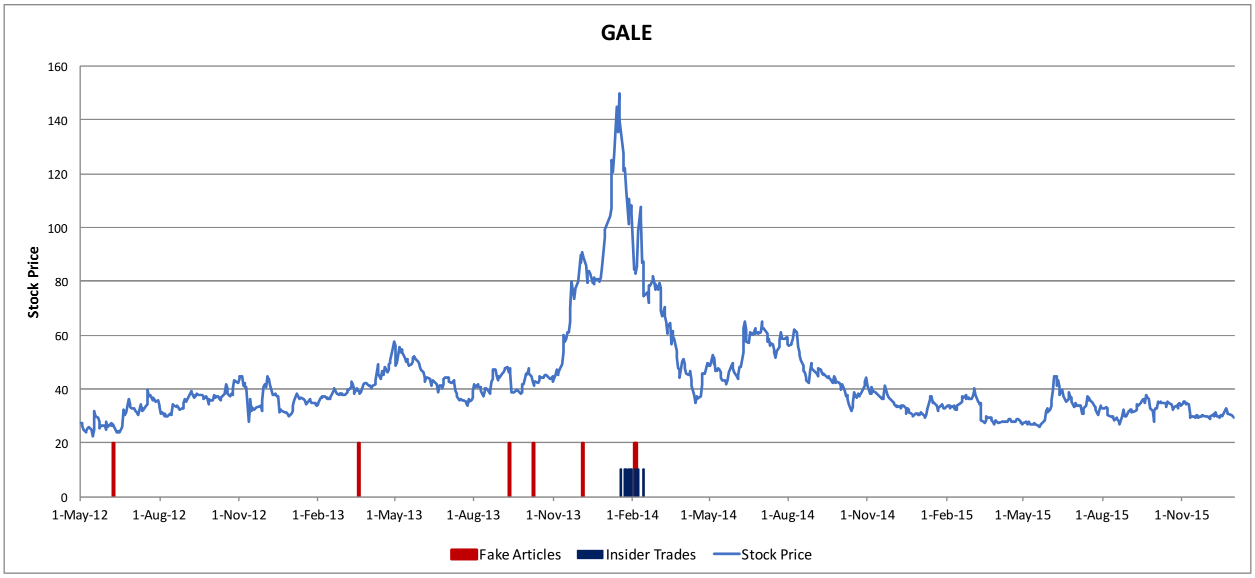
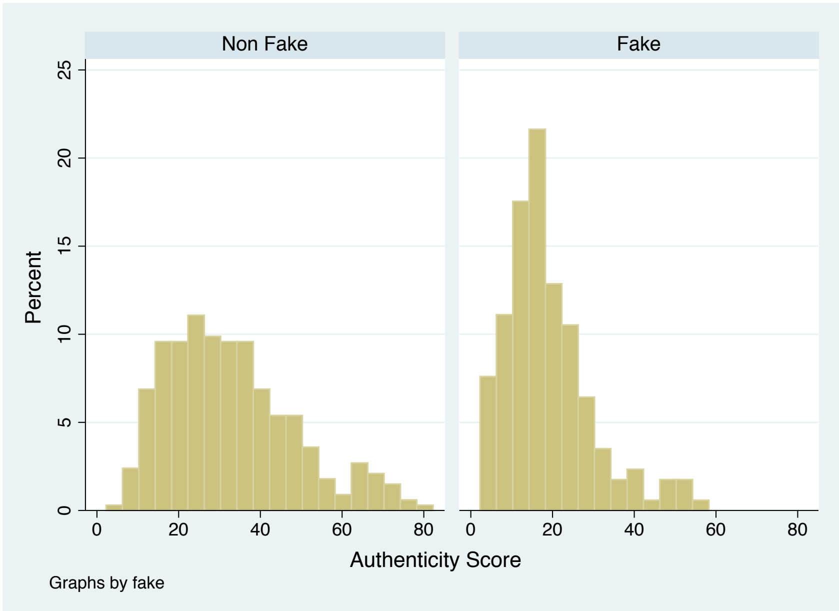


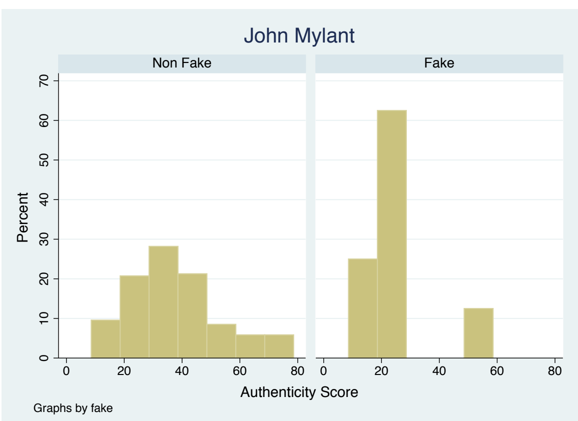
Figure 2. Authenticity Scores

This figure depicts the distribution of authenticity scores for fake and non-fake articles. In Panel A, we plot authenticity scores for all the articles in our validation sample of 171 fake and 334 non-fake articles. In Panel B, we plot authenticity scores for two authors in our validation sample with the most articles

Panel A:



Panel B:



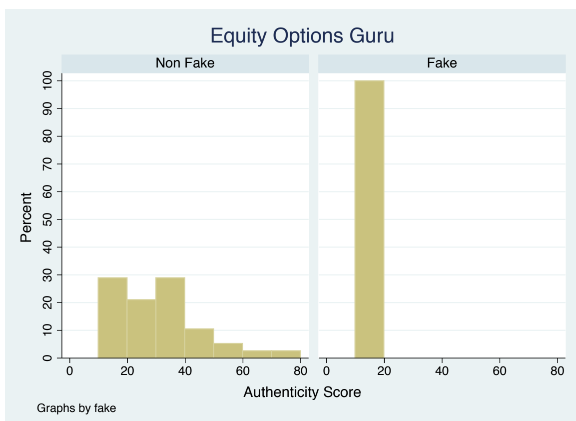


Figure 3. Authenticity score and the probability of being fake

This figure depicts the relationship between LIWC authenticity scores (S) and the conditional probability of being fake ($Prob(F|S)$).

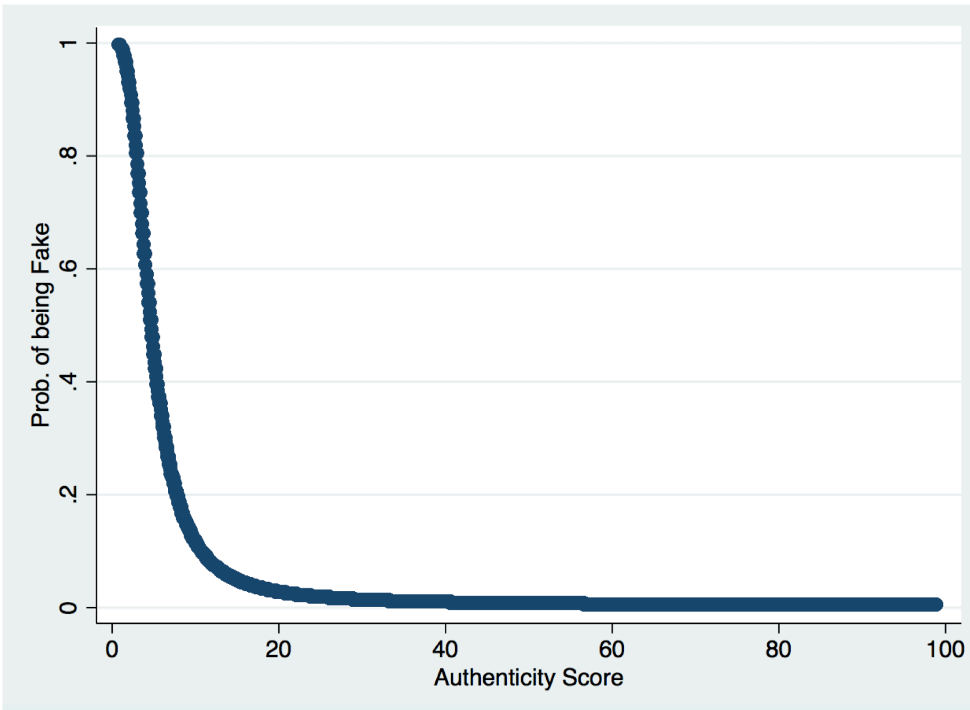


Figure 4. Abnormal Returns for For-sure Fake Articles

The figure depicts the progression of cumulative abnormal returns (measured as equal-weighted 4-factor residuals) for for-sure fake articles provided to us by Rick Pearson and that were subpoenaed by the SEC. The cumulative returns are measured starting with the day after the article was published until the 251 trading days after the article was published. Small firms are defined as firms in the bottom 10th percentile of NYSE firms, medium firms are defined as firms in the 20th-90th percentile of NYSE firms.

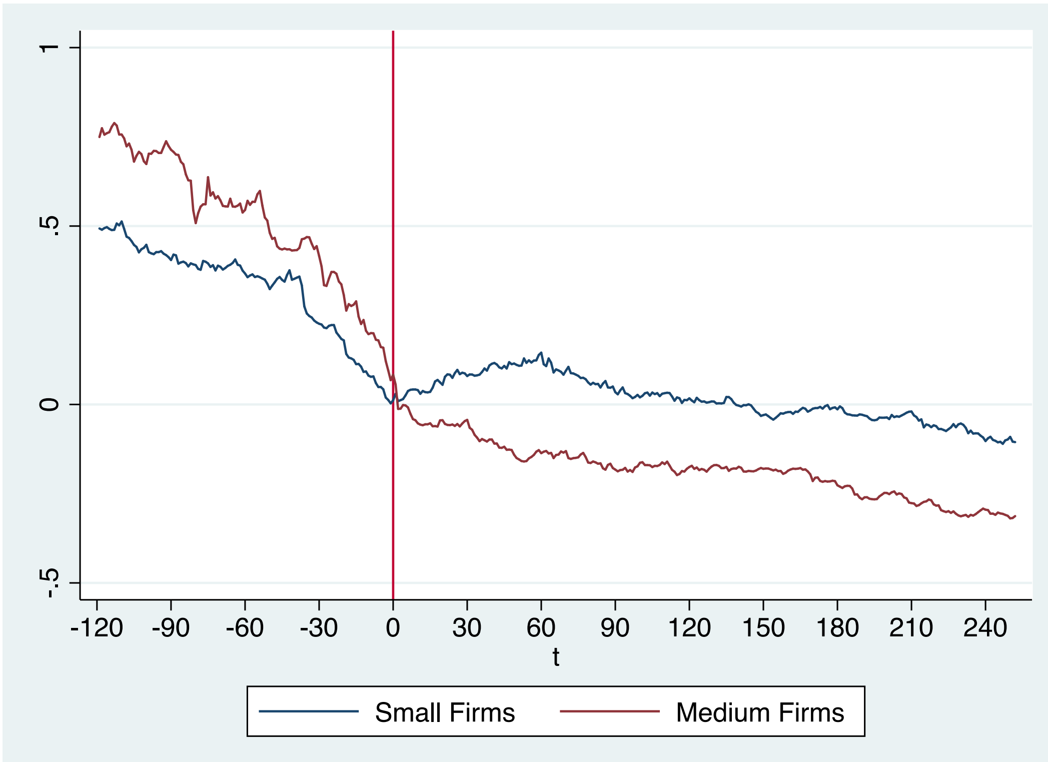


Figure 5. Abnormal Returns for Fake vs. non-Fake Articles

The figure depicts the difference in cumulative abnormal returns (measured as equal-weighted 4-factor residuals) between days with fake articles and days with non-fake articles separately for small, mid-size, and large firms in our sample. We designate a given day t for company i to have a fake article, if the probability of being fake, associated with the average authenticity score for all articles about firm i on day t, is greater than 20%. Similarly, we designate a day t for company i as not having any fake articles, if the probability of being fake, associated with the average authenticity score for all articles about firm i on day t, is less than 1%. The cumulative returns are measured starting with the day after the article was published until the 251 trading days after the article was published. Small firms are defined as firms in the bottom 10th percentile of NYSE firms, mid-size firms are defined as firms in the 20th-90th percentile of NYSE firms, and large firms are defined as firms above the 90th percentile of NYSE firms.

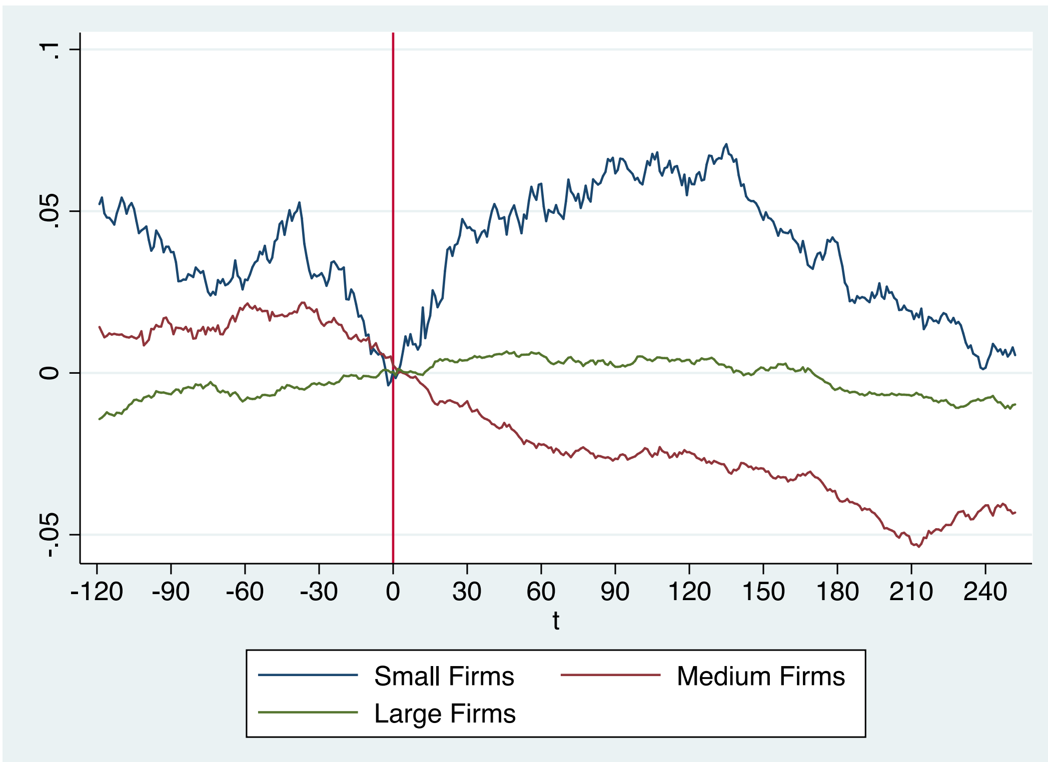


Table 1. LIWC and Fake Articles

This table presents the summary statistics for various LIWC textual measures for promotional articles that were shared with us by Rick Pearson, articles that were subpoenaed by the SEC that Seeking Alpha shared with us, and non-promotional articles that were written by the same authors. *For-sure Fake* articles are articles that have been shared with us by Rick Pearson and that were subpoenaed by the SEC. *Non-fake* articles, are articles that were written by the same authors, but about larger firms, that are unlikely to be promotional. We display the number of articles in each category as well as the mean of the *Authenticity* measure from LIWC. We also report the means of several other variables provided by LIWC to help better understand the authenticity score. In particular we display the means of the average of the *1st person singular* measure (examples: I, me, mine), *Insight* measure (examples: think, know), *Relativity* measure (examples: area, bend, exit), *Time* measure (examples: end, until, season), *Discrepancy* measure (examples: should, would), and the average number of words per sentence.

	For-sure Fake	Non-Fake
Number of articles	171	334
Authentic	19.08	32.77
	(10.66)	(15.73)
1st pers singular	0.42	0.76
	(0.43)	(0.67)
Words per sentence	57.54	65.23
	(33.30)	(31.73)
Insight	1.52	1.67
	(0.59)	(0.71)
Relativity	12.92	15.11
	(2.04)	(2.56)
Time	4.97	5.35
	(1.26)	(1.45)
Discrepancy	1.40	1.05
	(0.73)	(0.45)
Clout	58.25	52.31
	(9.78)	(8.63)

Table 2. Summary Statistics

This table presents the summary statistics for various LIWC textual measures and firm characteristics of the covered firms, for different types of articles on Seeking Alpha and Motley Fool. *For-sure Fake Articles* are articles that have been shared with us by Rick Pearson, or that were subpoenaed by the SEC and shared with us by Seeking Alpha. *Seeking Alpha Articles* and *Motley Fool Articles* are regular articles that we downloaded from Seeking Alpha and Motley Fool. Of those articles, *Fake* articles are articles whose probability of being fake was higher than 20%, *Non Fake* articles are articles with probability of being fake less than 1%, and the rest are classified as *Other*, which are not used in our main analysis.
In Panel A, we display the number of articles in each category as well as the mean of the *Authenticity* measure that we use to construct the probabilities of being fake. We also report the means of several other variables provided by LIWCvariables provided by LIWC to help better understand the authenticity score. In particular we display the means of the average of the *1st person singular* measure (examples: I, me, mine), *Insight* measure (examples: think, know), *Relativity* measure (examples: area, bend, exit), *Time* measure (examples: end, until, season), *Discrepancy* measure (examples: should, would), and the average number of words per sentence. In Panel B, we display the average probability of being fake, for each of the article categories. In Panel C, for the firms that are covered in the respective article groups, we provide the average fraction of retail investors, the average number of analysts covering the firm, and the average firm size (in Millions of dollars).

	Rick Pearson & SEC		Seeking Alpha			Motley Fool		
	For-sure Fake	Non Fake	Fake	Non Fake	Other	Fake	Non Fake	Other
Panel A: LIWC variables								
Number of articles	171	334	3,933	116,289	83,323	1,368	78,943	67,605
Authentic	19.09	32.79	5.44	50.71	22.51	5.71	46.75	21.96
1st pers singular	0.42	0.76	0.25	0.98	0.54	0.20	0.53	0.23
Words per sentence	57.55	65.23	23.89	21.76	22.18	31.23	19.28	19.39
Insight	1.52	1.67	1.43	1.75	1.63	1.62	2.08	1.84
Relativity	12.92	15.11	9.90	17.37	13.53	9.20	16.57	13.29
Time	4.97	5.35	3.40	6.34	4.68	3.34	6.54	5.23
Discrepancy	1.41	1.05	1.40	1.12	1.22	0.76	1.08	1.11
Panel B: Probability of being Fake								
Prob(Fake)	0.08	0.02	0.45	0.01	0.03	0.42	0.01	0.03
Panel C: Firm characteristics								
Percent of retail investors	76.66%	50.15%	42.32%	42.46%	44.96%	40.88%	36.78%	38.99%
Numer of Analysts	6.96	16.76	16.83	18.33	16.67	23.21	19.84	20.34
Firm Size ($Mil)	7.36	58.43	44.12	51.72	45.17	101.9733	70.58	80.4

Table 3. Fake Articles and Industries

This table presents the distribution of articles by Fama-French 12 industries, for different types of articles on Seeking Alpha and Motley Fool. *For-sure Fake Articles* are articles that have been shared with us by Rick Pearson, or that were subpoenaed by the SEC and shared with us by Seeking Alpha. *Seeking Alpha Articles* and *Motley Fool Articles* are regular articles that we downloaded from Seeking Alpha and Motley Fool. Of those articles, *Fake* articles are articles whose probability of being fake was higher than 20%, *Non Fake* articles are articles with probability of being fake less than 1%, and the rest are classified as *Other*, which are not used in our main analysis.

	Rick Pearson & SEC		Seeking Alpha			Motley Fool		
Industry	For-sure Fake	Non-Fake	Fake	Non-Fake	Others	Fake	Non-Fake	Others
Consumer NonDurables	-	2.45%	2.57%	5.19%	4.53%	5.67%	5.19%	5.19%
Consumer Durables	-	4.49%	3.13%	3.52%	3.37%	6.66%	5.04%	4.04%
Manufacturing	2.30%	12.65%	4.55%	7.26%	5.82%	8.05%	9.98%	8.09%
Energy	-	8.16%	4.9%	6.52%	6.17%	5.26%	5.66%	6.68%
Chemicals	1.15%	1.22%	1.46%	1.79%	1.78%	1.97%	2.44%	2.34%
Business Equipment	4.60%	27.35%	28.13%	23.66%	25.91%	26.87%	26.22%	25.39%
Telecom	-	2.86%	6.39%	4.77%	4.72%	4.35%	3.61%	3.87%
Utilities	-	-	1.11%	0.99%	1.46%	1.23%	1.66%	2.1%
Shops	-	2.86%	6.84%	12.19%	9.21%	13.72%	13.69%	11.62%
Healthcare	81.61%	17.14%	10.63%	5.38%	9.6%	7.81%	7.92%	10.4%
Finance	-	13.06%	22.2%	16.67%	16.42%	10.85%	6.49%	8.9%
Other	10.34%	7.76%	8.09%	12.06%	11.03%	7.56%	12.11%	11.38%

Table 4. Return Window Regressions – Unconditional

The table reports results from regressing 4-factor cumulative abnormal returns $Ret_{1,51}$, $Ret_{1,101}$, $Ret_{1,151}$, $Ret_{1,201}$, $Ret_{1,251}$ on a dummy variable for whether an article was fake. Small firms are defined as firms in the bottom 10th percentile of NYSE firms, mid-size firms are defined as firms in the 20th-90th percentile of NYSE firms, and Large firms are defined as the top 10th percentile of the NYSE firms.

	$Ret_{1,51}$	$Ret_{1,101}$	$Ret_{1,151}$	$Ret_{1,201}$	$Ret_{1,251}$
			Small Firms		
Fake	0.034	0.063***	0.055*	0.027	0.017
	(1.61)	(2.66)	(1.85)	(0.77)	(0.45)
Constant	-0.022***	-0.045***	-0.064***	-0.078***	-0.086***
	(-8.81)	(-12.62)	(-13.48)	(-13.20)	(-12.25)
Observations	11,622	11,622	11,622	11,622	11,622
R^2	0.000	0.000	0.000	0.000	0.000
			Medium Firms		
Fake	-0.017***	-0.020**	-0.028**	-0.045***	-0.038**
	(-3.04)	(-2.45)	(-2.50)	(-3.51)	(-2.50)
Constant	-0.006***	-0.012***	-0.017***	-0.025***	-0.031***
	(-7.76)	(-11.45)	(-12.21)	(-15.07)	(-16.05)
Observations	68,087	68,087	68,087	68,087	68,087
R^2	0.000	0.000	0.000	0.000	0.000
			Large Firms		
Fake	0.006*	0.004	0	-0.007	-0.011
	(1.71)	(0.90)	(0.06)	(-0.90)	(-1.33)
Constant	0.001**	0	-0.003***	-0.004***	-0.005***
	(2.26)	(-0.23)	(-2.77)	(-3.96)	(-4.02)
Observations	47,908	47,908	47,908	47,908	47,908
R^2	0.000	0.000	0.000	0.000	0.000

Table 5. Determinants of Article Impact on Volatility

The tables examines how firm-level idiosyncratic volatility (measured using 4-factor residuals) changes after the release of articles (days $[t+1, t+3]$) as a function of article and author characteristics. Number of comments measures the number of comments an article received, number of email recipients is the number of Seeking Alpha users that signed up to receive article about the firm, title length is the number of words in the article length, lagged author impact is the average variance observed for the author after the release of all prior articles, author tenure is the measure (in days) of the length of time from the first article observed for the author, author article count is the lagged number of articles published by the author, Fake news is a dummy equal to 1 if the probability of fake news is $> 20\%$, 1 if the probability of fake news is $< 1\%$, and missing otherwise.

Dependent variable	$[t+1, t+3]$ day idiosyncratic variance (ln)						
Num of comments (ln)	0.0128**		0.0177***	0.0053			
	(2.19)		(2.89)	(0.67)			
Num of email rec (ln)	0.0718***		0.0786***	0.0864***			
	(12.56)		(13.42)	(10.8)			
Title length (ln)	-0.0934***		-0.0766***	0.0013			0.0028
	(-5.61)		(-4.50)	(0.06)			(0.13)
Lagged auth impact		0.1450***	0.1226***	0.1191***	0.1483***		
		(31.23)	(23.14)	(16.7)	(23.65)		
Auth tenure (ln)		0.0011	-0.0075	-0.0055			
		(0.24)	(-1.29)	(-0.72)			
Auth article count (ln)		0.0109***	0.0274***	0.0172***			
		(3.88)	(6.15)	(2.96)			
Fake News				0.1033**	0.9097***	0.1312**	1.0922***
				(2.04)	(3.13)	(2.57)	(4.30)
Fake News ×					0.1034***		
Lag auth impact					(2.94)		
Auth lagged fake news count (ln)						1.7052***	
						(7.87)	
Fake News ×						-1.8952***	
Auth lagged fake news count (ln)						(-5.68)	
Fake News ×							-0.4885***
Title length (ln)							(-3.86)
Observations	193,190	332,673	184,149	108,006	189,686	189,686	115,359
R^2	0.23	0.19	0.23	0.24	0.2	0.2	0.24
Year, Firm FEs	X	X	X	X	X	X	X

Table 6. Fake News and Firm Announcements (Weekly Level)

In this table, we examine whether there are more likely to be fake news in the weeks around and contemporaneous with insider trading. At the weekly level, we regress a dummy variable for whether a firm had predominantly fake news in a given week (w = 0) on whether the firm was a net buyer or net a seller in the previous week (w-1), the contemporaneous week (w=0), and the following week (w=1), and a dummy variable for whether the firm issued a press release in weeks w-1, w=0, or w+1. *Net Buyer* (*Net Seller*) is an indicator for whether insiders bought more shares in dollar value than they sold in a given week (sold more shares than they bought). We define a dummy variable (*Fake Article*) for whether a firm had predominantly fake articles in a given week as 1 if the probability of being fake associated with the average authenticity score for articles written about the firm in the given week is great than 20%. *PR* is an indicator variable for whether the firm issues at least one press release in a given week. We perform our analysis separately for small, mid-size, and large firms. Small firms are defined as firms in the bottom 10th percentile of NYSE firms, mid-size firms are defined as firms in the 20th-90th percentile of NYSE firms, and large firms are defined as firms above the 90th percentile of NYSE firms. Standard errors are double-clustered at the year-month and firm level.

	Fake Article								
	Small Firms			Mid-size Firms			Large Firms		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Press Release (week-1)	0.0011**			0.0002			-0.0010*		
	(2.07)			(0.85)			(-1.75)		
8K filing (week-1)	0.0013**			0.0001			0.0003		
	(2.50)			(0.29)			(0.49)		
Press Release (week=0)		0.0026***			0.0020***			0.0008	
		(3.78)			(6.36)			(0.76)	
8K filing (week=0)		0.0018***			0.0017***			0.0002	
		(2.92)			(4.57)			(1.07)	
Press Release (week+1)			0.0002			0.0002			0.0001
			(0.36)			(0.71)			(0.18)
8K filing (week+1)			0.0014**			0.0004			0.0006
			(2.49)			(1.51)			(0.88)
Observations	137,560	137,998	137,719	406,508	407,379	406,593	86,956	87,104	86,946
R-squared	0.010	0.011	0.011	0.007	0.008	0.007	0.013	0.013	0.013
Year-month, Firm FEs	X	X	X	X	X	X	X	X	X

Table 7. Insider Trading and Fake News (Weekly Level)

In this table, we examine whether there are more likely to be fake news in the weeks around and contemporaneous with insider trading. At the weekly level, we regress a dummy variable for whether a firm had predominantly fake news in a given week (w = 0) on whether the firm was a net buyer or net a seller in the previous week (w-1), the contemporaneous week (w=0), and the following week (w=1), and a dummy variable for whether the firm issued a press release in weeks w-1, w=0, or w+1. *Net Buyer* (*Net Seller*) is an indicator for whether insiders bought more shares in dollar value than they sold in a given week (sold more shares than they bought). We define a dummy variable (*Fake Article*) for whether a firm had predominantly fake articles in a given week as 1 if the probability of being fake associated with the average authenticity score for articles written about the firm in the given week is great than 20%. *PR* is an indicator variable for whether the firm issues at least one press release in a given week. We perform our analysis separately for small, mid-size, and large firms. Small firms are defined as firms in the bottom 10th percentile of NYSE firms, mid-size firms are defined as firms in the 20th-90th percentile of NYSE firms, and large firms are defined as firms above the 90th percentile of NYSE firms. Standard errors are double-clustered at the year-month and firm level.

	Fake Article								
	Small Firms			Mid-size Firms			Large Firms		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Buyer (week-1)	0.0025*			0.0020*			-0.0062		
	(1.92)			(1.81)			(-1.41)		
Seller (week-1)	0.0008			0.0004			-0.0008		
	(0.72)			(0.90)			(-0.91)		
Buyer (week=0)		0.0051***			0.0040***			-0.005	
		(3.03)			(4.15)			(-0.62)	
Seller (week=0)		0.0017			0.0004			-0.0013	
		(1.56)			(1.05)			(-1.50)	
Buyer (week+1)			0.0058***			0.0022*			-0.0013
			(3.27)			(1.94)			(-0.55)
Seller (week+1)			0.0010			0.0006			0.0005
			(1.04)			(1.64)			(0.53)
Observations	137,593	137,998	137,721	406,575	407,379	406,595	86,959	87,104	86,946
R-squared	0.010	0.011	0.011	0.007	0.007	0.007	0.013	0.013	0.013
Year-month, Firm FEs	X	X	X	X	X	X	X	X	X