First to "Read" the News: News Analytics and Algorithmic Trading

Bastian von Beschwitz*	Donald B. Keim**	Massimo Massa***
Federal Reserve Board	Wharton School	INSEAD

May 5, 2017

Abstract

We investigate whether providers of news analytics affect the stock market. We exploit a unique identification strategy based on inaccurate news analytics that were released to the market. We document a causal effect of news analytics on the market, irrespective of the informational content of the news. Coverage in news analytics speeds up the market reaction in terms of stock price response and trading volume. The market reacts to inaccurate news analytics but these price distortions are generally small and corrected quickly. Furthermore, we document that traders learn dynamically about the precision of news analytics.

JEL classification: G10, G12, G14

Keywords: Stock Price Reaction, News Analytics, Information, Algorithmic Trading.

* Bastian von Beschwitz, Federal Reserve Board, International Finance Division, 20th Street and Constitution Avenue N.W., Washington, D.C. 20551, tel. +1 202 475 6330, e-mail: bastian.vonbeschwitz@insead.edu (corresponding author).

** Donald B. Keim, Wharton School, University of Pennsylvania, Philadelphia, PA 19104; keim@wharton.upenn.edu

*** Massimo Massa, INSEAD, Finance Department, Bd de Constance, 77305 Fontainebleau Cedex, France, tel. + 33-(0)160-724-481, email: massimo.massa@insead.edu

An earlier version of this paper was titled "Media-Driven High Frequency Trading : Evidence from News Analytics". We are grateful to RavenPack for providing their data, and Malcolm Bain in particular for his expertise on different RavenPack releases. Thanks also to the technical personnel at WRDS, especially Mark Keintz, for making the construction of the intraday-market indexes possible. We thank Joseph Engelberg, Nicholas Hirschey, Todd Gormley, Markus Leippold, Joel Peress, Ryan Riordan, Paul Tetlock, Sarah Zhang and conference participants at the NBER Microstructure Meeting, European Winter Finance Summit, FIRS, and DGF for valuable comments. We acknowledge the financial support of the Wharton-INSEAD Center for Global Research and Education. All remaining errors are our responsibility. The views in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

1. Introduction

The recent decade has witnessed three major phenomena. The first has been the rise of algorithmic and high frequency trading (HFT). HFT now accounts for nearly 50% of trading volume (Gerig, (2015)) and the race for higher execution speeds has driven the latency of the fastest traders down to the nanosecond level (Gai, Yao and Ye, (2013)). The second is the rise of algorithmic processing of news releases ("news analytics"). The third is an increase in the number of "flash crashes" – i.e., sudden strong deviations of prices from fundamentals that are quickly reversed (e.g. Brogaard et al. (2015), Golub, Keane and Poon (2012)). This raises the question of whether there is a link between these phenomena. In particular, given that one of the main benefits of financial markets is the aggregation of information and assimilation into prices, which role does the information generated by algorithmic news processing play in a world dominated by HFT and algorithmic trading?

The question is tricky as HFTs trade mainly in reaction to quotes and prices – i.e., they react to information that is already inside the market system. In contrast, news analytics allow its users to react faster to events that are not yet reflected in asset prices. For example, RavenPack, a major provider of news analytics whose data we use in this study, uses computer algorithms to determine for each article in the Dow Jones Newswire its relevance to each company mentioned in it, and whether it is positive or negative. This processed content is then electronically delivered to RavenPack's subscribers within a third of a second. While this is slow compared to the speed with which HFTs can react to price movements, news analytics companies such as RavenPack provide the fastest way to react to information that is not yet reflected in asset prices.

In this paper, we study how news analytics affect the way financial markets incorporate public information. Two features make news analytics particularly interesting. On the upside, news

analytics likely increase the speed with which markets incorporate information and thus increase their efficiency. On the downside, an algorithm "reading" news inaccurately can lead to unintended consequences as trading programs automatically initiate trades on an incorrect assessment of the news content. For example, in April 2013, an incorrect twitter feed about a White House explosion caused a mini flash crash. Some quickly blamed algorithmic trading for the reaction, while others argued that human traders were mainly responsible.[1] In any case, news-reading algorithms might be more likely to misinterpret news than humans. This raises the question of whether or not news analytics contribute to mini flash crashes.

The question of how news analytics affect the stock market is important, especially for policy considerations; but their effect is not easy to isolate because the *response to news analytics normally cannot be distinguished from the reaction to the news itself*. We are able to address this distinction by exploiting a unique identification strategy based on inaccuracies in news analytics that are revealed by comparing older and newer versions of RavenPack. We use the back-filled analytics of increasingly more sophisticated versions of RavenPack to identify inaccuracies in the old version that was released to the market. Finding evidence that markets react to such inaccuracies would suggest a causal impact of RavenPack on the stock market.

We test the following hypotheses. First, we ask whether inaccuracies in news analytics lead to price distortions analogous to "mini" flash crashes – i.e., whether they trigger price reactions that are subsequently reversed (*Hypothesis 1*). Second, we ask whether news analytics increase the speed with which traders react to public signals and thus the speed at which the market incorporates information (*Hypothesis 2*).

[1] See for example "The Trading Robots Really Are Reading Twitter"- http://finance.yahoo.com/news/trading-robots-really-reading-twitter-124443495.html and "#hashcrash: The anatomy of an investment panic" http://goinfront.com/blog/article/497

Finally, we ask whether high frequency traders dynamically learn about the signal precision of news analytics. If they do, we expect a lower price reaction to news analytics for stocks in which news analytics have been less informative in the past (*Hypothesis 3*). This question has interesting implications for how the market reacts to inaccurate news analytics, which are especially uninformative. If traders learn dynamically about the quality of RavenPack's signals, then the market reaction to them will be reduced following prior inaccurate Ravenpack signals about the stock. Thus, inaccurate news analytics can change the way in which the market reacts to future news analytics.

To identify inaccuracies in news analytics, we focus on differences in RavenPack's "relevance score", which measures the importance of an article for a certain company. The relevance score is very important: highly relevant articles that are positive (negative) are followed by positive (negative) stock returns, while there is almost no reaction to articles with a low relevance score. Differences in relevance scores between the old and new RavenPack versions are due to improvements in the algorithm when identifying companies in the article and determining the article's relevance to the company.

We use these differences in relevance scores to define three categories of articles: High-relevance articles Released as High-relevance articles (HRH) are articles that were correctly released to the market; Low-relevance articles Released as High-relevance articles (LRH) are false positives, i.e. articles that are wrongly attributed to a company; and High-relevance articles Released as Low-relevance articles (HRL) are false negatives, i.e. articles that the old version of RavenPack failed to attribute to the correct company.

To study Hypothesis 1, we focus on LRH articles. We find that the market indeed reacts to such false positives, but the effect does not persist. The market initially overreacts to the incorrect

information, realizes the inaccuracy, and quickly corrects after 30 seconds, thus confirming Hypothesis 1 and the causal effect of RavenPack on stock prices. While the magnitude of the reaction is too small to classify as a flash crash, larger magnitudes may be expected if news analytics become used even more widely.

To test the remaining hypotheses, we focus on the comparison between HRH and HRL articles. These two article types are of similar relevance according to the most recent version of RavenPack, but only HRH articles were correctly classified, and released to the market, as highly relevant. In contrast, HRL articles were incorrectly released as not relevant in the old version and thus for these articles there should be no causal effect of RavenPack on stock prices. Comparing the market response to HRH and HRL articles provides another way to assess the causal effect of RavenPack.

We find that the market reacts differently to HRH and HRL articles. The share of stock price reaction concentrated in the first 5 seconds after an article, compared to the total reaction over 120 seconds, is significantly greater for articles released as highly relevant (HRH) than for those highly-relevant articles originally released as having low relevance (HRL). This difference in speed of the stock price response is 1.3 percentage points or *10% relative to the mean*.

Not only does the market react faster, but it also reacts in the sentiment direction indicated by RavenPack. Indeed, the sentiment direction of an article as determined by RavenPack predicts the stock price reaction to the article better when RavenPack consistently identifies the article as having high relevance (HRH) than when the old technology mislabelled it as having low relevance (HRL). This implies that traders use RavenPack to trade in the direction of the sentiment indicator provided by the news analytics.

In addition to the faster stock price response, we also document an increase in the share of trade volume concentrated in the first 5 seconds compared to the two minutes after an article. This increase is consistent with the theoretical prediction that investors with a speed advantage trade aggressively on signals that they can exploit before other traders (e.g., Foucault, Hombert, and Rosu (2016)). Taken together, these findings confirm *Hypothesis 2*.

Finally, we document that high frequency traders dynamically learn about the signal precision of RavenPack. More specifically, the causal effect of RavenPack on 5-second announcement returns is stronger if RavenPack has been more informative in the past – i.e., if RavenPack's sentiment scores accurately predicted 2-minute announcement returns in the past for that industry. A one standard deviation increase in informativeness almost doubles the causal effect of RavenPack's sentiment score on 5-second returns. These findings suggest that algorithmic traders learn dynamically about the precision of RavenPack, and that they rely more heavily on RavenPack's sentiment scores if these scores have been informative in the past. Such learning could be programmed into their algorithms (machine learning) or can come from manually updating their algorithms over time. This finding confirms *Hypothesis 3*.

A series of robustness checks confirm our results. First, one potential concern is that our results are driven by the fact that HRH articles are systematically different from HRL articles. We address this issue in two ways. First, we show that HRH and HRL articles are similar in terms of long-run stock price reactions and several other characteristics. Second, we use the fact that RavenPack has back-filled the data of all versions to February 2004 to conduct placebo tests during the time before RavenPack went live. If our results are driven by actual differences between the two article types, rather than a causal impact of RavenPack, then we should find significant differences in price reactions before RavenPack went live. However, for all tests we report

insignificant differences in price reactions (between HRH and HRL articles) before RavenPack went live. Moreover, the stock price reactions to HRH and HRL articles start to diverge precisely when RavenPack went live, and the resulting increase in the difference between HRH and HRL articles is significant. All of this suggests a direct causal impact of RavenPack on the stock market.

Overall, our tests show that news analytics have a significant impact on the market in terms of returns and trading volume in a manner predicted by several models. This effect goes beyond the underlying influence of the news itself. While our study can only detect the effect of RavenPack, there are other providers of news analytics and traders may conduct algorithmic news processing in house. Thus, the total effect of algorithmic news processing is likely much larger than the effect of RavenPack measured in this paper.

Importantly, our results have normative implications with respect to discussions about the regulation of high-speed sources of information and the effects of algorithmic trading.[2] Our results suggest that news analytics are generally favourable for market conditions. They allow the market to incorporate information more quickly and thereby improve efficiency.[3] While inaccuracies in news analytics can lead to stock price reactions that are unrelated to fundamental news, these reactions are generally small and quickly reversed. However, one may be concerned about larger reactions if the usage of news analytics would further increase. In this case, erroneous news analytics may indeed lead to "mini" flash crashes.

[2] "FBI joins SEC in computer trading probe", Financial Times March 5, 2013.
[3] It is questionable, whether the increased efficiency yields sufficient welfare gains to justify the investments in fast trading technology. For a theoretical paper on the welfare effects of high frequency trading see Biais, Foucault and Moinas (2015). Also, improved price efficiency can lead to lower incentives to gather private information (Weller, 2015).

Our results contribute to three major strands of literature. First, we contribute to the growing empirical literature on algorithmic and high frequency trading.[4] Several papers show that high frequency traders in general improve price efficiency (e.g. Brogaard, Hendershott and Riordan (2014), Chaboud et al. (2013), Boehmer, Fong and Wu (2015)). In contrast to these studies, we are able to examine one specific channel of their informational advantage and provide evidence of an increased speed of price adjustment to that information.

Second, we contribute to the literature on news analytics. Prior papers in this literature study the correlation between the market and news analytics without passing judgment on whether there is a causal impact of news analytics on the market (e.g. Riordan, Storkenmaier, Wagener, and Zhang (2013), Gross-Klugmann and Hautsch (2011), Sinha (2012), Heston and Sinha (2016), Zhang (2013)). In contrast, our paper is the first to show the *causal* impact of news analytics on stock markets.

Third, our results are consistent with recent models of high frequency trading in which some traders have an informational advantage. For example, Foucault, Hombert, and Rosu (2016) model a situation in which a speculator receives information one period ahead of the market maker in a set-up similar to Kyle (1985); in Martinez and Rosu (2013) some agents have a short-lived informational advantage; and in Dugast and Foucault (2017), speculators face a trade-off between processing a signal faster or more accurately. Faster traders in these models make markets more informationally efficient, but also more unstable. We find support for both effects.

[4] Examples of this literature include Hendershott and Riordan (2013), Hendershott, Jones, and Menkveld (2011), Baron, Brogaard and Kirilenko (2014), Menkveld (2013), Jovanovic and Menkveld (2010), Riordan and Storkenmaier (2012), Boehmer, Fong, and Wu (2015), Hasbrouck and Saar (2013), Benos and Sagade (2012), Clark-Joseph (2013), Hirschey (2013), Brogaard et al. (2014), Chordia, Green, and Kottimukkalur (2015). A survey of this literature is provided by Jones (2013).

Fourth, we contribute to the literature on the causal effect of media on the stock market.[5] Methods to address the endogeneity of media coverage include exogenous scheduling of journalists (Dougal, Engelberg, Garcia, and Parsons, 2011), local media coverage and its delay due to extreme weather (Engelberg and Parsons, 2011) and newspaper strikes (Peress, 2014). We add to this literature in two ways. First, we study news analytics – i.e. meta information, rather than news articles themselves. Second, we study the effect of such news analytics on algorithmic traders rather than private investors. This focus increases the policy relevance of our findings in a regulatory environment that is increasingly focused on news analytics.

2. Test design, identification strategy, and data sources

In this section we first describe the RavenPack news analytics data and how it is used in our identification strategy and tests. After briefly describing our stock market data, we then present summary statistics for the variables used in our tests. Variable definitions are in Appendix 1.

2.1 RavenPack

RavenPack provides real-time news analytics based on the Dow Jones (DJ) Newswire. This service analyzes all the articles on the DJ Newswire with a computer algorithm and delivers article-level relevance and sentiment metrics to its users. It determines which companies are mentioned in the article, how relevant the article is to the company and reports different sentiment indicators about whether the article is good or bad news for the company. The latency – i.e. the time from the release of the DJ Newswire to the release of the RavenPack metrics – is approximately 300 milliseconds. RavenPack claims it has the "timeliest company sentiment indicators in the

[5] There is a wider literature on media and stock markets including for example Chan (2003), Tetlock (2007, 2011), Fang and Peress (2009), Griffin, Hirschey, and Kelly (2011), Boudoukh et al. (2012), Loughran and McDonald (2013), Garcia (2013), Ferguson et al. (2015)). For a review on textual analysis in finance see Kearney and Liu (2014).

marketplace."[6] As such, RavenPack is ideally suited for the use of traders engaging in algorithmic news trading. These traders could be co-located HFTs, but more likely they are hedge funds engaged in news trading.[7]

2.1.1 Ravenpack – definition of variables

We extract from RavenPack the following variables. *Article Category* is a variable determining the topic of the article and the role played by the company in the article. For example, *Article Category* might be "acquisition – completed – acquirer" for a company announcing the completion of an acquisition of another company or "rating – change – negative – rater" for a rating company that just downgraded another company. The identification of the news topic is based on a purely algorithmic approach, and a large percentage of articles cannot be classified in this way. *Article Category Identified* is a dummy variable equal to 1 if *Article Category* is identified by RavenPack, and zero otherwise.

There are two major sentiment scores in RavenPack. The *Composite Sentiment Score* (CSS) is based on several individual RavenPack sentiment measures. It takes a value ranging from 100 (positive) to 0 (negative), where 50 is a neutral article. It is available for each article. The *Event Sentiment Score* (ESS) is coded in the same way as CSS, but available only if the category of the article can be identified. We aggregate these two scores into a single sentiment variable called *Sentiment Direction*, which is primarily based on ESS and uses CSS only if ESS is either missing or equal to 50 (neutral).

[6] "RavenPack Enables Trading Programs with Sentiment on 10,000 Global Equities," RavenPack press release from May 28, 2009.

[7] Confidential discussions with RavenPack managers provided us with a very consistent overview of market penetration, suggesting that major institutional investors are in fact users of this service.

Relevance is an index provided by RavenPack that indicates the relevance of an article to the company. *Relevance* takes values ranging from 0 (least relevant) to 100 (most relevant). If the type of the article can be identified and the company plays an important role in the main context of the story – e.g. is an acquirer or announces a buyback – then *Relevance* is 100. If the company is mentioned in the title, but the type of article cannot be identified, then *Relevance* ranges between 90 and 100. If the company is mentioned, but plays an unimportant role, then it gets a low *Relevance* score – e.g., a bank advising an acquisition might get a score of 20. We would not expect such articles to affect the bank's or news agency's stock prices very much.

In line with this, RavenPack recommends "filtering for *Relevance* greater than or equal to 90 as this helps reduce noise in the signal". To examine this claim, Figure 1 plots the market reaction to news as a function of *Relevance*. We plot the cumulative returns relative to news events from April 1, 2009 to September 10, 2012. We multiply returns by the article's sentiment direction. The articles with *Relevance* greater than 90 do indeed have an important effect on stock prices, but there is no reaction to articles with *Relevance* below 90. Thus, we will refer to articles with *Relevance* below 90 as low relevance. This analysis suggests that RavenPack is good at identifying both relevance and sentiment of an article.

That the reaction to high relevance articles starts about 60 seconds before the article suggests that *some* of the news events are covered in other news sources before they are covered in the DJ Newswire (used by RavenPack). Cases where the DJ Newswire is not the first to report an event should only work against us by making it more difficult to find a causal impact of RavenPack. We have no reason to believe that this issue should bias the results because it should be unrelated to whether RavenPack makes a mistake interpreting the article. While some trades in the 5 seconds after a RavenPack article may be due to human traders reacting to earlier coverage of the news

elsewhere, this trading should affect both HRL and HRH articles. Thus the *additional* trading following HRH articles (relative to HRL articles) should only be due to algorithmic traders reacting to the coverage in RavenPack itself.

2.1.2 Ravenpack – test design using different product versions

RavenPack released its first version (v. 1.0) to the market on April 1, 2009,[8] [9] and a revised version of the service (v. 2.0) with additional features on June 6, 2011. The most recent version we use (v. 3.0) was released on September 10, 2012. RavenPack has provided us with data from each of the release-specific algorithms, each having been back-filled to February 2004. RavenPack does not continuously update its algorithm, so as not to distort its customers' trading strategies which might be based on specific variable definitions. Rather, RavenPack rolls out any changes to its algorithm when releasing a new version, meaning that stock-specific metrics from the three releases can sometimes differ.[10] These differences are often related to the way companies are identified in an article and how the relevance of an article to a company is determined.[11] Thus, there are articles that might be associated with a particular company in one RavenPack release, but not in another. Such differences in the relevance of articles to companies in different versions provide the basis for our tests. Assuming the most recent version of RavenPack (v. 3.0, hereafter New RavenPack) is the most accurate, we can identify inaccuracies in RavenPack 1.0 and RavenPack 2.0 (hereafter

[8] Even though the official release date of the RavenPack service was May 2009, some customers had access to the service as early as from April 1, 2009. Thus, we refer to April 1, 2009 as the introduction of RavenPack. Before April 2009 RavenPack had a pre-existing service that also released sentiment information on the Dow Jones News Wire. However, this service was meant for longer term news analysis, such as charting sentiment over several days. The prior service was not provided timely enough to be used at high frequency.

[9] RavenPack 1.0 was actually released on Sept 6, 2010. A predecessor to v.1.0, that was similar to v.1.0, is the version that was released on April 1, 2009. This predecessor version was not made available to us, but RavenPack confirmed that it was very similar to RavenPack 1.0.

[10] Because the algorithm is proprietary, we do not know exactly what changes RavenPack implemented but some examples of articles where the two versions disagree are provided in the Internet Appendix.

[11] In addition, the number of companies covered by RavenPack has also increased between releases. There are 156 companies (3%), which are only covered in New RavenPack. We ensure by using company fixed effects that this difference in coverage is not driving our results.

Old RavenPack) that were released to the market. If the market reacts to these inaccuracies, it is an indication of a causal effect of RavenPack on the stock market.

Our analysis can be thought of as assuming two types of traders: Algorithmic traders that subscribe to RavenPack and human traders that manually read the article to determine its content. Further, we assume that human traders can more precisely derive the relevant signal from the article, while algorithmic traders have an advantage in terms of speed (a setting modelled by Dugast and Foucault (2017)). This means that RavenPack allows its subscribers to trade faster on a possibly less precise signal. In the short run, when only algorithmic traders can react to news, RavenPack will have the largest impact; while in the long run human traders determine the price reaction because their signal is more precise.

In the empirical implementation we choose specific time intervals to constitute the short and long run. We define the short run to be 5 seconds, because this is long enough to capture the full reaction of algorithmic traders (and accommodates slower algorithmic traders that are not co-located and not trading within milliseconds), but is too short for a human trader to read an article, process it and make a trading decision based on it. We choose two minutes as the long run because this permits enough time to read an article and trade on it, whereas longer time windows will be more affected by noise. In the Internet Appendix, we provide robustness checks in which we use both 1 and 10 seconds for the short run and 5 minutes for the long run.

We define the following article types that we also list in Panel A of Table 1. *High relevance article Released as High relevance article* (HRH) is defined as an article classified as relevant in both Old and New RavenPack. We predict that such a correctly released article creates a fast and persistent market reaction. *High relevance article Released as Low relevance article* (HRL) is defined as an article with high relevance in New RavenPack, but incorrectly assigned low

relevance in Old RavenPack. Low relevance means either the article was not assigned to the company or the relevance score was below 90. We expect an HRL article to have a similar long-run market reaction as an HRH article because they are of similar relevance according to New RavenPack. However, we would expect a slower market reaction to an HRL article as it was not released originally as relevant. *Low relevance article Released as High relevance article* (LRH) is an article that was incorrectly released to investors as having high relevance but has low relevance in New RavenPack. For these articles we expect an initial overreaction of algorithmic traders, which might later be reversed by human traders. Examples of all three article groups are provided in the Internet Appendix. A fourth article category is *Low relevance articles Released as Low relevance articles* (LRL); these articles have a relevance score below 90 in both versions.[12] We do not expect much market reaction to LRL articles.

These predictions allow for two possible empirical set-ups. First, we could study overreaction to *false positives* by comparing the differential market reaction to low relevance articles released as either high or low relevance, i.e. comparing LRH and LRL. Second, we could study underreaction to *false negatives* by comparing the differential market reaction to high relevance articles released as either high or low relevance, i.e. comparing HRH and HRL.

In both cases, we would assume that Old RavenPack contains no information on the relevance of the article over and above that contained in New RavenPack. This is a fairly strong assumption. Fortunately, this assumption is testable. Because we have data from 2004 and RavenPack went "live" in 2009, we can examine the market impact to the different types of articles during the time period when RavenPack could not have had any causal market impact, because it was not yet

12 LRL articles also include articles that have a relevance score below 90 in either Old RavenPack or New RavenPack and are not assigned to the company in the other version.

"live". To do this, we regress absolute return and turnover in the two minutes after the article's release on dummy variables equal to 1 for HRH, HRL and LRH (with LRL being the omitted category). To control for firm- and time-specific effects, we include firm, date and hour-of-the-day fixed effects.

The results are presented in Panel B of Table 1. The coefficient on LRH is large and significantly positive for both absolute returns and turnover, suggesting LRH articles are significantly more important than LRL articles. This implies that a test of overreaction comparing LRH and LRL articles is not possible because the two article types are fundamentally different. Thus, instead of comparing LRH and LRL articles, we rely on graphical evidence to compare the reaction to LRH articles before and after RavenPack went "live".

When we test the significance of the difference between the coefficients on HRH and HRL, however, we find the difference is small and insignificant for both turnover and absolute returns. Therefore, most of our tests are based on the comparison between these two groups. To ensure further that differences in importance between HRH and HRL are not driving our results, we control for many article characteristics and study the speed of market reaction, i.e. the size of the short run reaction relative to the long-run reaction. We also conduct placebo checks for all our tests showing that our results are not driven by differences between HRH and HRL articles, but by the causal effect of RavenPack.

2.2 Stock market data

We use intraday quotes and trade data from TAQ.[13] We use the TAQ National Best Bid and Offer (NBBO) file provided by WRDS for quotes. As a first step, we aggregate trading volume at the

[13] We use the usual filters of excluding all trades with zero size, negative prices, correction code different from 0 and bid ask quotes where the bid is above the asked.

frequency of one second, and compute second-by-second returns based on end-of-second bid-ask midpoints. We use bid-ask midpoints rather than trading prices to avoid bid-ask bounce effects. Even after this aggregation, the data for *all* stocks in our 8-year sample is far too large to be used in a second-by-second panel set-up. But we are interested only in the market reaction around specific company news events, so we limit our analysis to a few minutes around only these events. This simplification allows us to study all US common stocks over the full 8-year sample period.

To control for the overall market movements taking place during this period, we compute a second-by-second intraday market index from the total TAQ universe. We compute second-by-second returns, turnover and value-weighted volatility for the market index. We also compute returns for industry-specific indices for the 12 Fama French industries. The details of the index construction are explained in Internet Appendix 1. To control for stock-specific information, we use the CRSP daily stock file and compute the prior month's return, volatility, turnover, Amihud (2002) illiquidity measure, and market capitalization.

We employ the following filters: To be included in our sample, a stock must be covered in CRSP and TAQ, must have SHRCD 10 or 11, must have a beginning of the day stock price of at least $1 and must have a beginning of the day percentage bid-ask spread of less than 10%. We exclude articles that occur outside trading hours or in the first or last 20 minutes of trading in the day. To avoid distortions from overlapping windows around articles, we exclude stale news defined as articles for which the company had an article in the prior 15 minutes. We also exclude four companies that appear in articles mainly as information providers: McGraw-Hill, NASDAQ, CME and Moody's. Because we need an initial bid-ask midpoint to compute a first return and because we want to avoid a stock's turnover influencing the stock price we measure, we use

seconds *t−480* to *t−1* as a burn-in period. Only articles for which the stock has a quote in those 8 minutes before the article are included in our analysis.

2.3 Summary statistics and comparison between HRH and HRL

The final sample consists of 321,912 article-firm combinations, starting with the release of RavenPack 1.0, over the period April 1, 2009 to September 10, 2012. In Panel A of Table 2, we report descriptive statistics for all our variables for the combined sample of articles classified as HRH and articles classified as HRL.

As alluded to previously, a concern is that the information content of HRH and HRL articles might be different. Therefore, we compare their difference in terms of observable variables in Panel B. For this purpose, we regress each article characteristic on a dummy variable, D(HRH) equal to 1 if the article is HRH, and Relevance, Category, Hour and Date Fixed Effects. We report the coefficient of D(HRH) as well as a t-statistic clustered at the firm level. There is no statistical significant difference between the HRH and HRL articles in terms of firm size, sentiment scores, time since the last article, turnover and illiquidity. Most importantly, we find no evidence that HRH are more important than HRL articles: the absolute returns both over the 2 minutes following an article and on the full trading day of the article are actually (insignificantly) lower for HRH articles. The only evidence of a significant difference is that stocks that are the subjects of HRH articles have a slightly lower return (0.03%) and volatility (1.5%) in the prior month than those associated with HRL articles, and that HRH articles cover fewer firms per article. However, these differences are small in economic terms (0.05, 0.09 and 0.22 standard deviations) and we account for these differences with control variables in all our subsequent tests. In addition, we show in the Internet Appendix that our results are robust to including fixed effects for the number of firms mentioned in the article. That the characteristics of HRH and HRL articles are similar alleviates

worries that our results are driven by differences in the article types. In addition, we run placebo tests to confirm that unobservable differences are not driving our results.

3. Results

This section reports the empirical results of our paper. Each subsection is dedicated to one of the hypotheses outlined in the introduction.

3.1 News analytics and temporary price distortions

Here we examine *Hypothesis 1*, whether inaccuracies in news analytics lead to price distortions analogous to "mini" flash crashes, i.e. to an overreaction in stock price that is afterwards reversed. As explained in Section 1.1.2, we expect the market to overreact to LRH articles, i.e. articles that New RavenPack identifies as having low relevance, but that were incorrectly released as having high relevance in Old RavenPack.

We first consider graphical evidence. In Figure 2, we compare the market reaction of articles consistently released as relevant (HRH) with those originally released as relevant, but assigned low relevance in New RavenPack (LRH). We focus on the cumulative returns from t−60 to t+120 seconds around the news events (measured relative to t=0). We multiply returns with the sentiment direction of the article to be able to combine positive and negative news in one analysis. We exclude articles with neutral sentiment. Figure 2 shows that the market overreacts to LRH articles. In the short-run these articles have a price reaction that is very similar to HRH articles. However, after approximately 30 seconds – a reasonable time for a fast human trader to process the article – the stock price reaction to LRH articles starts to revert. After approximately 2 minutes, a large part of the short-run reaction to these articles has reversed.[14] In contrast, articles classified as HRH

[14] We would never expect full mean reversion given that LRH articles contain at least some information (see Table 1 Panel C).

have a longer-term effect on price, lasting more than two minutes. This finding is consistent with a causal effect of RavenPack that leads high frequency traders to trigger an initial overreaction to the article that is then corrected by human traders. While this price distortion is analogous to a flash crash, it has a much smaller magnitude reaching only about 1 basis point on average. This is due to the fact that the average price reaction to company-specific news is fairly small. The average absolute return reaction in the first 2 minutes after an article is only 11.4 basis points and the average signed return is only 1.9 basis points (see Table 2).

Next, we provide a multivariate analysis. The problem in studying LRH articles in a regression set-up is that we do not have an appropriate control group for these articles as they are more relevant than LRL articles, but less relevant than HRH articles (see Table 1 Panel C). Therefore, we use LRH articles from the period before RavenPack went "live" as the control group. We confirm in IA-Table 1 in the Internet Appendix that the relevance of LRH articles does not change after the introduction of RavenPack. However, after the introduction of RavenPack, LRH articles can have a causal effect on the market, which should lead to a short-run overreaction to the article. Thus, we study whether LRH articles have a stronger short run stock price impact and a larger reversal after RavenPack goes "live", as compared to before. We run the following article-level regression including only LRH articles:

$$Return\ (t-1, t+5)_a = \beta_1 * Sentiment\ Direction_a * RavenPack\ Release_t + \beta_2 * Sentiment\ Direction_a + \beta_3 * RavenPack\ Release_t + \gamma * Controls + \varepsilon_a$$

where *Sentiment Direction* indicates whether the article is positive or negative news and *RavenPack Release* is a dummy variable equal to 1 after RavenPack went "live" in April 2009. The coefficient of interest is the interaction between *Sentiment Direction* and *RavenPack Release*. In addition, we include various combinations of control variables and fixed effects. To control for

stock-specific information, we use its market capitalization, return, volatility and turnover measured over the prior month, and our illiquidity measure based on Amihud (2002). For brevity, the coefficients on these control variables are reported only in the Internet Appendix. To control for characteristics of the news announcement, we include the sentiment and article-specific variables defined in section 1.1.1. Appendix 1 contains a description of all variables.

We present the results in Table 3. In regressions 1-3, the dependent variable is the short-run stock return from 1 second before to 5 seconds after the article. We find that the short-run return associated with LRH articles is significantly more positively correlated with the sentiment of the article after RavenPack went "live" in 2009. In that the relevance of LRH articles was the same both before and after the introduction of RavenPack, this finding indicates an overreaction to these articles. Indeed, it seems plausible that algorithms would trade in the direction of the article's sentiment, because RavenPack (incorrectly) labelled such articles as highly relevant.

Next, we ask whether this overreaction is subsequently reversed. In regressions 4-6, we use the stock price reaction from 6 to 120 seconds after the LRH article as the dependent variable. We find that it is more *negatively* correlated with the article sentiment after RavenPack went "live", consistent with a reversal. While this result is not statistically significant, the negative magnitude of this coefficient is about the same as the positive magnitude of the coefficient in regressions 1-3, implying that almost all of the short run overreaction is reversed in the two minutes after the article. The fact that this result is not significant can be explained by the fact that two-minute returns are much more noisy than the five-second returns studied in regressions 1 to 3.

Taken together, our graphical and regression analyses of LRH articles results provide evidence in favor of *Hypothesis 1* that inaccuracies in news analytics can cause short term overreaction that is afterwards reversed.

3.2 News analytics and speed of stock price response

In this section, we study *Hypothesis 2*, whether news analytics improve market efficiency by increasing the speed with which stock prices and traders react to news.

3.2.1 Preliminary graphical evidence

As a first step, we conduct a purely time-series analysis and examine whether the market reaction to news is faster after RavenPack was introduced in April 2009. For this purpose, we focus only on articles reported as highly relevant in both versions (HRH) and compare the market reaction to these articles both before and after RavenPack went live. We study the reaction in terms of cumulative returns within the first 120 seconds after an article. We multiply returns by the sentiment direction to be able to combine positive and negative news in one analysis.

We plot the results in Figure 3. Because the news before and after the release of RavenPack differ in average importance, we standardize the average cumulative returns in each group by the total average cumulative return for that group after 120 seconds. Thus, the graphs show how much of the total reaction happens within a certain time period. In Panel A, we plot this share of stock price reaction separately before and after RavenPack went live. We observe a faster reaction after the introduction of RavenPack. After 10 seconds, 35.7% of the total reaction is incorporated into prices when RavenPack is live, while it is only 28.4% before RavenPack.

For a better illustration, we display the difference between the two series in Panel B. It is striking to see that the faster reaction after RavenPack went live occurs mainly in the first 5 seconds after an article is released, an interval in which only a computer could react to an article. From seconds 5 to 20, the difference stays more or less constant. After 20 to 30 seconds, it starts to decline and it is reduced to zero after 60 seconds, a time in which a fast human trader could react to an article. This finding suggests that the speed of reaction to news increases after April 2009.

While these observations are consistent with news analytics improving market efficiency by increasing the speed of the market response after an article, the increase in market efficiency after April 2009 is not necessarily due only to newswire services such as RavenPack. Rather, it might also be caused by the rise of high frequency trading or any other phenomenon happening at the same time. The ideal experiment would be to randomly select a set of articles each day and not report news analytics for them. In our regression analysis in the next section, we come close to this idea by studying relevant articles that were released as having low relevance in Old RavenPack (HRL articles). This allows us to control for general time effects.

3.2.2 Regression analysis – speed of stock price response

For the remainder of the paper, we focus on false negatives, i.e. HRL articles that are highly relevant according to New RavenPack, but were released as having low relevance in Old RavenPack. For these articles we have a good control group in the form of articles that have been reported as having high relevance in both versions (HRH). Comparing the market reaction to those two article groups allows us to see whether the market underreacts to relevant news when RavenPack does not classify it as relevant. In this case, the market will react quicker to a relevant article that is also reported as highly relevant (HRH).

We consider two alternative analyses for market reaction. First, we examine whether stock prices respond faster to HRH articles irrespective of the direction of the reaction. Then we study whether the sentiment of HRH articles predicts the directional stock price response better than the sentiment of HRL articles. For the first analysis, we define *Speed of Stock Price Response* as:

$\frac{\text{Abs(Return t}-1,\text{t+5)}}{\text{Abs(Return t}-1,\text{t+5)}+\text{Abs(Return t+6,t+120)}}$ over the 120 seconds around the news event.[15] This variable measures the amount of the two-minute price change that takes place in the first five seconds after the news release. It is in the spirit of DellaVigna and Pollet (2008). It captures the degree of under-reaction by decomposing the market reaction into its short- and long-term components. The higher the value of *Speed of Stock Price Response*, the more the reaction to the news event concentrates in the first few seconds after the event – i.e., the less under-reaction. We run the following article-level regression including only HRH and HRL articles:

$$Speed\ of\ Stock\ Price\ Response_a = \alpha_t + \alpha_f + \beta_1 * D(HRH)_a + \gamma * Controls + \varepsilon_a$$

where *D(HRH)* is a dummy variable that equals one if the article was released to the market as highly relevant (HRH) and zero if it was (incorrectly) released as having low relevance (HRL). The regression is estimated at the article level, thus allowing for both HRH and HRL articles that were released for the same firm or on the same day. This allows us to control in all regressions for unobserved heterogeneity with firm fixed effects and daily fixed effects (α_t and α_f).

We report the results in Table 4. In regressions 1 to 3, we estimate our main specification during the time in which RavenPack was live (Apr 1, 2009 – Sept 10, 2012). In regressions 4 to 6, we estimate a placebo test during the period before RavenPack was live. In regressions 2, 3, 5, and 6, we add fixed effects for the article category (e.g. mergers and acquisitions), the relevance score (from 90 to 100), and the hour during the day in which the article was released. In regressions 3 and 6, we add as additional controls the absolute return, turnover and volatility each for industry

[15] We use $\text{Abs(Return t}-1,\text{t}+5)+\text{Abs(Return t}+6,\text{t}+120)$ rather than $\text{Abs(Return t}-1,\text{t}+120)$ in the denominator to constrain the variable between 0 and 1 rather than to allow it to approach infinity in cases where $\text{Abs(Return t}-1,\text{t}+120)$ is close to zero.

and market and for the two horizons from t−1 to t+5 and t−1 to t+120 seconds around the article. All standard errors are clustered at the firm level.

The results for regressions 1 to 3 show a positive and significant relation between *Speed of Stock Price Response* and *D(HRH)*, indicating that the stock price response is much quicker for an HRH article than for an HRL article. This result holds across all the different specifications and samples. It is not only statistically significant, but also economically relevant. If we focus on the main specification (specification 3), we find that HRH articles increase the *Speed of Stock Price Response* by 1.3 percentage points or 10% relative to the mean. We find similar results if we compute *Speed of Stock Price Response* using market-adjusted and industry-adjusted returns (reported in the Internet Appendix). This finding supports *Hypothesis 2* that news analytics increase market efficiency by increasing the speed of reaction to news.

Although we showed in section 2.3 that HRH and HRL articles are similar along most dimensions, one potential concern in this set-up is that the results are driven by the two article categories having different informational content. To address this issue, we use the fact that RavenPack has back-filled the data to February 2004. If our results are driven by general differences in the two categories, then there should be a difference in stock price reaction before RavenPack went live. In regressions 4 to 6, we report the results of this placebo test in the time period where RavenPack was not yet released to investors (February 1, 2004 – March 31, 2009). In contrast to the results in regressions 1 to 3 for the period when RavenPack was live, the placebo test does not show a statistically significant relation between *D(HRH)* and the *Speed of Stock Price Response*, thereby confirming that our main test is appropriate.

Another potential concern is that there might be a general trend in the difference of informational content between HRH and HRL articles, and that this trend is driving our results

rather than the causal effect of RavenPack coverage on the market. To address this concern, we examine the relation between *Speed of Stock Price Response* and *D(HRH)* for different years before and after RavenPack went live. To implement this, we follow Gormley and Matsa (2011) and plot in Figure 4 the point estimates of a modified version of regression 3 in Table 4. In this modified regression set-up, we allow the effect of *D(HRH)* to vary by year. The control variables and the fixed effects are the same as in the main specification. Because RavenPack went "live" in the second quarter of 2009, we assign the first quarter of every year to the prior year. This way, years 2004 to 2008 were entirely before RavenPack went live, while years 2009 to 2011 were completely after RavenPack went live. We report the plot for this specification with one-year dummy variables in Panel A. In Panel B, we report the same regression but interacting *D(HRH)* with two-year dummy variables (with the first quarter shifted backwards as described above). We report 95% confidence intervals for the coefficients in both panels. In Panel C, we report the simple difference between *Speed of Stock Price Response* for HRH and HRL articles without any controls over different years (with the first quarter shifted backwards).

It is evident in the plots that the release of RavenPack magnifies the reaction to differences in versions. Before the introduction of RavenPack, the difference between HRH and HRL hovers around zero and there is no obvious time trend. After the introduction of RavenPack, the difference is much larger. This suggests the delivery of news analytics by RavenPack has an impact on the market that is separate and distinct from the underlying informational content of the news. It also suggests that our results are not driven by a spurious trend.

3.2.3 Regression analysis – directional stock price response

We now ask whether there is a relation between the stock price response and the sentiment direction of the news. That is, does the magnitude of the RavenPack-related stock price response

(via correctly-labelled HRH articles) depend on whether the news is positive, negative, or neutral? For this purpose, we ask whether the sentiment indicator in RavenPack better predicts the short run stock price reaction if an article is correctly classified as relevant (HRH) in RavenPack. We use the following article-level regression specification:

$$Return\ (t-1,t+5)_a = \alpha_t + \alpha_f + \beta_1 * D(HRH)_a * Sentiment\ Direction_a + \beta_2 * Sentiment\ Direction_a + \beta_3 * D(HRH)_a + \gamma * Controls + \varepsilon_a$$

We use the same fixed effects as before, but exclude any sentiment-related control variables as the effect of sentiment will be captured by *Sentiment Direction*.

We report the results in Table 5. In Regressions 1 to 3, we estimate our main specification during the period when RavenPack was live (Apr 1, 2009 – Sept 10, 2012). The results show a significant positive relation between returns and the interaction between *D(HRH)* and *Sentiment Direction*. That is, the RavenPack-induced stock price reaction is significantly more positive for positive news stories than for negative news stories. This result holds across all the different specifications. (Similar results for market- and industry-adjusted returns are reported in the Internet Appendix.) As before, the placebo test in Regressions 4 to 6 indicates there is not a statistically significant effect on returns before RavenPack was live. These results confirm that news analytics have a directional impact on stock prices over and above the one of the underlying news.

3.3 News analytics and trade volume response

We showed above that news analytics increase the speed of price adjustment after news is publicly released via the DJ Newswire. While the DJ Newswire constitutes a public signal, RavenPack enables faster reaction to such signals for its subscribers. Such a reaction speed advantage is modelled by Foucault, Hombert, and Rosu (2016) who predict that investors trade very aggressively when they receive a signal earlier than other market participants.

Therefore, we investigate whether the faster stock price response to an HRH article is accompanied by a faster trade volume response as well. We define *Speed of Trade Volume Response* as: $\frac{Turnover\ t-1,t+5}{Turnover\ t-1,t+120}$. The variable is defined using the same intervals as *Speed of Stock Price Response*. It captures the amount of trade volume that is concentrated in the first 5 seconds after the news event relative to the trading volume in the two minutes following the news event. We regress *Speed of Trade Volume Response* on *D(HRH)* using the same fixed effects and control variables defined above. The specification is identical to the specification for *Speed of Stock Price Response* employed in Table 4.

We report the results in Panel A of Table 6. In regressions 1 to 3, we estimate our main specification during the period in which RavenPack was live (Apr 1, 2009 – Sept 10, 2012). As in the case of *Speed of Stock Price Response*, we find a strong positive and significant relation between *Speed of Trade Volume Response* and *D(HRH)*. This result holds across all specifications. *Speed of Trade Volume Response* is 0.5 percentage points larger for HRH articles than for HRL articles, or 9% relative to the mean. In regressions 4 to 6, we estimate a placebo test in the period before RavenPack was live (Feb 1, 2004 – Mar 31, 2009). As was the case for *Speed of Stock Price Response*, the placebo test shows no significant difference in the speed of trade volume response between HRH and HRL articles before RavenPack went live.

We expect traders using RavenPack to trade in the direction of the article sentiment, so we now examine directional trading volume. Using the methodology of Lee and Ready (1991), we first determine whether a trade is initiated in the direction of the article, i.e. buyer initiated for positive articles and seller initiated for negative articles.[16] We then define *Speed of Directional*

[16] Different from Lee and Ready (1991), we use the quote at the end of the previous second as the prevailing quote rather than the quote 5 seconds ago to account for the faster trading processes today.

Trade Volume Response as $\frac{Turnover\ in\ Direction\ of\ Article\ t-1,t+5}{Turnover\ t-1,t+120}$, and use this as the dependent variable using the same regression specification (Panel B, Table 6). *Speed of Directional Trade Volume Response* is 0.4 percentage points larger for HRH than for HRL articles. Comparing it to the 0.5 percentage point increase in *Speed of Trade Volume Response* suggests that close to 80 percent of the increase in trading volume in the 5 seconds after the article is due to trading in the direction of the article. This finding suggests that RavenPack triggers fast and informed trading.

The results in these last two sections show that stock prices react faster and traders trade more aggressively after the release of articles covered in RavenPack. Combined, these results confirm that news analytics have a measurable impact on stock prices in addition to the information content of the news itself and improve price efficiency, as posited by *Hypothesis 2*.

3.4 Learning about precision in news analytics

In this section, we study whether RavenPack users dynamically learn about the signal precision of RavenPack. Such learning could be programmed into their algorithms (machine learning) or come from manual updates to the algorithms. If algorithmic traders learn about the precision of RavenPack, we expect them to rely more on RavenPack's sentiment indicators if these indicators were more informative in the past. In that case, there should be a stronger price reaction to news analytics for stocks in which news analytics have been informative in the past (*Hypothesis 3*).

This raises two questions: how to best measure the informativeness of RavenPack and whether this informativeness is persistent and thus predictable. We choose to measure informativeness by studying how well the *Sentiment Direction* of RavenPack predicts stock returns in the two minutes

following an article.[17] In IA-Table 2 in the Internet Appendix, we show that informativeness is persistent within industry. In particular, *Sentiment Direction* predicts two-minute post-article returns better if it was better in predicting these returns in the prior 3-6 months in the same industry. Thus, for an article related to industry *k*, we define *Past Informativeness* as the average *signed* two-minute post-article return for all articles related to industry *k* during the previous six months.

We study whether RavenPack subscribers trade more on articles with higher *Past Informativeness* by estimating the following article-level regression specification:

$$\begin{aligned} Ret\ (t-1,t+5)_a = {} & \alpha_t + \alpha_f + \beta_1 * Sentiment\ Direction_a * \ D(HRH)_a * Past\ Informativeness_a + \beta_2 * D(HRH)_a \\ & * Past\ Informativeness_a + \beta_3 * Sentiment\ Direction_a * Past\ Informativeness_a + \beta_4 * D(HRH)_a \\ & * Sentiment\ Direction_t + \beta_5 * D(HRH)_a + \beta_6 * Sentiment\ Direction_t + \ \beta_7 * Past\ Informativeness_a + \gamma \\ & * Controls + \varepsilon_a \end{aligned}$$

The explanatory variable of interest is the triple interaction between *Sentiment Direction*, *D(HRH)* and *Past Informativeness*. Its coefficient tests whether RavenPack's causal effect on 5-second announcement returns (Table 5) is stronger if RavenPack has been more informative in the past. This regression specification is the specification of Table 5 interacted with *Past Informativeness*.

The results are in Table 7. We define *Past Informativeness* over six months and use the 12 Fama-French industry classifications. We find a significant increase in the causal effect of RavenPack sentiment information on 5 second stock returns if *Past Informativeness* is high. A one standard deviation increase in *Past Informativeness* increases this effect by 57% relative to the average result reported in Table 5.[18] In regressions 4 to 6 of Panel A, we show in a placebo test that this

[17] It may seem more intuitive to use the number of "mistakes" RavenPack makes in assigning the relevance score, however this information is not available to the market, because it uses the New RavenPack dataset, which was only released years later. Therefore, traders could not have conditioned their trading on the number of mistakes.

[18] For Past Informativeness 6 month 12 FF: effect of 1 standard deviation: $0.225 * 1.15 = 0.25\ bp$, which is relative to the average effect from Table 5: $\frac{0.25}{0.452} = 57\%$.

effect does not occur before Ravenpack went live. In IA-Table 3 in the Internet Appendix, we report robustness checks using different definitions of *Past Informativeness*. In particular, we use 30 industries instead of 12 industries and 3 months instead of 6 months. Furthermore, we use informativeness based on whether the *direction* of the news and two-minute stock price reaction agree. This last robustness check shows that our results are not driven by any persistence in volatility. For all three alternative measure of informativeness, we obtain similar results as those presented in Table 7.

In total, these results suggest that algorithmic traders learn dynamically about the precision of RavenPack and base their trades more on RavenPack's sentiment scores if these scores have been informative in the past for that stock's industry, thereby confirming *Hypothesis 3*. This finding has interesting implications for how the market reacts to inaccurate news analytics. Because inaccurate news analytics are uninformative, subscribers will base their future trades less on RavenPack's sentiment score. Thus, inaccurate news analytics can potentially reduce the market's responsiveness to news analytics for several months following the inaccurate news analytic.

4. Robustness checks

4.1 Difference in difference specification

Until now we have mainly focused on the significant effect of RavenPack on the stock market during the period when it was live and showed in placebo tests that there is no effect for the pre-RavenPack period. However, it is possible that the placebo tests might not find significant results because of weak power. Therefore, we estimate a difference-in-difference specification for our entire sample period (February 1, 2004 – September 10, 2012) to study whether the difference between the pre- and post-RavenPack periods is statistically significant.

We report the results in Table 8. In Regressions 1-2 and 3-4, the dependent variables are *Speed of Stock Price Response* and *Speed of Trade Volume Response*, respectively. In Regressions 1 to 4, we estimate the following difference-in-difference specification:

$$DepVar_a = \alpha_t + \alpha_f + \beta_1 * D(HRH)_a * RavenPack\ Release_t + \beta_2 * D(HRH)_a + \gamma * Controls + \varepsilon_a$$

where α_t and α_f are firm and date fixed effects; *RavenPack Release* is a dummy variable equal to 1 after the release of RavenPack on April 1, 2009, and zero otherwise; and *D(HRH)* is a dummy equal to 1 for HRH articles and 0 for HRL articles.

For specifications 1-4, the explanatory variable of interest is the interaction between *D(HRH)* and *RavenPack Release*. Its coefficient is significantly positive, implying the effect of an HRH article on the speed of reaction increased significantly after the introduction of RavenPack. This is in line with our previous findings.

In regression 5 and 6, the dependent variable is the return from 1 second before to 5 seconds after the article. Because our baseline analysis used an interaction, the difference-in-difference specification uses a triple interaction:

$$\begin{aligned} Ret\ (t-1, t+5)_a = {} & \alpha_t + \alpha_f + \beta_1 * Sentiment\ Direction_a * \ D(HRH)_a * RavenPack\ Release_t + \beta_2 * D(HRH)_a * RavenPack\ Release_t \\ & + \beta_3 * Sentiment\ Direction_a * RavenPack\ Release_t + \beta_4 * D(HRH)_a * Sentiment\ Direction_t + \beta_5 * D(HRH)_a \\ & + \beta_6 * Sentiment\ Direction_t + \gamma * Controls + \varepsilon_a \end{aligned}$$

The explanatory variable of interest is this triple interaction between *D(HRH)*, *RavenPack Release* and *Sentiment Direction*. Its coefficient is statistically significantly positive, suggesting an increase in the effect of *Sentiment Direction* on returns for articles classified as HRH after RavenPack went live and confirming our previous finding.

In sum, the results of our difference-in-difference analysis are in line with our previous findings: there is a causal effect of RavenPack on the market rather than a spurious correlation.

4.2 Alternative placebo tests

Our base sample for the placebo test is Feb 2004 − Apr 2009. This time period includes the financial crisis and the introduction of Regulation National Market System (Reg NMS) which brought several changes to market structure that increased high frequency trading (Hasbrouck and Saar (2013)) and fragmentation of U.S. markets (O'Hara and Ye (2011)). To address the possible effect of these events on our analysis, we conduct two additional placebo tests: one that covers the period Feb 1, 2004 to Dec 31, 2007, thereby excluding the financial crisis; and one that covers the period July 9, 2007 to April 1, 2009, thereby excluding the time before the introduction of Reg NMS. We report the results in Panel A and B of IA-Table 4 in the Internet Appendix. We use the same regression set-ups as Tables 4 to 6. For both alternative placebo tests and for all specifications, we find no significant effects and the coefficients of interest are generally small. This suggests that the absence of significant results in our placebo sample is not driven by the inclusion of the financial crisis or the pre-Regulation NMS time period.

4.3 "Old RavenPack" definition: RavenPack 1.0 versus RavenPack 2.0

In our main specification, Old RavenPack included both RavenPack 1.0 and RavenPack 2.0. A concern is that the difference in reaction before and after the release of New RavenPack might be related to the transition from v.1.0 to v.2.0 in July 2011. Therefore, our next robustness check focuses only on RavenPack 1.0. We re-estimate the same specifications as before, but include only the period when RavenPack1.0 was live, i.e. April 1, 2009 to July 6, 2011. We report results in Panel C of IA-Table 4 in the Internet Appendix with the same regression set-up as in Panels A and B. All specifications confirm previous results and are similar in terms of economic magnitude.

4.4 Adding fixed effects for number of firms mentioned in the article

As we have shown in Section 2.3, HRL articles mention somewhat more firms than HRH articles. Thus it is a concern that this difference in the number of firms drives the difference in market reaction to these two article types. In our main tests, we address this issue by including a control variable for the number of firms and running placebo tests. In this robustness check, we control even more carefully for this issue using fixed effects. In particular, we exclude articles that mention more than 3 firms and use fixed effects for whether the article mentions 1, 2, or 3 firms. The results are presented in Panel D of IA-Table 4. All specifications confirm previous results and are similar in terms of economic and statistic magnitude.

4.5 Alternative length of event window

In our analyses we compare the stock price reaction in the short run, during which only algorithmic traders can react to an article, to the stock price reaction in the long run during which human traders will have read and traded on the article. In all of our prior analyses, we used 5 seconds as the short-run window and 120 seconds as the long-run window. Here we consider other window lengths.

In IA-Table 5 in the Internet Appendix, we show robustness to choosing longer windows lengths. In particular, we use 10 seconds as the alternative short window and 300 seconds as the alternative long window. We show robustness for Tables 4 to 6 using all three possible additional combinations of event windows: 10 seconds / 120 seconds; 5 seconds / 300 seconds; and 10 seconds / 300 seconds. The results vary slightly in magnitude, but remain statistically significant.

5. Conclusion

We study how news analytics companies affect the stock market and, in particular, market efficiency. We exploit an identification strategy based on inaccuracies in news analytics that were released to the market by RavenPack, a major provider of news analytics for algorithmic traders. Comparing the market reaction to similar news items depending on whether the news has been correctly released to customers or not, we are able to determine the causal effect of news analytics on stock prices, irrespective of the informational content of the news.

We show that news analytics have a significant impact on the market that is separate from the information contained in the news. The speed of adjustment of both stock prices and trade volume in response to a highly-relevant article is faster if the article was originally released by RavenPack as being relevant than if it was incorrectly released as not relevant.

We also consider the market response to low relevance articles that were released as having high relevance. We find that the market temporarily overreacts to these articles, although the reaction is small and reverts after 30 seconds. Thus, we show that inaccuracies in news analytics can lead to short-run price distortions. Furthermore, we provide evidence that algorithmic traders learn about the informativeness of news analytics dynamically. Thus, inaccuracies in news analytics can reduce the market's sensitivity to news analytics for a particular stock. A series of econometric robustness checks (e.g., difference-in-difference specifications, different samples, placebo tests) confirm the results.

Our findings have normative implications in terms of the recent regulatory debate on high-speed information and the effects of algorithmic and high-frequency trading. We show that news analytics improve market efficiency by speeding up the market reaction to news, while causing temporary distortions when they are wrong.

References

Amihud, Y., 2002. Illiquidity and Stock Returns: Cross-Section and Time-Series Effects. Journal of Financial Markets 5, 31–56.

Baron, Matthew, Jonathan Brogaard and Andrei Kirilenko, 2014, The Risk and Return in High Frequency Trading, Working Paper.

Benos, Evangelos and Satchit Sagade, 2012, High-frequency trading behaviour and its impact on market quality: evidence from the UK equity market, Working Paper.

Biais, Bruno, Thierry Foucault and Sophie Moinas, 2015, Equilibrium Fast Trading, Journal of Financial Economics.

Boehmer, Ekkehart, Kingsley Fong, and Julie Wu, 2015, International evidence on algorithmic trading, Working Paper.

Boudoukh, Jacob, Ronen Feldman, Shimon Kogan and Matthew Richardson, 2012, Which News Moves Stock Prices, Working Paper.

Brogaard, Jonathan, Terrence Hendershott, and Ryan Riordan, 2014, High frequency trading and price Discovery,The Review of Financial Studies, 27, n. 8. .

Brogaard, Jonathan, Al Carrion, Thibaut Moyaert, Ryan Riordan, Andriy Shkilko, Konstantin Sokolov, 2015, High-Frequency Trading and Extreme Price Movements, Working Paper.

Chaboud, Alain, Ben Chiquoine, Erik Hjalmarsson, and Clara Vega, 2013, Rise of the Machines: Algorithmic Trading in the Foreign Exchange Market, Journal of Finance, forthcoming.

Chan, W. S., 2003, Stock price reaction to news and no-news Drift and reversal after headlines, Journal of Financial Economics, 702, 223-260.

Chordia, Tarun, T. Clifton Green, and Badrinath Kottimukkalur, 2015, Do High Frequency Traders Need to be Regulated? Evidence from Algorithmic Trading on Macro News, Working Paper.

Clark-Joseph, Adam D., 2013, Exploratory Trading, Working Paper.

Das, Sanjiv R., and Mike Y. Chen, Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, Management Science 53, 1375-1388.

DellaVigna, S., and Pollet, J., 2009, Investor Inattention, Firm Reaction, and Friday Earnings Announcements, Journal of Finance, 64, 709-749.

Dougal, Casey, Joseph Engelberg, Diego Garcia and Christopher Parsons, 2012, Journalists and the Stock Market, Review of Financial Studies, 25, 639-679.

Dugast, Jerome and Thierry Foucault, 2017, Data Abundance and Asset Price Informativeness, Working Paper.

Engelberg, Joseph and Parsons, Christopher A., 2011, The Causal Impact of Media in Financial Markets, Journal of Finance.

Fang, Lily H. and Peress, Joel, 2009, Media coverage and the cross-section of stock returns, Journal of Finance, 64, 2023-2052.

Ferguson, Nicky J., Dennis Philip, Herbert Y.T. Lam and Jie Michael Guo, 2015, Media Content and Stock Returns: The Predictive Power of Press, Multinational Finance Journal 19, 1-31.

Foucault, Thierry, Johan Hombert and Ioanid Rosu, 2016, News Trading and Speed, Journal of Finance 71, 335-382.

Gai, Jiading, Chen Yao and Mao Ye, 2013, The Externalities of High Frenquency Trading, Working Paper.

Garcia, Diego, Sentiment during Recessions, Journal of Finance 68, 1267–1300.

Gerig, Ausin, 2015, High-Frequency Trading Synchronizes Prices in Financial Markets, Working Paper.

Golub, Anton, John Keane and Ser-Huang Poon, 2012, High Frequency Trading and Mini Flash Crashes, Working Paper.

Gormley, Todd A., and David A. Matsa, 2011, Growing Out of Trouble? Corporate Responses to Liability Risk, Review of Financial Studies 24, 2781-2821.

Griffin, John M., Nicholas H. Hirschey, and Patrick J. Kelly, How Important Is the Financial Media in Global Markets?, Review of Financial Studies 24, 3941-3992.

Groß-Klußmann, Axel and Nikolaus Hautsch, 2011, When machines read the news: Using automated text analytics to quantify high frequency news-implied market reactions, Journal of Empirical Finance, 18, 321-340.

Hasbrouck, Joel and Gideon Saar, 2013, Low-Latency Trading, Journal of Financial Markets 16, 646–679.

Hendershott, Terrence and Ryan Riordan, 2013, Algorithmic Trading and the Market for Liquidity, Journal of Financial and Quantitative Analysis 48, 1001-1024.

Hendershott, Terrence, Charles M. Jones and Albert J. Menkveld, 2011, Does Algorithmic Trading Improve Liquidity?, Journal of Finance, 66, 1-33.

Heston, Steven L. and Nitish R. Sinha, 2016, News versus Sentiment: Predicting Stock Returns from News Stories, Working Paper.

Hirschey, Nicholas, 2013, Do High-Frequency Traders Anticipate Buying and Selling Pressure?, Working Paper.

Hu, Grace X., Jun Pan and Jiang Wang, 2013, Early Peek Advantage?, Working Paper.

Jegadeesh, Narasimhan, and Di Wu, 2013, Word power: A new approach for content analysis, Journal of Financial Economics 110, 712-729.

Jones, Charles, 2013, What do we know about high-frequency trading?, Working Paper.

Jovanovic, Boyan and Albert J. Menkveld, 2011, Middlemen in Limit-Order Markets, Working Paper.

Kyle, Albert, 1985, Continuous auctions and insider trading, Econometrica 53, 1315-1336.

Lee, Charles M. and Mark J. Ready, Inferring Trade Direction from Intraday Data, Journal of Finance 46, 733-746.

Loughran, Tim and Bill McDonald, 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks, Journal of Finance 66, 35-65.

Martinez, Victor H., and Ioanid Rosu 2013, High Frequency Traders, News and Volatility, Working Paper.

Menkveld, Albert, 2013, High frequency trading and the *new market* makers, Journal of Financial Markets, 16, 712-740.

O'Hara, Maureen and Mao Ye, 2011, Is market fragmentation harming market quality?, Journal of Financial Economics, 100, 459–474.

Peress, Joel, 2014, The Media and the Diffusion of Information in Financial Markets: Evidence from Newspaper Strikes, Journal of Finance, 69, 2007-2043.

Riordan, Ryan, Andreas Storkenmaier, Martin Wagener and S. Sarah Zhang, 2013, Public information arrival: Price discovery and liquidity in electronic limit order markets, Journal of Banking and Finance, 37, 1148-1159.

Riordan, Ryan and Andreas Storkenmaier, 2012, Latency, liquidity and price discovery, Journal of Financial Markets, 15, 416-437.

Sinha, Nitish Rajan, 2012, Underreaction to News in the US Stock Market, Working Paper.

Tetlock, Paul, 2007, Giving content to investor sentiment The role of media in the stock market, Journal of Finance, 62, 1139–1168.

Tetlock, Paul, 2011, All the News That's Fit to Reprint: Do Investors React to Stale Information?, Review of Financial Studies, 24, 1481-1512.

Weller, Brian, 2015, Efficient Prices at Any Cost: Does Algorithmic Trading Deter Information Acquisition?, Working Paper.

Zhang, Sarah S., 2013, Need for Speed: An Empirical Analysis of Hard and Soft Information in a High Frequency World, Working Paper.

Figure 1: Market reaction by Relevance Score

This figure displays cumulative signed return (relative to the time of the article) from 60 seconds before to 120 seconds after the article for news events from April 1, 2009 to September 10, 2012 (the time when RavenPack was live). Signed returns are returns are multiplied with the sentiment direction of the article. We exclude articles with neutral sentiment. *Low Relevance* refers to articles with a Relevance Score below 90 in both RavenPack versions, while *High Relevance* refers to articles that have a Relevance Score greater or equal than 90 in both RavenPack versions.

Cumulative signed return

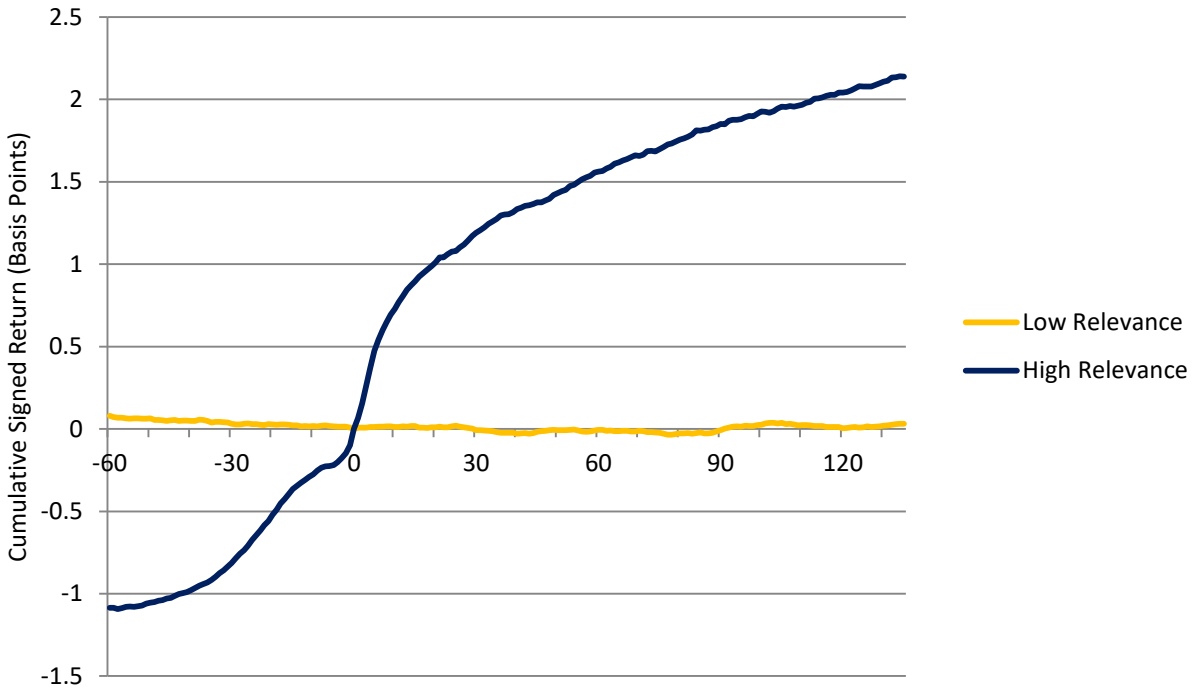


Figure 2: Difference in Stock Price Response between HRH and LRH Articles

This figure displays cumulative signed return (relative to the time of the article) from 60 seconds before to 120 seconds after the article for news events from April 1, 2009 to September 10, 2012 (the time when RavenPack was live). Returns are multiplied with the sentiment direction of the article. We exclude articles with neutral sentiment. HRH refers to articles that have a relevance scores greater or equal 90 in both RavenPack versions, while LRH refers to articles that had a relevance score greater or equal 90 in the old RavenPack version while having Relevance below 90 in the new RavenPack version.

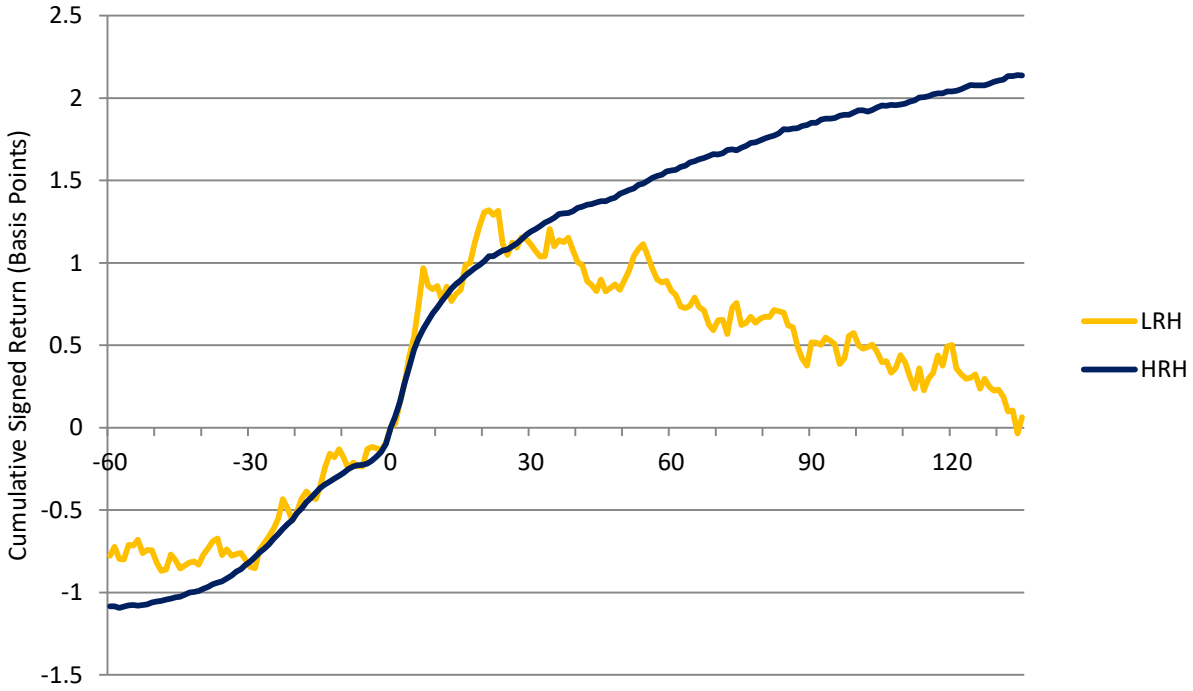
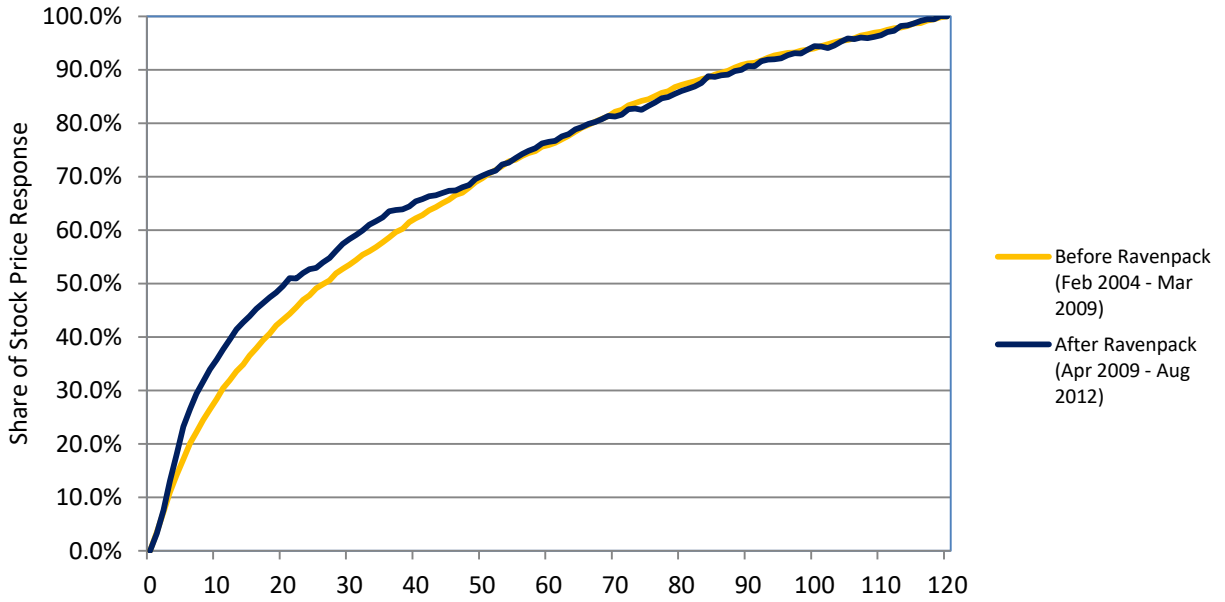


Figure 3: Difference in market reaction after RavenPack is live

The figure in Panel A displays the share of the total stock price response to news within the first 120 seconds after an article. We compare the reaction to articles before (Feb 2004 – Mar 2009) and after RavenPack went live (April 2009 to 10 September 2012). Returns are multiplied with the sentiment direction of the article. We exclude articles with neutral sentiment. We standardize the average cumulative return within each group by dividing it by the total average cumulative return for that group after 120 seconds. We only include articles that are consistently reported as relevant (HRH) in both versions. In Panel B, we display the difference between the two series from Panel A.

Panel A: Share of Stock Price Reaction before vs. after RavenPack is live



Panel B: Difference in Share of Stock Price Reaction before vs. after RavenPack is live

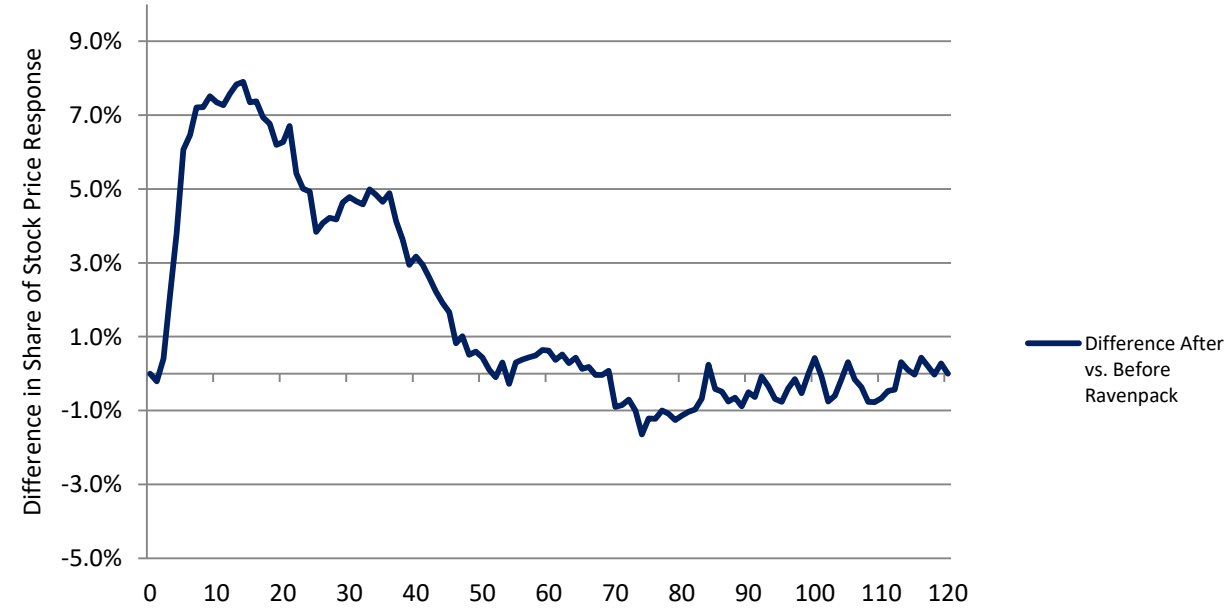
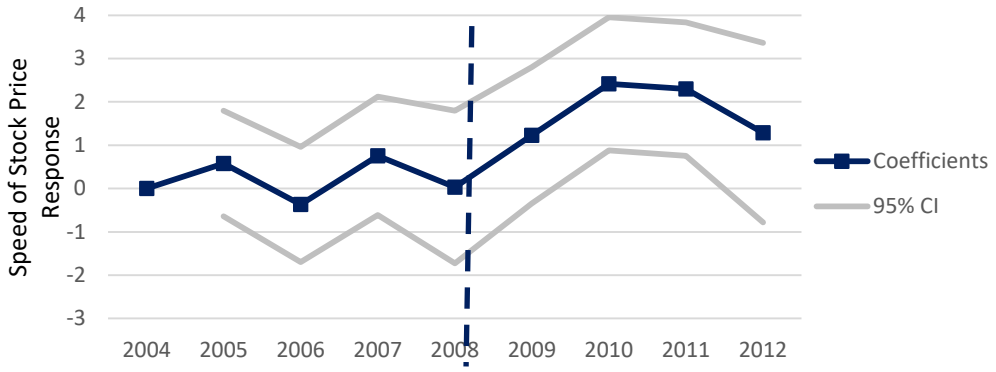


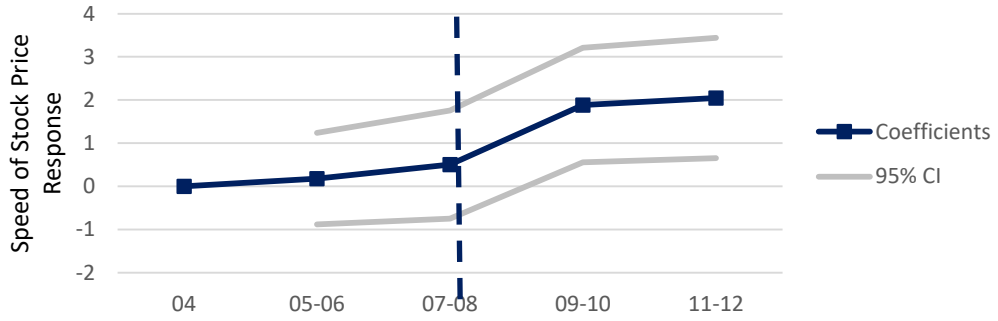
Figure 4: Difference in Speed of Stock Price Response Over Time

The figure in Panel A reports the point estimates from an OLS regression of Speed of Stock Price Response ($\frac{\text{Abs(Return t}-1,t+5)}{\text{Abs(Return t}-1,t+5)+\text{Abs(Return t}+6,t+120)}$) on D(HRH) interacted with yearly dummy variables from 2004 to 10 September 2012. We assign the first quarter of a year to the prior year, i.e. the 2009 dummy covers a time period from 1 April 2009 to 1 April 2010. Controls and fixed effects are the same as in table 3 regression 3. The vertical line indicates the introduction of RavenPack on 1 April 2009. In Panel B, we report the same regression but interacting the HRH dummy variable with two-year dummy variables (with the first quarter shifted backwards). In Panel C, we report the difference between Speed of Stock Price Response for HRH and HRL articles over different years (with the first quarter shifted backwards).

Panel A: Estimate of coefficient on D(HRH) interacted with yearly dummies



Panel B: Estimate of coefficient on D(HRH) interacted with two-year dummies



Panel C: Comparing the difference in mean

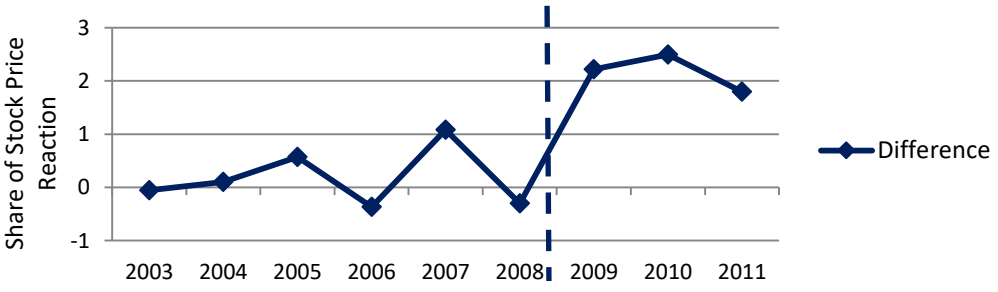


Table 1: Overview of Four Article Types

In Panel A, we present our predictions for the market reaction to different articles. In Panel B, we present the results of article-level regressions that examine the market reaction to different types of articles in the time period where RavenPack was not yet sold to investors. The dependent variables are the absolute returns and turnover in the two minutes after the article. Returns are based on mid-quotes. The explanatory variable of interest are D(HRH), D(HRL) and D(RLH), which are dummy variables for these article categories (LRL is the omitted category). At the bottom of the table we display the t-statistic for the difference between HRH and HRL articles. All standard errors are clustered at the firm level. T-statistics are below the parameter estimates in parenthesis; ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Panel A: Predictions for the market reaction of the different article types

		New RavenPack	
		High Relevance Article	**Low Relevance Article**
Old RavenPack	**High Relevance Article**	HRH: Fast and persistent market reaction	LRH: Fast market reaction that mean-reverts (overreaction).
	Low Relevance Article	HRL: Slow market reaction (underreaction)	LRL: No stock price reaction

Panel B: Stock price reaction to the different article types BEFORE RavenPack went "live"

Dependent Variable:	Absolute Return t−1,t+120	Turnover t−1,t+120
D (HRH)	3.235***	0.335***
	(36.37)	(33.75)
D (HRL)	3.118***	0.325***
	(11.74)	(9.08)
D (LRH)	2.494***	0.311***
	(9.24)	(9.34)
Number of Observations	2214726	2214726
R^2	0.151	0.184
Hour Fixed Effects	Yes	Yes
Date and Firm Fixed Effects	Yes	Yes
Difference coefficients of D(HRH) and D(HRL)	0.117	0.010
(t-stat)	(0.43)	(0.29)

Table 2: Summary Statistics – Relevant articles, Apr 2009 to Sept 2012

This table displays summary statistics for the 321,912 article-company combinations after RavenPack went "live" (April 1, 2009 to September 10, 2012). These article-company observations are classified as relevant in the new RavenPack (i.e. they are HRH or HRL). *Market capitalization* is the number of shares outstanding multiplied by the prior day closing price. *Average volatility prior month* is the mean of daily squared return in the 20 trading days before the article. *Average turnover prior month* is the mean of daily trading volume divided by shares outstanding in the 20 trading days before the article. *Absolute return t−1, t+5* is the absolute stock return from 1 second before to 5 seconds after the article. *Speed of Stock Price Response* is defined as $\frac{\text{Abs(Return t}-1,t+5)}{\text{Abs(Return t}-1,t+5)+\text{Abs(Return t}+6,t+120)}$. *Turnover t−1, t+5* is trading volume divided by shares outstanding from 1 second before to 5 seconds after the article. *Speed of Trade Volume Response* is defined as $\frac{\text{Turnover t}-1,t+5}{\text{Turnover t}-1,t+120}$. *Return on trading day* is the stock return over the entire trading day that the article was released. *Absolute return on trading day* is its absolute value. *Time since last company article* is the time since the company was last mentioned in an article. *Number of firms in article* defines the number of companies mentioned in the article. *Composite Sentiment Score* is a sentiment score that is provided by RavenPack and takes a value from 100 (positive) to 0 (negative). *Absolute Composite Sentiment Score* is defined as Abs (Composite Sentiment Score – 50). *Neutral Composite Sentiment Score* is a dummy variable equal to 1 if the *Composite Sentiment Score* equals 50. *Article Category Identified* is a dummy variable equal to 1 if the article category (e.g. merger and acquisitions) is identified by RavenPack. *Event Sentiment Score* is a sentiment score that is provided by RavenPack and takes a value from 100 (positive) to 0 (negative); this is available only for articles for which the category is identified. *Absolute Event Sentiment Score* is defined as Abs (Event Sentiment Score – 50). *Neutral Event Sentiment Score* is a dummy variable equal to 1 if the *Event Sentiment Score* equals 50. In Panel A, we report descriptive statistics. In Panel B we report the difference between articles that were consistently released as relevant in both RavenPack versions (HRH) and those that were released as having low relevance (HRL). The difference is defined as the regression coefficient of D(HRH) in a regression of the respective variable on D(HRH) and Relevance, Category, Hour and Date Fixed Effects. D(HRH) is a dummy equal to 1 if the article is HRH. We also report t-statistics for the coefficient clustered at the firm level. ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Panel A: Descriptive Statistics

	Mean	25th Percentile	Median	75th Percentile	Standard Deviation
Market capitalization ($ million)	13185.0	157.4	1782.9	30027.4	37016.1
Average return prior month (%)	0.12	-0.57	0.10	0.79	0.65
Average volatility prior month (%)	9.69	1.19	4.79	20.4	17.7
Average turnover prior month (%)	1.17	0.27	0.83	2.29	1.23
Absolute Return t−1,t+5 (basis points)	1.95	0	0	4.43	9.46
Absolute Return t−1,t+120 (basis points)	11.4	0	5.00	27.4	21.7
Speed of Stock Price Response (%)	13.2	0	0	50	24.7
Signed Return t−1,t+5 (basis points)	0.60	-1.38	0	1.97	10.2
Signed Return t−1,t+120 (basis points)	1.89	-15.1	0	18.5	25.6
Turnover t−1,t+5 (basis points)	0.041	0	0	0.084	0.14
Turnover t−1,t+120 (basis points)	0.86	0	0.24	1.81	2.22
Speed of Trade Volume Response (%)	5.80	0	0	16.8	13.9
Return on trading day (%)	0.23	-3.29	0.056	3.92	4.02
Absolute return on trading day (%)	2.48	0.22	1.45	5.65	3.18
Time since last company article (hours)	32.2	0.49	6.42	103.1	57.6
Number of companies in article	2.14	1	1	3	4.30
Composite Sentiment Score	50.0	47	50	52	4.19
Absolute Composite Sentiment Score	2.07	0	2	5	3.65
Neutral Composite Sentiment Score	0.47	0	0	1	0.50
Article category identified	0.35	0	0	1	0.48
Event Sentiment Score	51.8	37	50	67	12.9
Absolute Event Sentiment Score	3.83	0	0	13	6.71
Neutral Event Sentiment Score	0.69	0	1	1	0.46
Past Informativeness 6 month 12FF	1.67	0.58	1.39	3.46	1.15
Past Informativeness 3 month 12FF	1.58	0.41	1.22	3.29	1.23
Past Informativeness 6 month 30FF	1.69	0.39	1.34	3.66	1.46
Number of Observations	321,912				

Panel B: Comparison between Accurately Classified as Relevant (HRH) vs. Misclassified (HRL)

	Standard Deviation	Difference between HRH and HRL after fixed effects	T- Statistic	Difference in terms of Standard Deviations
Market capitalization ($ million)	37016.1	-921.79	-0.25	-0.0249
Average return prior month (%)	0.65	-0.0319***	-3.05	-0.04908
Average volatility prior month (%)	17.7	-1.526*	-1.70	-0.08621
Average turnover prior month (%)	1.23	-0.0773	-0.94	-0.06285
Average illiquidity prior month (percentile)	26.4	-2.0874	-0.84	-0.07907
Absolute Return t−1,t+120 (basis points)	21.7	-0.3433	-0.77	-0.01582
Turnover t−1,t+120 (basis points)	2.22	0.0731	1.19	0.032928
Return on trading day (%)	4.02	-0.0953	-1.58	-0.02371
Absolute return on trading day (%)	3.18	-0.1242	-0.93	-0.03906
Time since last company article (hours)	57.6	3.32	1.24	0.057639
Number of companies in article	4.30	-0.95***	-3.23	-0.22093
Composite Sentiment Score	4.19	-0.0679	-0.63	-0.01621
Absolute Composite Sentiment Score	3.65	0.0105	0.06	0.002877
Neutral Composite Sentiment Score	0.50	-0.0353	-0.67	-0.0706
Event Sentiment Score	12.9	-0.7440	-1.09	-0.05767
Absolute Event Sentiment Score	6.71	-0.0986*	-1.77	-0.01469
Neutral Event Sentiment Score	0.46	-0.0020	-0.97	-0.00435

Table 3: Overreaction to News Analytics (LRH articles)

This table contains the results of article-level regressions that examine how well the sentiment direction of LRH articles predicts stock returns before and after the release of RavenPack. In regressions 1 to 3, the dependent variable is the return from 1 second before to 5 seconds after the article (measured in basis points). In regressions 4 to 6, we study the return from 6 to 120 seconds after the article to determine a potential reversal of the short run reaction. Returns are based on mid-quotes. The explanatory variable of interest is an interaction between *RavenPack Release* and *Sentiment Direction*. RavenPack Release is a dummy variable equal to 1 during the time in which RavenPack was "live" (April 1, 2009 – September 10, 2012) and equal to 0 before RavenPack was "live" (February1, 2004 – March 31, 2009). *Sentiment Direction* is a variable indicating the sentiment of the article derived from RavenPack sentiment indices; it takes the value +1 for positive sentiment, 0 for neutral sentiment and −1 for negative sentiment. In all regressions we include the following firm specific control variables: Company size, Return prior month, Volatility prior month, Turnover prior month, Illiquidity prior month. In regressions 2, 3, 5 and 6 we add fixed effects for the article category (e.g. mergers and acquisitions), the relevance score (from 90 to 100) and the hour during the day in which the article was released. In regressions 3 and 6, we include absolute return, turnover, and volatility each for industry and market from t−1 to t+5 seconds around the article. All variables are defined in Appendix 1. All standard errors are clustered at the firm level. T-statistics are below the parameter estimates in parenthesis; ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Dependent Variable:	Return t−1, t+5			Return t+6, t+120		
	(1)	(2)	(3)	(4)	(5)	(6)
RavenPack Release * Sentiment Direction	0.465**	0.533**	0.563***	-0.602	-0.666	-0.660
	(2.31)	(2.48)	(2.60)	(-1.24)	(-1.32)	(-1.31)
RavenPack Release	-0.252	-0.258	-0.309	-1.739***	-1.659***	-1.643***
	(-1.50)	(-1.44)	(-1.62)	(-3.92)	(-3.71)	(-3.42)
Sentiment Direction	0.258**	0.098	0.099	1.532***	1.517***	1.538***
	(2.28)	(0.80)	(0.81)	(5.10)	(4.58)	(4.64)
Article category identified		-1.177***	-1.217***		0.043	-0.316
		(-3.14)	(-3.04)		(0.04)	(-0.27)
Time since last article		0.112**	0.107**		0.373***	0.357***
		(2.40)	(2.30)		(2.91)	(2.78)
Number of firms in article		-0.179**	-0.171**		-0.397*	-0.415*
		(-2.19)	(-2.12)		(-1.70)	(-1.79)
Number of Observations	20588	20588	20588	20588	20588	20588
R^2	0.003	0.009	0.014	0.007	0.013	0.018
Relevance, Category and Hour Fixed Effects	No	Yes	Yes	No	Yes	Yes
Market control variables	No	No	Yes	No	No	Yes
Firm specific control variables	Yes	Yes	Yes	Yes	Yes	Yes

Table 4: Speed of Stock Price Response to News Articles

This table contains the results of article-level regressions that examine the effect of an article covered in RavenPack on stock price, measured by absolute returns. The dependent variable is *Speed of Stock Price Response* (in percent) defined as $\frac{\text{Abs(Return t}-1,t+5)}{\text{Abs(Return t}-1,t+5)+\text{Abs(Return t}+6,t+120)}$ and measured in seconds around an article. Returns are based on mid-quotes. The explanatory variable of interest is D(HRH), a dummy variable equal to 1 if an article was consistently released as highly relevant in both RavenPack versions and 0 if it was originally released as having low relevance (HRL). In regressions 1 to 3, we estimate the various specification during the time in which RavenPack was "live" (April 1, 2009 – September 10, 2012). In regressions 4 to 6, we run a placebo test for the time period where RavenPack was not yet sold to investors (February1, 2004 – March 31, 2009). In all regressions we include firm and date fixed effects and the following firm specific control variables: Company size, Return prior month, Volatility prior month, Turnover prior month, Illiquidity prior month. In regressions 2, 3, 5 and 6 we add fixed effects for the article category (e.g. mergers and acquisitions), the relevance score (from 90 to 100) and the hour during the day in which the article was released. In regressions 3 and 6, we include additional controls: the absolute return, turnover, and volatility each for industry and market and for the two horizons from t−1 to t+5 and t−1 to t+120 seconds around the article. All variables are defined in Appendix 1. All standard errors are clustered at the firm level. T-statistics are below the parameter estimates in parenthesis; ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Dependent Variable:	Speed of Stock Price Response					
	Main Test - RavenPack is "live"			Placebo Test - Before RavenPack is "live"		
	(1)	(2)	(3)	(4)	(5)	(6)
D(HRH)	1.469***	1.333***	1.321***	-0.048	-0.058	-0.012
	(3.50)	(3.15)	(3.19)	(-0.16)	(-0.20)	(-0.04)
Absolute Composite Sentiment Score		-0.004	-0.002		-0.032**	-0.032**
		(-0.22)	(-0.11)		(-2.29)	(-2.31)
Neutral Composite Sentiment Score		-0.107	-0.115		-0.338***	-0.357***
		(-0.87)	(-0.96)		(-3.30)	(-3.52)
Article category identified		0.522	1.395		-3.574	-4.141
		(0.10)	(0.28)		(-0.54)	(-0.68)
Absolute Event Sentiment Score		0.098***	0.089***		0.021*	0.022*
		(5.13)	(4.78)		(1.70)	(1.85)
Neutral Event Sentiment Score		-0.818	-0.958		-1.202***	-1.150***
		(-1.29)	(-1.56)		(-2.92)	(-2.82)
Time since last article		0.099***	0.086**		0.069***	0.062**
		(2.85)	(2.54)		(2.72)	(2.44)
Number of firms in article		-0.060	-0.058		-0.149***	-0.170***
		(-0.70)	(-0.69)		(-2.65)	(-3.06)
Number of Observations	249065	249065	249065	400303	400303	400303
R^2	0.035	0.039	0.084	0.032	0.033	0.049
Relevance, Category and Hour Fixed Effects	No	Yes	Yes	No	Yes	Yes
Date and Firm Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Market control variables	No	No	Yes	No	No	Yes
Firm specific control variables	Yes	Yes	Yes	Yes	Yes	Yes

Table 5: Directional Stock Price Response to Article Sentiment

This table contains the results of article-level regressions that examine how well the sentiment direction of an article predicts the 5-second return reaction to an article depending on whether the article is covered in RavenPack. The dependent variable is the return from 1 second before to 5 seconds after the article (measured in basis points). Returns are based on mid-quotes. The explanatory variable of interest is an interaction between D(HRH) and *Sentiment Direction*. D(HRH) is a dummy variable equal to 1 if an article was consistently released as highly relevant in both RavenPack versions and 0 if it was originally released as having low relevance (HRL). *Sentiment Direction* is a variable indicating the sentiment of the article derived from RavenPack sentiment indices; it takes the value +1 for positive sentiment, 0 for neutral sentiment and −1 for negative sentiment. In regressions 1 to 3, we estimate the various specification during the time in which RavenPack was "live" (April 1, 2009 – September 10, 2012). In regressions 4 to 6, we run a placebo test for the time period where RavenPack was not yet sold to investors (February1, 2004 – March 31, 2009). In all regressions we include firm and date fixed effects and the following firm specific control variables: Company size, Return prior month, Volatility prior month, Turnover prior month, Illiquidity prior month. In regressions 2, 3, 5 and 6 we add fixed effects for the article category (e.g. mergers and acquisitions), the relevance score (from 90 to 100) and the hour during the day in which the article was released. In regressions 3 and 6, we include absolute return, turnover, and volatility each for industry and market from t−1 to t+5 seconds around the article. All variables are defined in Appendix 1. All standard errors are clustered at the firm level. T-statistics are below the parameter estimates in parenthesis; ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Dependent Variable:	Return t-1, t+5					
	Main Test - RavenPack is "live"			Placebo Test - Before RavenPack is "live"		
	(1)	(2)	(3)	(4)	(5)	(6)
D(HRH) * Sentiment Direction	0.407***	0.452***	0.452***	0.081	0.116	0.114
	(3.09)	(3.39)	(3.40)	(0.79)	(1.10)	(1.09)
D(HRH)	0.187*	0.125	0.125	0.137	0.103	0.102
	(1.86)	(1.19)	(1.18)	(1.28)	(0.94)	(0.93)
Sentiment Direction	0.118	-0.010	-0.009	0.421***	0.184*	0.187*
	(0.92)	(-0.08)	(-0.07)	(4.23)	(1.81)	(1.84)
Article category identified		1.151**	0.932*		0.304	0.370
		(1.97)	(1.86)		(0.70)	(0.84)
Time since last article		0.043***	0.043***		0.237***	0.239***
		(3.38)	(3.40)		(13.20)	(13.28)
Number of firms in article		-0.150***	-0.145***		-0.221***	-0.220***
		(-5.05)	(-4.90)		(-9.37)	(-9.31)
Number of Observations	321860	321860	321860	481939	481939	481939
R^2	0.063	0.066	0.069	0.057	0.062	0.063
Relevance, Category and Hour Fixed Effects	No	Yes	Yes	No	Yes	Yes
Date and Firm Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Market control variables	No	No	Yes	No	No	Yes
Firm specific control variables	Yes	Yes	Yes	Yes	Yes	Yes

Table 6: Speed of Trade Volume Response to News Articles

This table contains the results of article-level regressions that examine the effect of an article covered in RavenPack on the market for a stock, measured by turnover. In Panel A, the dependent variable is *Speed of Trade Volume Response* (in percent), defined as the turnover from 1 second before the article to 5 second after the article divided by the turnover from 1 second before the article to 120 seconds after the article. The explanatory variable of interest is D(HRH), a dummy variable equal to 1 if an article was consistently released as highly relevant in both RavenPack versions and 0 if it was originally released as having low relevance (HRL). In regressions 1 to 3, we estimate the various specification during the time in which RavenPack was "live" (April 1, 2009 – September 10, 2012). In regressions 4 to 6, we run a placebo test for the time period where RavenPack was not yet sold to investors (February1, 2004 – March 31, 2009). In all regressions we include firm and date fixed effects and the following firm specific control variables: Company size, Return prior month, Volatility prior month, Turnover prior month, Illiquidity prior month. In regressions 2, 3, 5 and 6 we add fixed effects for the article category (e.g. mergers and acquisitions), the relevance score (from 90 to 100) and the hour during the day in which the article was released. In regressions 3 and 6, we include additional controls: the absolute return, turnover, and volatility each for industry and market and for the two horizons from t−1 to t+5 and t−1 to t+120 seconds around the article. In panel B, we split the trading volume by who initiated the trade using the algorithm of Lee and Ready (1991). In Panel B, the dependent variable is *Speed of Directional Trade Volume Response*, defined as the turnover initiated in direction of the article from 1 second before the article to 5 second after the article divided by the turnover from 1 second before the article to 120 seconds after the article. The control variables in Panel B are the same as in Panel A, but not reported for brevity. All variables are defined in Appendix 1. All standard errors are clustered at the firm level. T-statistics are below the parameter estimates in parenthesis; ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Panel A: Trade Volume

Dependent Variable:	Speed of Trade Volume Response					
	Main Test - RavenPack is "live"			Placebo Test - Before RavenPack is "live"		
	(1)	(2)	(3)	(4)	(5)	(6)
D(HRH)	0.656***	0.516**	0.533**	0.033	-0.002	0.022
	(3.12)	(2.40)	(2.52)	(0.24)	(-0.02)	(0.16)
Absolute Composite Sentiment Score		-0.008	-0.006		-0.018***	-0.017**
		(-0.94)	(-0.74)		(-2.65)	(-2.52)
Neutral Composite Sentiment Score		-0.171**	-0.164**		-0.085	-0.089*
		(-2.55)	(-2.48)		(-1.64)	(-1.70)
Article category identified		-4.034***	-3.973***		-0.716	-0.690
		(-2.93)	(-3.37)		(-0.47)	(-0.45)
Absolute Event Sentiment Score		0.060***	0.056***		0.023***	0.024***
		(5.70)	(5.37)		(3.57)	(3.70)
Neutral Event Sentiment Score		-0.973***	-1.018***		-0.394*	-0.368*
		(-2.84)	(-3.01)		(-1.82)	(-1.71)
Time since last article		0.109***	0.101***		0.091***	0.087***
		(5.74)	(5.38)		(6.01)	(5.83)
Number of firms in article		-0.107***	-0.112***		-0.168***	-0.173***
		(-2.61)	(-2.75)		(-6.35)	(-6.55)
Number of Observations	272215	272215	272215	418252	418252	418252
R^2	0.029	0.032	0.059	0.026	0.027	0.038
Relevance, Category and Hour Fixed Effects	No	Yes	Yes	No	Yes	Yes
Date and Firm Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Market control variables	No	No	Yes	No	No	Yes
Firm specific control variables	Yes	Yes	Yes	Yes	Yes	Yes

Panel B: Trade volume in direction of article

Dependent Variable:	Speed of Directional Trade Volume Response					
	Main Test - RavenPack is "live"			Placebo Test - Before RavenPack is "live"		
	(1)	(2)	(3)	(4)	(5)	(6)
D(HRH)	0.472***	0.376**	0.385**	0.080	0.085	0.082
	(2.69)	(2.13)	(2.21)	(0.60)	(0.64)	(0.61)
Number of Observations	168499	168499	168499	272734	272734	272734
R^2	0.050	0.055	0.067	0.041	0.042	0.047
Relevance, Category and Hour Fixed Effects	No	Yes	Yes	No	Yes	Yes
Date and Firm Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Market control variables	No	No	Yes	No	No	Yes
Firm specific control variables	Yes	Yes	Yes	Yes	Yes	Yes

Table 7: Directional Stock Price Response conditional on Past Informativeness of RavenPack

This table contains the results of article-level regressions that examine how well the past performance of RavenPack affects the stock price impact of RavenPack. The dependent variable is the return from 1 second before to 5 seconds after the article (measured in basis points). Returns are based on mid-quotes. The explanatory variable of interest is a triple interaction between *D(HRH)* and *Sentiment Direction* and *Past Informativeness*. *Past Informativeness* is the average signed return (in basis points) from t-1 to t+120 seconds around articles over the previous 6 month for stocks within the same industry. D(HRH) is a dummy variable equal to 1 if an article was consistently released as highly relevant in both RavenPack versions and 0 if it was originally released as having low relevance (HRL). *Sentiment Direction* is a variable indicating the sentiment of the article derived from RavenPack sentiment indices; it takes the value +1 for positive sentiment, 0 for neutral sentiment and −1 for negative sentiment. We estimate our main specification using *Past Informativeness* measured over the previous six months and using the 12 industry categories of Fama French. In regressions 1 to 3, we estimate the various specification during the time in which RavenPack was "live" (April 1, 2009 – September 10, 2012). In regressions 4 to 6, we run a placebo test for the time period where RavenPack was not yet sold to investors (February1, 2004 – March 31, 2009). In IA-Table 3 in the Internet Appendix, we report a robustness check using 30 FF industry categories (instead of 12) and *Past Informativeness* measured over the previous three months (instead of a six). In all regressions we include firm and date fixed effects and the following firm specific control variables: Company size, Return prior month, Volatility prior month, Turnover prior month, Illiquidity prior month. In regressions 2, 3, 5 and 6, we add fixed effects for the article category (e.g. mergers and acquisitions), the relevance score (from 90 to 100) and the hour during the day in which the article was released. In regressions 3 and 6, we add additional controls: the absolute return, turnover, and volatility each for industry and market and for the two horizons from t−1 to t+5 seconds around the article. Control variables are defined in Appendix 1. All standard errors are clustered at the firm level. T-statistics are below the parameter estimates in parenthesis; ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Panel A: Main Test

Dependent Variable:	Return t−1, t+5					
	Main Test - RavenPack is "live"			Placebo Test - Before RavenPack is "live"		
	(1)	(2)	(3)	(4)	(5)	(6)
Past Informativeness 6 month 12FF * D(HRH) * Sentiment Direction	0.221***	0.213***	0.225***	-0.178*	-0.165	-0.168
	(2.92)	(2.77)	(2.85)	(-1.71)	(-1.60)	(-1.63)
Past Informativeness 6 month 12FF * Sentiment Direction	-0.025	-0.034	-0.049	0.354***	0.311***	0.310***
	(-0.35)	(-0.46)	(-0.65)	(3.51)	(3.10)	(3.10)
Past Informativeness 6 month 12FF * D(HRH)	0.044	0.026	0.037	0.050	0.049	0.051
	(0.77)	(0.45)	(0.63)	(0.57)	(0.56)	(0.58)
D(HRH) * Sentiment Direction	0.020	0.075	0.056	0.497**	0.507**	0.512**
	(0.11)	(0.42)	(0.31)	(2.50)	(2.51)	(2.55)
Past Informativeness 6 month 12FF	-0.078	-0.055	-0.145**	-0.043	-0.031	-0.098
	(-1.29)	(-0.90)	(-2.29)	(-0.48)	(-0.35)	(-1.11)
D(HRH)	0.112	0.083	0.064	0.002	-0.030	-0.037
	(0.83)	(0.61)	(0.46)	(0.01)	(-0.17)	(-0.21)
Sentiment Direction	0.155	0.048	0.075	-0.393**	-0.531***	-0.525***
	(0.90)	(0.27)	(0.42)	(-2.04)	(-2.71)	(-2.69)
Article category identified		1.262**	1.044**		0.336	0.422
		(2.14)	(2.07)		(0.76)	(0.94)
Time since last article		0.043***	0.044***		0.239***	0.240***
		(3.40)	(3.44)		(13.12)	(13.19)
Number of firms in article		-0.152***	-0.146***		-0.222***	-0.220***
		(-5.09)	(-4.93)		(-9.32)	(-9.26)
Number of Observations	321860	321860	321860	472827	472827	472827
R^2	0.064	0.066	0.070	0.058	0.063	0.064
Relevance, Category and Hour Fixed Effects	No	Yes	Yes	No	Yes	Yes
Date and Firm Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Market control variables	No	No	Yes	No	No	Yes
Firm specific control variables	Yes	Yes	Yes	Yes	Yes	Yes

Table 8: Difference in Difference Analysis

This table contains the results of article-level regressions implementing a difference in difference set-up for our whole sample from February 1, 2004 to September 10, 2012 as a robustness check to tables 4 to 6. In regressions 1 and 2, the dependent variable is *Speed of Stock Price Response* (in percent), defined as $\frac{\text{Abs(Return t}-1,t+5)}{\text{Abs(Return t}-1,t+5)+\text{Abs(Return t}+6,t+120)}$ and measured in seconds around an article. In regressions 3 and 4, the dependent variable is *Speed of Trade Volume Response* (in percent), defined as the turnover from 1 second before the article to 5 second after the article divided by the turnover from 1 second before the article to 120 seconds after the article. In regressions 1 to 4, the explanatory variable of interest is the interaction between *D(HRH)* and *RavenPack Release*. D(HRH) is a dummy variable equal to 1 if an article was consistently released as highly relevant in both RavenPack versions and 0 if it was originally released as having low relevance (HRL). *RavenPack Release* is a dummy variable taking the value of 1 for articles after RavenPack went "live" on April 1, 2009, and zero otherwise. In regressions 5 and 6, the dependent variable is the return (in percent) measured from 1 second before to 5 seconds after the article. The explanatory variable of interest is a triple interaction between *HRH*, *RavenPack Release* and *Sentiment Direction*, where *Sentiment Direction* is a variable indicating the sentiment of the article derived from RavenPack sentiment indices. It takes the value +1 for positive sentiment, 0 for neutral sentiment and −1 for negative sentiment. In all regressions we include firm and date fixed effects and the following firm specific control variables: Company size, Return prior month, Volatility prior month, Turnover prior month, Illiquidity prior month. In regressions 2, 4, and 6, we add fixed effects for the article category (e.g. mergers and acquisitions), the relevance score (from 90 to 100) and the hour during the day in which the article was released as well as additional controls: the absolute return, turnover, and volatility each for industry and market from t−1 to t+5 seconds around the article. In regression 2 and 4, we also include those values for t−1 to t+120 seconds around the article. All variables are defined in Appendix 1. All standard errors are clustered at the firm level. T-statistics are below the parameter estimates in parenthesis; ***, **, * indicate significance at the 1%, 5%, and 10% level, respectively.

Dependent Variable:	Speed of Stock Price Response		Speed of Trade Volume Response		Return t-1, t+5	
	(1)	(2)	(3)	(4)	(5)	(6)
RavenPack Release * D(HRH) * Sentiment Direction					0.336**	0.330**
					(2.11)	(2.06)
RavenPack Release * D(HRH)	1.702***	1.713***	0.501*	0.476*	0.101	0.109
	(3.71)	(3.81)	(1.71)	(1.66)	(0.73)	(0.77)
RavenPack Release * Sentiment Direction					-0.314**	-0.294*
					(-2.03)	(-1.89)
D(HRH) * Sentiment					0.072	0.114
					(0.70)	(1.09)
D(HRH)	-0.053	-0.070	0.082	0.044	0.095	0.025
	(-0.19)	(-0.25)	(0.57)	(0.31)	(0.95)	(0.24)
Sentiment Direction					0.430***	0.222**
					(4.33)	(2.20)
Absolute Composite Sentiment Score		-0.025**		-0.013**		
		(-2.38)		(-2.52)		
Neutral Composite Sentiment Score		-0.269***		-0.111***		
		(-3.43)		(-2.70)		
Absolute Event Sentiment Score		0.044***		0.035***		
		(4.45)		(6.24)		
Neutral Event Sentiment Score		-1.019***		-0.536***		
		(-3.04)		(-2.99)		
Article category identified		-1.660		-2.200**		0.580**
		(-0.36)		(-2.04)		(2.13)
Time since last article		0.049**		0.082***		0.165***
		(2.43)		(7.04)		(13.33)
Number of firms in article		-0.117**		-0.136***		-0.201***
		(-2.49)		(-6.15)		(-10.98)
Number of Observations	649368	649368	690467	690467	803799	803799
R^2	0.026	0.051	0.019	0.037	0.046	0.052
Relevance, Category and Hour Fixed Effects	No	Yes	No	Yes	No	Yes
Date and Firm Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes
Market control variables	No	Yes	No	Yes	No	Yes
Firm specific control variables	Yes	Yes	Yes	Yes	Yes	Yes

Appendix 1: Variable Definitions

This table displays the variable definitions for all variables used in the regressions. Article variables (sentiment scores, relevant scores, etc.) are based on RavenPack 3. When we winsorize, we set outliers to the allowed extreme value; e.g., "smaller 10" means that any value below 10 is set to 10. For all variables, winsorizing affects less than 1% of observations on either side.

Variable Name	Definition	Winsorizing
HRH	High relevance article Released as High relevance article. Dummy variable equal to 1 if an article has a relevance of 90 or higher in both RavenPack versions. When used in regressions, it is equal to 0 if an article has a relevance score of 90 or higher in the new RavenPack version (RavenPack 3), but was not covered or had a relevance score below 90 in the old RavenPack version. (the old RavenPack version is RavenPack 1 until July 6, 2011 and RavenPack 2 afterwards).	*None*
HRL	High relevance article Released as Low relevance article. Dummy variable equal to 1 if an article has a relevance score of 90 or higher in the new RavenPack version (RavenPack 3), but was not covered or had a relevance score below 90 in the old RavenPack version. (the old RavenPack version is RavenPack 1 until July 6, 2011 and RavenPack 2 afterwards).	*None*
LRH	Low relevance article Released as High relevance article. Dummy variable equal to 1 if an article has a relevance score below 90 or is not covered in the new RavenPack version (RavenPack 3), but had a Relevance Score greater or equal than 90 in the old RavenPack version. (the old RavenPack version is RavenPack 1 until July 6, 2011 and RavenPack 2 afterwards).	*None*
Company size	Log(prior day closing price * shares outstanding)	*Smaller 10*
Volatility prior month	Average of daily squared returns of the stock in the prior 20 trading days	*Larger 2%*
Turnover prior month	Average of daily volume divided by shares outstanding in the prior 20 trading days	*Larger 10%*
Return prior month	Average return in the prior 20 trading days	*Larger 3% & Smaller -3%*
Illiquidity prior month	Percentile rank of all article-firm combinations of a day according to Amihud Illiqudity $=$ $\text{mean}_{\text{over past 20 trading days}}\left(\frac{\lvert \text{ret}_{\text{daily}} \rvert}{\text{dollar volume}_{\text{daily}}}\right)$. *The most illiquid firms are assigned 100 the most liquid 1.*	*None*
Relevance	Score provided by RavenPack that indicates the relevance of an article to a company and takes values from 0 (least relevant) to 100 (most relevant).	
Event Sentiment Score	Sentiment score that is provided by RavenPack; takes a value from 100 (positive) to 0 (negative). It is available only for articles for which the category is identified.	*None*
Absolute Event Sentiment Score	Abs (*Event Sentiment Score* – 50)	*None*
Neutral Event Sentiment Score	Dummy variable equal to 1 if *Event Sentiment Score* equals 50 or if it is missing.	*None*
Composite Sentiment Score	Sentiment score that is provided by RavenPack; takes a value from 100 (positive) to 0 (negative). It is available for each article.	*None*
Absolute Composite Sentiment Score	Abs (*Composite Sentiment Score* – 50)	*None*
Neutral Composite Sentiment Score	Dummy variable equal to 1 if *Composite Sentiment Score* equals 50.	*None*
Article category identified	Dummy variable equal to 1 if the category (e.g. "merger") of the article is identified	*None*
Number of firms in article	Log (Number of firms in article)	*None*
Time since last article	Log (Time since last article in seconds)	*None*
Sentiment Direction	Variable indicating the sentiment of the article based on RavenPack sentiment indices. It can take the values 1 (positive sentiment), 0 (neutral sentiment) and −1 (negative sentiment). It is first based on *Event Sentiment Score* (*ESS*). If *ESS* is larger 50, this variable is 1, if *ESS* is smaller than 50, it is −1. If *ESS* is missing or 50, we consult *Composite Sentiment Score* (*CSS*). If *CSS* is greater than 50 we set this variable to 1, if *CSS* is smaller than 50 we set it to −1, if *CSS* equals 50 we set it to zero.	*None*
Return t-1, t+5	Stock return from 1 second before to 5 seconds after the article. Returns are computed from mid-quotes.	*Larger 2%, smaller -2%*
Return t+6, t+120	Stock return from 6 seconds after to 120 seconds after the article. Returns are computed from mid-quotes.	*Larger 2%, smaller -2%*

Speed of Stock Price Response	$\frac{Abs(Return\ t-1,t+5)}{Abs(Return\ t-1,t+5)+Abs(Return\ t+6,t+120)}$	*None*
Speed of Stock Price Response – Market Adjusted	$\frac{Abs(Market\ Adjusted\ Return\ t-1,t+5)}{Abs(Market\ Adjusted\ Return\ t-1,t+5)+Abs(Market\ Adjusted\ Return\ t+6,t+120)}$ Set to missing if: $Abs(Return\ t-1,t+5)+Abs(Return\ t+6,t+120)=0$.	*None*
Speed of Stock Price Response – Industry Adjusted	$\frac{Abs(Industry\ Adjusted\ Return\ t-1,t+5)}{Abs(Industry\ Adjusted\ Return\ t-1,t+5)+Abs(Industry\ Adjusted\ Return\ t+6,t+120)}$ Set to missing if: $Abs(Return\ t-1,t+5)+Abs(Return\ t+6,t+120)=0$.	*None*
Speed of Trade Volume Response	$\frac{Turnover\ t-1,t+5}{Turnover\ t-1,t+120}$	*None*
Speed of Directional Trade Volume Response	$\frac{Turnover\ in\ Direction\ of\ the\ Article\ t-1,t+5}{Turnover\ t-1,t+120}$, *where Turnover in Direction of the Article is buyer initiated turnover for articles with positive Sentiment Direction and seller initiated turnover for articles with negative Sentiment Direction.* The direction of a trade is determined using the Lee and Ready (1991) methodology, but using the quote at the end of the previous second as the prevailing quote rather than the quote 5 seconds ago.	*None*
Signed Return t-1, t+120	$Return_{t-1,t+120} * Sentiment\ Direction$ This variable is set to missing if Sentiment Direction is equal to zero.	*Larger 2%*
Past Informativeness 6 month 12FF	$Mean(Signed\ Return_{t-1,t+120})$, the mean is taken over the prior six calendar months within the same industry following 12 Fama French industry classification	*None*
Past Informativeness 3 month 12FF	Same definition as *Past Informativeness 6 month 12FF*, but using 3 month instead of 6 month.	*None*
Past Informativeness 6 month 30FF	Same definition as *Past Informativeness 6 month 12FF*, but using 30 Fama French industry classification instead of 12 Fama French industry classification.	*None*
Direction-Based Past Informativeness	Mean (D(Return and Sentiment agree), the mean is taken over the prior six calendar months within the same industry following 12 Fama French industry classification. D(Return and Sentiment agree) is a dummy variable equal to one if the sign of article sentiment and *Return t+6, t+120 agree* (sign can be -1, 0 or 1).	
Market return t-1, t+5	Value-weighted return of all common stocks in TAQ (which are also in CRSP) from 1 second before to 5 seconds after the article. Returns are computed from mid-quotes.	*None*
Industry return t-1, t+5	Value-weighted return of all common stocks in the same 12 Fama French Industry from 1 second before to 5 seconds after the article. Returns are computed from mid-quotes.	*None*
Market turnover t-1, t+5	Total dollar trading volume of all common stocks in TAQ (which are also in CRSP) from 1 second before to 5 seconds after the article divided by total market capitalization at t-2.	*None*
Market volatility t-1, t+5	Value weighted average squared second return of all common stocks in TAQ (which are also in CRSP) averaged from 1 second before to 5 seconds after the article.	*Larger 20 bp*
Market adjusted return t-1, t+5	*Return* (*t-1, t+5*) − *Market Return* (*t-1, t+5*)	*Larger 2%, smaller -2%*
Industry adjusted return t-1, t+5	*Return* (*t-1, t+5*) − *Industry Return* (*t-1, t+5*)	*Larger 2%, smaller -2%*