News Articles and the Invariance Hypothesis*

Albert S. Kyle
Robert H. Smith School of Business
University of Maryland
akyle@rhsmith.umd.edu

Anna A. Obizhaeva
Robert H. Smith School of Business
University of Maryland
obizhaeva@rhsmith.umd.edu

Nitish Ranjan Sinha
College of Business Administration
University of Illinois at Chicago
nrsinha@uic.edu

Tugkan Tuzun
Board of Governors of the Federal Reserve System
tugkan.tuzun@frb.gov

First Draft: August 5, 2010; This Draft: January 3, 2012

Abstract

Using a database of news articles from Thomson Reuters for 2003-2008, we investigate how the arrival rate of news articles mentioning an individual stock varies with the level of trading activity in that stock. Defining trading activity W as the product of dollar volume and volatility, we estimate that the arrival rate of news articles is proportional to $W^{0.68}$. Market microstructure invariance predicts that the stock trading process unfolds in "business time" which passes at a rate proportional to $W^{2/3}$. Since the estimated exponent of 0.68 is close to 2/3, we conclude that information in news articles flows into the market in the same units of business time that microstructure invariance predicts to govern the trading process for stocks. The arrival of news articles is well approximated by a negative binomial process with the over-dispersion parameter of 2.11.

*The views expressed herein are those of the authors and do not necessarily reflect the views of the Board of Governors or the staff of the Federal Reserve System.

1 Introduction

The market microstructure invariance hypothesis of Kyle and Obizhaeva (2011a) makes precise predictions about how business time governs the the trading process for individual stocks. In this paper, we examine whether the same business time also governs the arrival rate of information into the market for individual stocks. We use counts of news articles from Thomson Reuters during the period 2003-2008 to approximate the arrival rate of information. We thus generalize microstructure invariance from being a hypothesis about the trading process alone to being a hypothesis about both the trading process and the information process associated with trading. The empirical results about news articles in this paper—combined with the empirical results about portfolio transitions in Kyle and Obizhaeva (2011a, 2011b) and emprical results about TAQ data prints in Kyle, Obizhaeva and Tuzun (2011)—suggest that the same business time clock governs both the trading process and the information process for individual stocks.

According to the invariance hypothesis, traders participate in trading games which are the same across assets, except for the speed with which the games are played. The business time clock runs at a faster rate for active stocks than for inactive stocks. Defining W as the product of daily dollar volume and the percentage standard deviation of daily returns, the invariance hypothesis implies that the speed of the trading game is proportional to $W^{2/3}$. The exponent of precisely 2/3 follows from the invariance hypothesis that the risk transferred by a bet is constant per unit of business time (not calendar time).

When playing trading games, traders make trades based on the flow of information into the market. It is therefore natural to hypothesize that the rate of information flow is also proportional to $W^{2/3}$, or the rate at which business time passes. Invariance then implies that the number of bets per news article is constant across stocks, and the standard deviation of dollar gains and losses on a bet between the arrival of one news article and the next news article is also constant across stocks. We can imagine a world of trading in which traders bet on a flow of information approximated by a flow of news articles. Across stocks with different levels of trading activity and different rates of flow of information and news articles, microstructure invariance conjectures that a constant amount of money changes hands on average per news article. This addresses a fundamental question about the role of time in financial markets, discussed in the important work of Mandelbrot and Taylor (1967), Clark (1973), and Hasbrouck (1999).

Before stating the hypotheses and results in this paper, we provide a context by summarizing the empirical results from Kyle and Obizhaeva (2011b) and Kyle, Obizhaeva, and Tuzun (2011) concerning three hypotheses of market microstructure invariance about the trading process for stocks:

- **Trading Game Invariance:** Between each tick on the business time clock, the distribution of the risks transferred by a bet is the same across assets and across time. When trading activity W increases by one percent, the arrival rate of bets increases by 2/3 of one percent and the distribution of bet sizes shifts upwards by 1/3 of one percent.

- **Market Impact Invariance:** The expected market impact cost of a bet is the same across assets and across time. When trading activity W increases by one percent,

the expected market impact cost (per dollar traded in volatility units) incurred by executing a bet equal to a given fraction of average daily volume (say one percent) increases by 1/3 of one percent.

- **Bid-Ask Spread Invariance:** The expected bid-ask spread cost of a bet is the same across assets and across time. When trading activity W increases by one percent, the expected bid-ask spread cost (per dollar traded in volatility units) decreases by 1/3 of one percent.

To derive empirically testable hypotheses which generalize microstructure invariance from hypotheses about the trading process to hypotheses about the information process, we make the following two empirical conjectures.

- **Information Flow Invariance:** Both public and private information are expected to arrive at a rate proportional to the rate at which the business-time clock ticks, with a proportionality constant which is the same across assets and across time. When trading activity W increases by one percent, the flow of public and private information speeds up by 2/3 of one percent.

- **News Article Invariance:** News articles are expected to arrive at a rate proportional to the rate at which public information arrives, with a proportionality constant which is the same across assets and across time. When trading activity W increases by one percent, the number of news articles increases by 2/3 of one percent.

These empirical hypotheses are parallel to the hypotheses of trading game invariance, market impact invariance, and bid-ask spread invariance put forth in Kyle and Obizhaeva (2011a). The proportionality constants are examples of market microstructure invariants.

Similar to Kyle and Obizhaeva (2011a), we consider two alternative models: the model of invariant bet frequency and the model of invariant bet size. Since these models do not have a natural concept of a time clock, we make assumptions consistent with their general spirit. In the first model, we assume that the number of news articles about firms over a given period of time is expected to be the same across assets, regardless of trading activity. In the second model, we assume that the expected number of articles about firms over a given period of time is proportional to the number of bets placed by traders. According to these models, the number of articles is therefore either constant across assets or increases proportionately with the trading activity. The predictions of all three models are nested into one specification with different exponents: Letting μ denote the expected arrival rate of news articles per month, then $\mu \sim W^{\gamma}$, with $\gamma = 2/3$ for the invariance hypothesis and $\gamma = 0$ or $\gamma = 1$ for the two alternatives.

We test the models using news data provided by Thomson Reuters from the beginning of 2003 to the end of 2008. We implement several empirical tests based on log-linear regressions and count-data regressions with the arrival rate of news articles specified either as a Poisson or a negative binomial processes. The Poisson model assumes that the arrival rate is a constant proportional to W^{μ}. The negative binomial model assumes that the arrival rate is a random variable having a gamma distribution with mean W^{μ} and variance given by an "over-dispersion" parameter. Note that the Poisson model is a special case of the negative binomial model when data is not "over-dispersed." In the context of the the invariance

hypothesis, over-dispersion is consistent with the intuition that some stocks generate news not related to stock market trading as a multiplicative factor of news relevant for stock market trading.

For the entire sample period 2003-2008, the estimated exponent of 0.68 (with standard error 0.024) is close to the value of 2/3 predicted by the invariance hypothesis. Fixing the exponent at a level of 2/3, we calibrate a negative binomial model with expected arrival rate of μ news articles per month. Letting $\tilde{G}(\alpha)$ denote a Gamma random variable with mean of one and variance of α, we estimate

$$\mu(W) = 7.17 \cdot \Big(\frac{W}{W^*}\Big)^{2/3} \cdot \tilde{G},$$

where the variance of $\tilde{G}(\alpha)$ is given by $\alpha = 2.11$ (with standard error 0.238). The scaling constant $W^* = 40 \cdot 10^6 \cdot 0.02$ corresponds to the trading activity of a benchmark stock with price of \$40 per share, trading volume of one million shares per day, and volatility of 2% per day; this hypothetical benchmark stock would be at the bottom of S&P500. This calibration implies that there are on average 7.17 news articles per month for the benchmark stock. The formula shows how to extrapolate this estimate to assets with different levels of trading activity.

The estimated over-dispersion parameter $\alpha = 2.11$ is statistically different from $\alpha = 0$ corresponding to the Poisson model. The negative binomial model describes the data much better than the Poisson model. The negative binomial model allows the number of news articles in a month to vary for the three reasons: (1) the variation in the Poisson arrival rate associated with different levels of trading activity, as predicted by the invariance hypothesis, (2) an additional component of variation in the stochastic Poisson arrival rate associated with otherwise unmodeled features captured by the Gamma distribution, and (3) the random variation in the actual number of Poisson events for the given Poisson arrival rate determined by the particular level of the trading activity and the realization of a Gamma random variable. In our further tests, we find that the variation unexplained by the invariance hypothesis might be related to differences in market capitalization, book-to-market ratios, past returns, and the square value of trading activity.

Monthly estimates of parameters show that there is a structural break in the middle of 2005. Around this time, conversations with Thomson Reuters employees indicate that Thomson Reuters made changes in response to requests from its clients to broaden news coverage. These changes resulted in more news articles for smaller companies. The average number of news articles for the benchmark stock increased from 6.50 news articles per month in the first half of the sample to 8.20 news articles in the second half. The estimated exponent decreased from $\gamma = 0.78$ before 2005 to 0.61 after 2005. Although the estimate of $\gamma = 0.68$ for the entire sample period 2003-2008 is close to the value of $\gamma = 2/3$ predicted by invariance, there is substantial variation in γ during the period. An increased propensity to cover every firm in the sample could also explain why the over-dispersion parameter dropped from 2.96 in the first half of the sample to 1.39 in the second half.

In the database, news articles are tagged with topics, and one news article frequently carries tags for multiple topics. For example, if a news article talks both about the downgrade of a firm's debt and the worsened forecasts of its earnings, it has two tags. The most frequent tag categories are "regulations, additions and deletions from indices, new listings, delistings,"

"corporate results," "changes of ownership," "forecasting of corporate financial results," "major breaking news," and "corporate analysis." When we use the number of news tags instead of the number of news articles in our regressions, we obtain an estimated exponent of 0.71 (with standard error 0.025), which is only slightly higher than the predicted value of 2/3. This slight shift upwards in the estimates exists because news articles are usually tagged with at least two tags and so news tags tend to occur in pairs.

We also estimate invariance exponents γ for different categories of news tags. The estimated exponents range from 0.60 to 1.23. The lowest exponent is for the "corporate results" category and the highest for the "major breaking news" category. These results are not surprising. Small firms with low levels of trading activity receive a high percentage of their news from the company's announcements of "corporate results," a news category which includes corporate financial results, tabular and textual reports, dividends, accounts, and annual reports. In contrast, large firms with high levels of trading activity receive a disproportionate share of articles in the "major breaking news" category, which includes articles of interest to a wide audience These are news stories that are expected to appear in the financial and general headlines of the worlds major newspapers, web sites, television and radio networks.

Several papers have tested predictions of the invariance hypothesis for trading data. For example, Kyle and Obizhaeva (2011b) document evidence concerning the distribution of order sizes, price impact, and bid-ask spread using the sample of portfolio transitions. Kyle, Obizhaeva and Tuzun (2011) implement tests based on the transactions in the Trades and Quotes (TAQ) dataset. Our paper suggests that not only that the trading processes unfold in a business time, but that the information flow conforms to the same time clock. This finding validates the internal consistency of the invariance hypothesis.

Berry and Howe (1994) and Mitchell and Mulherin (1994) study the relationship between the number of news releases and market activity for the aggregate market. They suggest a small positive time-series relationship between public information and trading volume as well as an insignificant relationship between public information and price volatility. Our paper shows a strong cross-sectional relationship based on information flow for individual stocks rather than the aggregate market.

A growing body of literature has recently documented that measures of trading activity—such as volume, volatility and returns—are related to various news events. Examples include the analysis of the stock messages on internet boards in Antweiler and Frank (2004), economic news announcements in Green (2005), CEO interviews on CNBC in Mescke (2004), information in Wall Street Journal columns in Tetlock (2007), corporate announcements in Chae (2005), as well as data in the Dow Jones news archives in Chan (2003), Tetlock, Saar-Tsechansky, and Macskassy (2008), and Tetlock (2010). In contrast to the previous literature, we test a specific quantitative prediction about the relationship between the number of news articles and the trading activity.

The remainder of the paper states the implications of the invariance hypothesis for the flow of information in Section 2, describes the data in Section 3, explains the design and results of empirical tests in Section 4 and Section 5, and finally suggests several directions for the future research in Section 6.

2 Implications of Invariance For News Data

In the context of the invariance hypothesis, traders are thought as playing trading games. They arrive to the market and execute orders, with innovations in their order flow referred to as "bets." Trading volume is the sum of "long-term" bet volume and "short-term" non-bet volume which intermediates bets. Trading games are similar across assets and across time, except for the speed with which they are being played. Each security has its own business time clock that ticks at a rate proportional to the arrival rate of bets. Active securities have a fast time clock, while inactive securities have a slow time clock.

Trading activity and information flow are synchronized, speeding up and slowing down in tandem. We hypothesize that both public and private information arrive in a business time and refer to this hypothesis as "information flow invariance." The conjecture that the amount of public news is effectively proportional to the amount of private information may appear unlikely, but there are good reasons to believe so. First, news reporters may write articles about the same firms for which traders are starting to acquire private information. Second, private information may arise due to the manner in which public information is processed. For example, asset managers may generate private information after earnings announcements, if they have special skills for interpreting available public information. A formal validation of the conjecture is ultimately an empirical question.

Public information comes in many forms, including new articles, press digests, TV news, earnings announcements, firms' filings, and analysts' reports. In this paper, we put forth the hypothesis of the "news article invariance" that news articles arrive at a rate proportional to the rate at which public information arrives. Information flow invariance and news article invariance together imply that the expected rate of news articles arrival is proportional to the business time clock.

Suppose there are two stocks. The business-time clock runs H times faster for active stock than for inactive stock. There expected to be μ^* and μ news articles per calendar for active stock and inactive stock, respectively,

$$\mu = \mu^* \cdot H. \tag{1}$$

The business-time clock H is unobservable, because it is difficult to identify independent bets in trading data, but Kyle and Obizhaeva (2011a) show how to relate this unobservable time clock to the observable measure of trading activity, defined as the product of daily volume V , share price P, and daily volatility σ,

$$W = V \cdot P \cdot \sigma. \tag{2}$$

The product of daily volume and volatility captures the amount of risk transfer taking place in the market during a calendar day.

The correspondence between the speed of the business-time clock and the trading activity W is non-linear. Speeding up the time clock ($H > 1$) affects the trading activity from W^* to W in two ways. First, there is the "volume effect" - the number of bets per day and therefore the dollar volume increase proportionately with H. Second, there is the "volatility effect" - returns variance increases proportionately with H, but the volatility (the square root of

variance) increases proportionately with $H^{1/2}$. The combination of both effects implies a non-linear relation between trading activity and time clock,

$$W = W^* \cdot H^{2/3}. \tag{3}$$

Plugging (3) into (1), we obtain the relationship between the expected arrival rates of news articles μ and trading activity W,

$$\mu = \mu^* \cdot \Big(\frac{W}{W^*}\Big)^{2/3}. \tag{4}$$

A one percent increase in trading activity comes with a two-thirds of one percent increase in the expected arrival rate of news articles. Equation (4) is the main relationship that we test in this paper.

As an illustrative example, imagine doubling the speed of the time clock ($H = 2$). The information flow speeds up: The analysts type twice faster their reports, the journalists publish twice more articles, the news service providers release twice more news items, and twice more news messages appear on the screens of traders. The same amount of information that used to arrive during a day now comes in half a day. The number of news articles released per day μ goes up by a factor of 2. The dollar volume goes up by a factor of 2, since investors trade twice as many shares each day. The variance doubles, or equivalently, the standard deviation increases by $2^{1/2}$. The trading activity increases by a factor of $2^{3/2}$. The changes in both trading activity and news articles arrival rate are consistent with equation (4).

Alternative Models. Kyle and Obizhaeva (2011a) consider two alternative models. Since these models do not have a well-defined concept of time, we suggest conjectures about information flow which are consistent with their general spirit.

The model of invariant bet frequency assumes that variation in trading activity comes entirely from variation in bet sizes, while the number of bets per day remains invariant across stocks. In a spirit of this model, we assume that the number of news articles per day is constant across stocks. Each news article leads to the same number of bets, but bets are larger for more active stocks, since the articles about these stocks have more valuable information, thus allowing traders who read them to place larger bets. The conjecture implies a testable prediction that the expected number of news articles μ does not vary with trading activity W,

$$\mu = \mu^* \cdot \Big(\frac{W}{W^*}\Big)^{0}. \tag{5}$$

The model of invariant bet size assumes that variation in trading activity comes entirely from variation in the number of bets placed per day, while the distribution of bet sizes over a calendar day remains the same across stocks. In a spirit of this model, we assume that the number of news articles varies across stocks proportionally to the number of bets. Each news article leads to a certain number of bets, similar in size. This assumption implies a testable prediction that the expected number of news articles μ is proportional to trading activity W,

$$\mu = \mu^* \cdot \Big(\frac{W}{W^*}\Big)^{1}. \tag{6}$$

These conjectures are ultimately related to our assumptions about how information is being processed institutionally. They are chosen in a somewhat ad hoc way and may be potentially replaced by other assumptions. In the context of the first model, for example, we have conjectured that stocks have the same number of news articles and bets per day. We could, however, have a situation when more active stocks have more news articles than inactive stocks and yet the same number of bets executed. Indeed, active stocks may be traded by large financial institutions with several people in different departments analyzing news articles about different segments of the market. If their decision-making processes are internalized inside the firm, then their collective efforts may lead to only one trading decision, just as one person can read a news article about a small stock and decide to make one trade.

In the context of the second model, we have conjectured that the number of news articles vary across stocks proportionally to the number of bets. But this model is also consistent with a situation when stocks have the same number of news articles published about them per day, but yet have different number of bets executed. For instance, active stocks may be followed by many traders, who often disagree with each other about how to interpret a news article and therefore place multiple independent bets reflecting their own views upon reading a single article.

All three models imply a specific relation between the expected number of news articles μ on left-hand side and the measure of trading activity W on the right-hand side, which can be nested into one specification,

$$\mu = \mu^* \cdot \Big(\frac{W}{W^*}\Big)^{\gamma}. \tag{7}$$

The three models differ only in their predictions about γ. The invariance hypothesis predicts that $\gamma = 2/3$, the model of invariant bet frequency predicts that $\gamma = 0$, and the model of invariant bet size predicts that $\gamma = 1$. Although our paper examines the extension of the invariance hypothesis to the news data rather than trading data, we chose to keep the original names of the three models as in Kyle and Obizhaeva (2011a), for simplicity of exposition.

3 Data

Thomson Reuters firm provided the news data from NewsScope dataset described in detail in Sinha (2011). The sample covers all news articles sent by the news service provider to its clients from January 2003 through December 2008. During the evaluation period, the data has been collected by the Reuters group. In 2008, the Reuters group and Thomson corporation have merged to form Thomson Reuters. We use the number of news articles shown on the screens of traders as a proxy for the arrival rate of public information.

Each news items has the following fields: the time stamp, the ticker of a company, the relevance indicator that measures how substantive the news item is for the company, the sentiment indicator that shows a prevailing tone of the news item, the probabilities of the news item having positive, negative, or neutral tone that provide a more granular sentiment, the news item type (alert, article, update, or correction), the headline indicator, the linked counts that show how many times this news has been mentioned in the past, and the topic code that describes the news item. The news dataset is matched with daily returns, prices, and daily volume from the CRSP data for common stocks listed on the NYSE, the Amex,

and the NASDAQ exchanges.

We apply several filters to identify new information. We omit all one-line alert messages, which are usually sent out by Thomson Reuters before important news articles appear in full. We exclude updates and corrections, since they usually do not contain new information, but rather provide more detail about original articles. We also exclude news items linked to more than one article in the sample, to make sure that this information did not appear in the sample before.

News items can mention multiple firms. If a news item is associated with several firms, this news story can be often irrelevant for some of them. Indeed, large companies are often mentioned as placeholders in news articles about small companies, just in a context of a general description of an industry in which both companies operate. For example, a story about a small technology firm can often mention other technology heavyweights like Intel, Apple, and Microsoft, but the news story does not have any new information about these companies. Thomson Reuters assigns a relevance parameter associated to each pair of a news item and a firm. The relevance parameter ranges from zero to one. This parameter is equal to one, if the news item is highly relevant for a particular firm, and zero otherwise. We include only those news items whose relevance parameter for a given firm is greater than 0.35. This threshold does not affect our results.

News stories may have information on the multiple dimensions of a firm. These stories are then tagged by Thomson Reuters with several topic codes. If we count these news items only once, we can potentially underestimate the amount of actual information. We therefore chose to consider two samples. In the first sample, we count each news item once. In the second sample, we count each news item as many times as it has been tagged by Thomson Reuters. For example, if the news item mentions an earnings announcement, a earnings forecast, and a merger announcement, it will be tagged by Thomson Reuters with three tags. This news item will be counted as one observation in the first sample and as three observations in the second sample.

Table 2 lists all topic codes with a brief descriptions and the proportion of news articles being tagged with a particular topic code. The three most commonly used topic codes are 'STX', 'RES', and 'MRG'. The topic code 'STX' indicates additions and deletions from stock indices, new listings, delistings and suspensions; it has been assigned to 15% of news articles in our sample. The topic code 'RES' indicates all corporate financial results, tabular and textual reports, dividends, annual and quarterly reports; it has been assigned to 14% of news articles in our sample. The topic code 'MRG' indicates mergers and acquisitions; it has been assigned to 12% of news articles. Most of remaining topic codes indicate economic news. For example, the topic code 'DBT' indicates news articles related to debt market, 'RESF' indicates news indicates results of corporate financial results, 'CORA' and 'RCH' indicate analysis of a company by a journalist and a broker, respectively. Other topic tags indicate behavior news. For example, the topic code 'HOT' indicates news articles about stocks that are on move, and the topic code 'NEWS' indicates news articles that are likely to lead to television or radio bulletins or make the front page of major international newspapers and web-sites. We focus on the firm-specific news articles and exclude news tags about an industry. We also exclude news tags about firms that could not match to any ticker symbol in the CRSP dataset.

We consider two samples. The first sample is the sample of "Thomson Reuters firms,"

which consists of firms covered by Thomson Reuters from the instance we observe the first news article about a given firm. If the firm does not have any news articles in a given month, then we count the number of news articles and news tags in that month as being equal to zero. Of course, the Thomson Reuters's decision to cover particular firms is endogenous. The small firms with a few news articles can be easily left out of the sample, and the rest of small stock covered by Thomson Reuters will appear to have "too many" news articles. To deal with a selection bias, we also implement our tests on the other sample. This sample is the sample of "CRSP firms," which includes all firms recorded in the CRSP from 2003 through 2008, with zero news items assigned for firms not covered by the Thomson Reuters firm.

In total, there are about 1.4 million news articles and about 3.4 million news tags in the database. These observations are spread over 72 months. The coverage has increased over time and converged to almost 100% by year 2006, as the news provider has responded to requests of its clients who demanded a broader coverage. As a result, most of our data is weighed more towards the later periods. The average number of firms in a given month is 3,820, ranging from 2,586 to 4,468 in both of our samples. There are 275,059 firm-month observations in the sample of Thomson Reuters firms, resulting from at least one match between a firm and a news article. There are 340,505 firm-month observations in the sample of all firms in the CRSP.

Descriptive Statistics. Table 1 provides a descriptive statistics for stocks in our sample. Statistics are calculated for all securities in aggregate as well as separately for the ten volume groups of stocks sorted by average dollar volume. Instead of dividing the securities into ten deciles with the same number of securities, the volume break points are set at the $30^{th}, 50^{th}, 60^{th}, 70^{th}, 75^{th}, 80^{th}, 85^{th}, 90^{th}$ and 95^{th} percentiles of trading volume for the universe of stocks listed in the NYSE with CRSP share codes of 10 and 11. Group 1 contains stocks in the bottom 30^{th} percentile. Group 10 contains stocks in the top 5^{th} percentile and approximately corresponds to the universe of S&P100. Smaller percentiles for the more active stocks make it possible to focus on the stocks which are economically more important. For each month, the thresholds are recalculated and stocks are reshuffled across groups.

Panel A of Table 1 reports the statistical properties of securities in our sample. The average daily volume is \$22 million, ranging from \$1 million for low-volume stocks to \$466 million for high-volume stocks. The average volatility of daily returns is equal to 3.10%, ranging from 3.30% for low-volume stocks to 2.30% for high-volume stocks. These numbers imply that trading activity—a product of dollar volume and volatility—varies by a factor of 315 between inactive stocks in group 1 and active stocks in group 10.

Panel B of Table 1 reports the statistics for the number of news articles in the Thomson-Reuters dataset. The average number of news articles per month varies from 0.58 news articles for low-volume stocks to 83 news articles for high-volume stocks. The median ranges from 0 to 46 news articles. The actual variation in the average number of news articles is bigger than predicted by the invariance hypothesis, according to which there should be only 46 times ($= 315^{2/3}$) fewer news articles for low-volume stocks than for high-volume stocks. As we discuss below, this may be attributed to the convexity in the news data.

For each volume group, the minimum number of news articles per month is zero, whereas

its maximum values vary from 143 to 3,344 news articles across volume groups. The significant variation reveals that releases of news articles about a given firm tend to cluster in time. Inactive stocks get no attention during most months, but when something happens - for example, a small firm is acquired by a large firm after developing a successful product - there will be a disproportionately large number of news articles released. Our estimation procedures will have to be adjusted for an excessive variation in the news arrival rates due to the news clustering.

Similar conclusions can be drawn from the statistics on the fraction of firms with no news articles during a given month. For the aggregate sample, about 58% of firms have no news articles in a given month. For high-volume stocks, only 5% of firms do not have any news articles during a given month (357 out of 7,143 pairs); about 2.70% of firm-month pairs are not covered by Thomson-Reuters at all (196 out of 7,143 pairs), and the other 2.30% of firms have no news articles reported by the news provider (161 out of 7,143 pairs). For low-volume stocks, 73% of firms do not have any news articles during a given month (162,456 out of 222,543); about 25% of firm-month pairs are not covered by Thomson-Reuters at all (55,864 out of 222,543 pairs), and 48% of firms have no news articles reported although they are in Thomson-Reuters sample (106,592 out of 222,543 pairs).

The data clearly has too many zeros and exhibits over-dispersion relative to a Poisson model. If a Poisson model were a correct model, then the fraction of firms with no news would be equal to $e^{-\mu}$, where μ is the average number of news articles per month in the table. Given the average arrival rate of 0.58 news articles per month for inactive stocks, we can infer that the fraction of low-volume stocks with no news articles would be 51% ($= e^{-0.58}$). Given the average arrival rate of 82.86 news articles per month for active stocks, we can infer that the fraction of high-volume stocks with no news articles would be 0% ($= e^{-82.86}$). Comparing these implied numbers of 51% and 0% with the actual numbers of 73% and 5%, we conclude that the data has "excess zeros," whose existence has important implications for model selection. It suggests that a negative binomial model, which allows to correct for over-dispersion, could be a better choice than a Poisson model.

Each news articles can be tagged with several news topics. In the table, statistics for the mean and maximum of news tags is about twice bigger than those for the news articles. This implies that one news articles is usually tagged to two news topics. The number of observations with only one news tag per month is very small, since usually there is either no news articles about a given firm at all or there is one news articles with two news tags attached. As a result, even though the arrival of news articles may be closely approximated by a Poisson model or a negative binomial model, the number of news tags will have to be described by a more complicated distribution.

Empirical Distributions of The Number of News. Figure 1 shows the distribution of the number of news articles per month for different volume groups across the news bins. The figure has three panels. The first panel shows the distribution for stocks in volume group 1. The second panel shows the distribution for stocks in volume groups 2 through 8. The third panel shows the distribution for stocks in volume groups 9 through 10. On each panel, observations are split into the twelve bin with $0, 1, 2, 3-4, 5-8, 9-16, 17-32, 33-64, 65-128, 129-256, 257-512, 513-1024$ news items per month, respectively. Except for the first

bins, most bins are such that their upper cutoff has the form of 2^i news items per month. These bins have finer grid on the left allowing to zoom in into a crucial area of densities for cases when no news events or only a few news events occur per month. The distributions are constructed based on the number of news articles per months (in dark blue) and based on the number of news tags per month (in light blue). Observations are pooled together across time and across stocks.

The figure shows three subplots for the three sub-samples: inactive stocks from the volume group one, medium stocks from the volume groups two through eight, and active stocks from the volume groups nine and ten. For inactive stocks in the lowest volume group, 73% of stocks have no news articles, 17% of stocks are mentioned in one article, and 6% of stocks are mentioned in two articles. For active stocks in the two highest volume groups, 6% of stocks have no news articles, 1% of stocks are mentioned in one news articles, 1% of stocks are mentioned in two articles, and the remaining observations are spread over higher news bins, with the biggest density in the news bin "seven" implying that actively traded stocks are typically mentioned in 17 to 32 news articles per month.

The density of news tags in light blue is shifted slightly to the right relative to the density of news articles in dark blue, since one article is tagged with at least one news tag. By definition, the densities for news articles and news tags are identical in the first no-news bin.

We examine next whether the invariance hypothesis can explain the cross-sectional differences in the distribution of the number of news articles and news tags, shown in the figure.

4 Estimation Procedures

For each stock i and month t, we observe the trading activity $W_{t,i}$ and the number of news items $N_{t,i}$. The trading activity $W_{t,i}$ is the product of average daily dollar volume and volatility calculated using the CRSP data. The number of news items $N_{t,i}$ is a count variable calculated using the news data; it is either the number of news articles or the number of news tags. We next implement the three estimation approaches to test (7): a log-linear model, a Poisson model, and a negative binomial model.

Log-linear model for averages. The simplest approach is to estimate a log-linear model for the average number of news items per month with trading activity being an explanatory variable. The main problem is that the number of news items is often equal to zero, since many firms do not generate any news. This makes the logarithm of the number of news being infinite. To avoid taking the logarithm of zero, we aggregate the data and work with the averages. Each month, we sort all stocks based on their trading activity into 30 groups such that each group has the same number of news items. We then calculate the average number of news items $\bar{N}_{t,j}$ and the average trading activity $\bar{W}_{t,j}$ in each group j. By construction, neither of these two numbers is zero. Finally, we regress the logarithm of the average number of news items $\bar{N}^*_{t,j}$, adjusted for the within-group variation in trading

activity, on the logarithm of the average trading activity $\bar{W}_{t,j}$ in each group j and month t,

$$\ln \bar{N}^*_{t,j} = \eta + \gamma \cdot \ln \big[\frac{\bar{W}_{t,j}}{W^*}\big] + \epsilon_{t,j}, \tag{8}$$

where a constant term $\eta = \ln \mu^*$ and the scaling constant $W^* = (40)(10^6)(0.02)$ corresponds to the trading activity of the benchmark stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. This hypothetical stock would be at the bottom of S&P500. We rescale the explanatory variable so that a constant term e^η quantifies the average number of news items reported per month about the benchmark stock.

For the log-linear specification, we need to make additional adjustment of the news items $\bar{N}_{t,j}$ for the within-group variation in trading activity. Suppose that the number of news items $N_{t,i}$ is modeled as,

$$N_{t,i} = e^{\eta + \gamma \ln[W_{t,i}/W^*]} \cdot \tilde{Z}_{t,i},$$

where $\tilde{Z}_{t,i}$ is a random variable with the mean equal to one; if its variance is equal to zero then it is a constant equal to one. The average number of news items in each group j with $M_{t,j}$ observations is a random variable $\bar{N}_{t,j}$,

$$\bar{N}_{t,j} = \frac{1}{M_{t,j}} \sum_{i=1}^{M_{t,j}} N_{t,i}.$$

Denoting $\bar{W}_{t,j} = 1/M_{t,j} \sum_{i=1}^{M_{t,j}} W_{t,i}$, we can write the average number of news item as follows,

$$E\bar{N}_{t,j} = e^{\eta + \gamma \ln \left[\frac{\bar{W}_{t,j}}{W^*}\right]} \cdot \frac{1}{M_{t,j}} \sum_{i=1}^{M_{t,j}} e^{\gamma(\ln W_{t,i} - \ln \bar{W}_{t,j})}.$$

$$\ln E\bar{N}_{t,j} = \eta + \gamma \cdot \ln \big[\frac{\bar{W}_{t,j}}{W^*}\big] + \ln \Big(\frac{1}{M_{t,j}} \sum_{i=1}^{M_{t,j}} e^{\gamma \cdot (\ln W_{t,i} - \ln \bar{W}_{t,j})}\Big).$$

The last equation suggests that we can not simply regress $\ln E\bar{N}_{t,j}$ on $\ln \bar{W}_{t,j}$ to obtain the estimate of γ, rather we need to adjust the average number of news items for the potential within-group variation in the trading activity, reflected in the last term. The adjustment term is always positive and potentially more significant for groups with lower trading activity, where variation in trading activity is more significant. The omitted adjustment term can introduce the systematic bias into our estimates. To avoid this bias, we calculate the adjusted average number of news $\bar{N}^*_{t,j}$ for group j and month t as,

$$\ln \bar{N}^*_{t,j} = \ln \bar{N}_{t,j} - \ln \Big(\frac{1}{M_{t,j}} \sum_{i=1}^{M_{t,j}} e^{2/3(\ln W_{t,i} - \ln \bar{W}_{t,j})}\Big), \tag{9}$$

assuming that $\gamma = 2/3$ in the adjustment term. We then regress this variable on the logarithm of the average trading activity $\bar{W}_{t,j}$ in group j and month t in the log-normal specification of our tests. Note that it is not necessary to implement this adjustment for the count data regressions, for which we use the actual news data, rather than the averages.

Poisson model. A better way to model count data is to assume a Poisson model for the number of news items. The Poisson model ensures that the left-hand side variable is always positive and allows to deal graciously with zeros. It implies that the distribution of the number of news items $N_{t,i}$ about stock i in month t has the following density function,

$$f(N_{t,i}|W_{t,i}) = \frac{e^{-\mu(W_{t,i})} \cdot \mu(W_{t,i})^{N_{t,i}}}{N_{t,i}!}, \tag{10}$$

where the expected number of news items $\mu_{t,i}$ per month is a non-linear function of trading activity $W_{t,i}$,

$$\mu(W_{t,i}) = e^{\eta + \gamma \cdot \ln\left[\frac{\bar{W}_{t,j}}{W^*}\right]}. \tag{11}$$

A constant term e^{η} quantifies the average number of news items reported per month about the benchmark stock.

The Poisson model assumes that the expected arrival rate is a non-stochastic function of the trading activity, i.e., all variation in arrival rates occurs only within the context of the Poisson distribution. From the properties of that distribution, we know that $\mu(W_{t,i}) = \mathrm{E}(N_{t,i}|W_{t,i}) = \mathrm{V}(N_{t,i}|W_{t,i})$. The Poisson model assumes that stocks with the same level of trading activity have the same expected number of news items and the same variance equal to $\mu(W)$. As discussed earlier, the descriptive statistics suggest that these assumptions may be too restrictive, because the news data exhibit over-dispersion, with the variance of the information flow being greater than its mean.

Negative binomial model. A negative binomial model allows the Poisson arrival rate to vary randomly, even for firms with the same level of trading activity. To model the additional variation, we use a continuous mixture of the Poisson distributions where the mixing distribution is modeled as the Gamma distribution,

$$\mu(W_{t,i}) = e^{\eta + \gamma \cdot \ln\left[\frac{\bar{W}_{t,j}}{W^*}\right]} \cdot \tilde{G}_{t,i}(\alpha). \tag{12}$$

The Gamma variable $\tilde{G}_{t,i}$ has the mean of $\kappa \cdot \theta$ and the variance of $\kappa \cdot \theta^2$. We impose the restrictions $\kappa = 1/\alpha$ and $\theta = \alpha$ to restrict the mean of the Gamma variable to be equal to one. Its variance then is equal to α ($\alpha = k \cdot \theta^2 = \theta$). The model parameter η then identifies the same mean as the mean in the Poisson model. The mixture does not affect the mean, but it affects the variance and other moments.

The negative binomial model nests the Poisson model as a special case when $\alpha = 0$. For a given mean, the negative binomial model allows the variance of the number of news items to be greater than the variance implied by the Poisson model. Higher values of parameter α indicate a more dispersed distribution of the arrival rates. If firms with similar levels of trading activity indeed have dramatically different numbers of news items per month, i.e., they vary across stocks too much to be explained by a simple Poisson model, then the negative binomial specification is a more reasonable model for describing the news data.

The negative binomial specification allows the number of news items in a month to vary for the three reasons: (1) the variation in the Poisson arrival rate associated with different levels of trading activity, (2) an additional component of variation in the stochastic

Poisson arrival rate associated with otherwise unmodeled features captured by the Gamma distribution, and (3) the random variation in the actual number of Poisson events given the Poisson arrival rate that is determined by the particular level of the trading activity and the particular realization of a Gamma random variable. For negative binomial specification, the Poisson arrival rate then varies randomly according to the realization from the Gamma distribution, even if two firms have with the same trading activity. Restricting the over-dispersion parameter α to be equal to zero, we obtain the Poisson specification that does not allow for the second source of uncertainty: The Poisson arrival rate is a non-stochastic function of the trading activity. Note that the log-linear model with data in bins also does not provide a statistical explanation of why, given two firms with similar levels of trading activity, one firm might have many news items in a given month and the other firm might have no news items in the same month.

We implement the empirical tests of the three models by estimating a coefficient γ and testing whether $\gamma = 2/3$ as predicted by the invariance hypothesis, $\gamma = 0$ as predicted by the model of invariant bet frequency, or $\gamma = 1$ as predicted by the model of invariant bet size.

The data might have a complex covariance structure of residuals. For each firm, the observations can be correlated across time; for example, a firm approaching the bankruptcy usually generates a large number of news articles over an extended period of time. Also, the observations for different firms can be correlated within each month; for example, unusually large number of news articles was released during the volatile months in the fall of year 2008. In negative binomial model, both the randomness in the Poisson arrival rates as well as the randomness in the mixing Gamma random variables might be interrelated. To adjust for these interdependencies, we implement the Fama-MacBeth procedure by estimating our models using the OLS regressions or the maximum-likelihood procedures for each of 72 months and then averaging the estimates across months. We also correct the standard errors using the Newey-West procedure with the three lags. Since this approach does not require specifying a particular form of interdependencies between residuals, it is a reasonable estimation strategy.

5 Results

We discuss next the results of our tests, starting with the estimation results for a log-linear specification and then reporting those for count-data models.

5.1 Log-Linear Models For Averages

Each month, we sort all stocks into 30 equally-sized groups in ascending order of their trading activity, from stocks with the lowest trading activity in the first group to stocks with the highest trading activity in the last group. Figure 2 shows the logarithms of the adjusted average number of news articles about firms per month, $\ln \bar{N}^*_{t,j}$, for a given group on the vertical axis and the logarithm of the average trading activity, $\ln \bar{W}_{t,j}$, on the horizontal axis, for each group j and month t. The six subplots contain observations for each of six

years from year 2003 through year 2008. Each subplot has 30×12 points. For each month, there are 30 points for each of 30 groups. For each group, there are twelve points for each of the twelve months in a year.

For the convenience, we superimpose the same fitted line with a slope fixed at 2/3 and an intercept of -6.74 is superimposed on each plot. We choose the slope to satisfy the invariance hypothesis and the intercept to be equal to the average number of news articles from the pooled sample. According to the invariance hypothesis, all observations are expected to be close to the fitted line.

The observations from the lowest group form a distinctive set of twelve points in the left tail in each of six subplots. As the trading activity increases from the first group to the last group, the monthly observations from the same group start to form tighter clouds of points. These patterns are consistent with our intuition that the within-group variation in the trading activity is the biggest for the first group and then decreases gradually when moving to groups with higher trading activity.

The scatter plots shows that the data exhibit patterns similar to those predicted by the invariance hypothesis. The observations pile up around the fitted line. The graph also has a visible "smile" indicating some convexity in the relationship between trading activity and the number of news articles. In comparison with the fitted line, the bins with very active and very inactive stocks have "too many" news articles, and the stocks in the middle have "too few" news articles. "Too many" news articles for inactive stocks may be due to the policy of the Thomson Reuters to expand its coverage and cover all firms in the economy, even though some smaller companies may have not much of actual new information about them. The goal of global coverage became especially important after year 2005, and indeed, the observations in the left tail are closer to the fitted lines in year 2003 and year 2004 than during subsequent years. "Too many" news articles for active stocks may be explained by a large number of news article simply referring to that stocks as "hot stocks," rather than carrying new information. The "smile" suggests that the explanatory power of the log-linear specification may be improved by adding a quadratic term.

Table 3 shows the estimates of the intercept η and the slope γ from the log-linear regression model (8) for the averages. We report the estimates based on the sample of all CRSP firms and the sample of firms in the Thomson Reuters universe. For each of the two samples, we provide estimates based on the number of news articles and the number of news tags. In total, the table contains four columns with four different sets of estimates.

The estimates of γ range from 0.65 to 0.75 across four columns. These estimates are economically close to 2/3 predicted by the invariance hypothesis and very different from 0 and 1 predicted by the alternative models. The F-tests for the hypothesis $\gamma = 2/3$ range from 0.03 to 0.79, indicating that the invariance hypothesis can not be rejected. At the same time, the F-tests strongly reject both alternative models. For the news articles, the estimates of η are 2.32 and 2.41 for the sample of all stocks and stocks covered by Thomson Reuters, respectively. The first estimate is lower than the second one, because the first sample differ from the second one by a set of firms with no news articles reported. For the number of news tags, the estimates of η are equal to 3.02 and 3.12, respectively. The value 3.02 and 3.12 are higher than 2.32 and 2.41, because there are more news tags than news articles, by definition.

The average R-squares range from 0.893 to 0.917. While relatively large R-squares indi-

cate that the log-linear specification explains most of the variation in the average number of news items across thirty bins, the "smile" in figure 2 suggests that much of the unexplained variance could have been captured by including a quadratic term into our log-linear specification.

5.2 Count-Data Models

Table 4 reports the estimates of the intercept η and the slope γ for the count-data specifications in (11) and (12). These specifications are more reasonable for modeling news events. The estimates are reported for the number of news articles and the number of news tags about firms from the two samples, the sample of all CRSP firms and the sample of Thomson Reuters firms. Our main sample is the sample of all CRSP firms, which is not affected by endogenous decisions of Thomson Reuters to cover particular firms.

The first two columns show the estimates based on the number of news articles for all CRSP firms, using the Poisson and negative binomial specifications. There are five facts to note about the estimates. First, the estimate 0.68 of γ for the negative binomial model is similar to the estimate of 0.70 for the log-linear model in table 3. This suggests that the log-linear model for the averages in the thirty bins provides a good way to model how the number of news items vary with trading activity. Second, the Newey-West standard error of 0.024 for the estimate of γ in the negative binomial model is sufficiently large so that the hypothesis $\gamma = 2/3$ can not be rejected; it is, however, sufficiently small so that the hypotheses of the alternative models, $\gamma = 0$ and $\gamma = 1$, are soundly rejected. Third, the estimate of γ for the negative binomial model is smaller than the estimate 0.81 of γ for the Poisson model; this indicates that the Poisson model produces the biased estimates of the average number of news events, since their arrival rates are likely to be non-constant in the data. Fourth, the estimates 2.11 and 2.01 of η are smaller than its estimate 2.32 in the log-linear model, since the log-linear model inflates the importance of close-to-zero observations whereas the count-data models properly account for the existing zeros. Fifth, the estimate 2.05 of the over-dispersion parameter α is statistically different from zero given its very small standard error of 0.218, thus indicating a strong statistical support for the negative binomial model over the Poisson model.

The third and fourth columns show the estimates based on the number of news articles for the subset of Thomson Reuters firms. The estimates are similar to those in columns one and two. The estimate of γ for the negative binomial model is equal to 0.65, being lower than its estimate of 0.86 for the Poisson model. The standard errors are such that the invariance hypothesis is not rejected, but the alternative models are rejected again. The estimates 2.19 and 2.08 of η are slightly bigger than 2.11 and 2.01 for the sample of all firms, since the difference between two samples is the subset of firms with no news articles. The estimate 1.63 of α (standard error 0.120) suggests that the data is too over-dispersed relative to the Poisson model.

The last four columns in table 4 show the estimates for the number of news tags. The four estimates of γ are equal to 0.86, 0.71, 0.84, and 0.66 for the number of news tags, being somewhat higher than the corresponding estimates 0.81, 0.68, 0.78, 0.65 for the number of news articles, but still close to 2/3 predicted by the invariance hypothesis. The estimated intercepts of 2.78, 2.66, 2.85 and 2.73 for the number of news tags are bigger than the

estimated intercepts of 2.11, 2.01, 2.19 and 2.08 by about 0.65; this implies that there are usually twice ($\approx \exp(0.65)$) more news tags than news articles, i.e., each article is usually tagged with two news topic codes.

The true model for the number of news tags is likely to be more complicated than the true model for the number of news articles. For example, consider firms for which the Poisson arrival rate is very low. Such firms usually have no news articles during a month, but occasionally they receive one news article and only rarely more articles. If it is common for news articles to have more than one topic code assigned, then the negative binomial model with news items counted by topic codes will be trying to fit to a Poisson distribution a different distribution with too many cases of two or more news events and not enough cases of one news event.

The Poisson and Negative Binomial Specifications. Figure 3 shows the residuals between the empirical distributions of the news arrival rates in figure 1 and the fitted count-data models calibrated using the estimates from table 4. The differences in the densities are plotted across the twelve bins. These bins are defined as containing observations with $0, 1, 2, 3-4, 5-8, 9-16, 17-32, 33-64, 65-128, 129-256, 257-512, 513-1024$ news items per month, respectively; except for the first bins, most bins have an upper cutoff of the form 2^i.

There are six plots organized into two rows and three columns. The three columns separate results for the three sub-samples: inactive stocks from the volume group one, medium stocks from the volume groups two through eight, and active stocks from the volume groups nine and ten. The top panels show the difference between the empirical frequencies and the fitted frequencies from the negative binomial model. The bottom panels show the same differences but for the Poisson model. The differences in distributions of the number of news articles are in dark color, and the differences in distributions of the number of news tags are in light color. In some sense, we show residuals from the estimated models, but further transformed into densities for convenience. The positive (negative) bar in a bin means that the data have more (fewer) observations in that bin relative to what the calibrated model predicts. The closer these bars are to zero, the better models fit the historical data. A "true" model would have all residuals equal to zero.

The top panels show the differences between the empirical densities and the fitted frequencies from the negative binomial model for the number of news articles (in dark blue). These differences are very small across all bins and stocks.

For active stocks, the largest deviations of -9.90% and -5.70% are in the first two bins, with the large negative values persisting till the sixth bin containing the density for observations of $5-8$ news articles per month, after which they turn positive. The negative binomial distribution thus overestimates how often less than eight news articles are published per month for active stocks and overestimates the frequency of observations with more than eight articles.

The plots show that the model fits the news data more accurately for inactive stocks than for active stocks. For inactive stocks, the largest deviations of -2.20% and 2.40% are in the bins with no news articles and with one news articles per month. Other residuals are even smaller. For medium stocks, the largest deviations of -1.50% and 1.50% are in the bins with one news article and two news articles per month. Intuitively, our estimation

procedures tend to fit the models so that they match observations for groups of inactive stocks with larger number of observations.

Note that the residuals of -2.20%, -1.20% and -9.90% in the no-news bins are negative in all three columns; the negative binomial model thus overestimates these densities relative to the historical data.

The bottom panels show the difference between the empirical frequencies and the fitted frequencies from the Poisson model. The differences are small in the first column and larger in the second and the third columns. The biggest deviations on these plots are -2.10% for inactive stocks, 25.40% for medium stocks, -14.30% for active stocks. The positive values of 1.60%, 25.30%, 5.70% in the first no-news bins show by how much the Poisson model underestimates the probability of these events in the data. These positive values implies that the actual news data is over-dispersed relative to the Poisson model.

The residuals for news tags (in light blue) are usually larger than those for news articles. For example, the negative binomial model overestimates the frequency of observations with one news tag per month, as shown by negative bars in the second bin in all three subplots. There are too few cases of one news event, since most news articles tend to have at least one topic code assigned to them. Even though the negative binomial model may accurately describe the data for the number of news articles, its fit to a more complicated distribution of the number of news tags is worse.

The comparison of residuals for the negative binomial model in the top row and the Poisson model in the bottom row suggests that the former model fits the historical data better than the latter one. We therefore implement most of our subsequent tests only for the negative binomial model.

Comparison of Three Models. Figure 4 further examines how accurately the empirical data is described by the invariance hypothesis relative to the two alternative models. There are nine plots organized into three rows and three columns. The three columns present results for the three samples of inactive stocks, medium stocks, and active stocks, respectively. The three rows correspond to the three models. In the first row, we fix the parameter $\gamma = 2/3$ for the invariance hypothesis. In the second and third rows, we fix the parameter $\gamma = 0$ for the model of invariant bet frequency and $\gamma = 1$ for the model of invariant bet size, respectively. We then estimate η and α in the negative binomial model (12) for the number of news articles. The nine figures show the difference between the empirical distributions and the fitted negative binomial distributions for the twelve bins, with the standard errors calculated using the bootstrap procedure. The smaller are residuals, the better corresponding model fits the data.

The plots in the first row suggest that the invariance hypothesis explains a large fraction of variation in the number of news articles across stocks. For inactive and medium stocks, the residuals are so small that, even despite small standard errors that are hardly seen on the charts, the invariance hypothesis is rejected. For active stocks, the residuals are large comparing to standard errors, and the model is formally rejected. The three plots almost perfectly repeat the three plots in the top row of figure 3. The difference is that to construct plots in figure 3, we do not restrict the value of γ to be equal to 2/3 predicted by the invariance hypothesis, but rather estimating this parameter together with η. Since

the estimated γ of 0.68 in table 4 is close to 2/3, the results for restricted and unrestricted specifications are very similar.

In the second row, the residuals of the first alternative model are bigger than the residuals of the invariance model. The model of invariant bet frequency assumes that the same number of news items is released per month about both inactive and active stocks. Effective, the model overestimates the number of news articles for inactive stocks and underestimates this number for active stocks. For inactive stocks in the first column, the residuals are positive in the left bins and negative in the right bins, i.e., the model predicts too few months with a few news events and too many months with a lot of news events. For active stocks in the last column, the residuals are negative in the left bins and positive in the right bins, i.e., the model predicts too many months with a few news events and too few months with a lot of news events.

In the third row, the residuals of the second alternative model are smaller in magnitude than the residuals of the first alternative model, but are somewhat comparable to the residuals of the invariance model, except for small stocks. The negative values in the left bins and positive values in the right bins indicate that, as our preferred model, this model underestimates the average number of news events. Note also that this model seems to fit the historical data for active stocks better than the invariance hypothesis. Since the exponent in the second alternative model is equal to 1 rather than 2/3, this is consistent with the convex patterns in figure 2, i.e., the information flow seems to speed up with trading activity faster than predicted by the invariance hypothesis. With these caveats, we conclude that overall the invariance hypothesis fits the historical frequencies better than alternatives.

Model Estimation. Table 5 shows the estimates for the restricted negative binomial model (12) with the exponent γ being fixed at 2/3 and the intercept η and the over-dispersion parameter α being estimated. As before, the four sets of estimates are reported in four columns for the two samples of stocks (all CRSP firms and Thomson-Reuters firms) and the two proxies for information flow (news articles and news tags).

The estimates 1.97, 2.11, 2.58, and 2.75 of η in the restricted specification are similar to the estimates 2.01, 2.08, 2.66, and 2.73 in the unrestricted specification in table 4. The estimates 2.11, 1.65, 3.30, and 2.54 of α are similar to corresponding estimates 2.05, 1.63, 3.17, and 2.49 of α in table 4 as well. Fixing $\gamma = 2/3$ also only slightly reduces the log-likelihood function. For instance, the log-likelihood decreases from -7,170 for the unrestricted specification to -7,216 for the restricted specification, if we consider our main sample of news articles about all firms and compare values in the first columns of table 4 and table 5. We interpret these result as implying that the negative binomial specification with the expected arrival rate of news events modeled according to the invariance hypothesis is a good description of the news data.

The estimate 1.97 of η for the number of news articles about all firms implies that, on average, seven news articles are released per month about a benchmark stock. The estimate 2.58 of η for the number of news tags says that these articles are tagged with about 13 topic codes, or about two tags per article. The positive estimates 2.11 and 3.30 of α confirm that the Poisson model would underestimate variation in the arrival rate of news articles and news tags.

To summarize, we find that the news data can be described by the negative binomial model with the expected arrival rate of μ news articles per month calibrated based on the sample of all CRSP firms as,

$$\mu(W) = 7.17 \cdot \Big(\frac{W}{40 \cdot 10^6 \cdot 0.02}\Big)^{2/3} \cdot \tilde{G},$$

where $\tilde{G}$ is the Gamma random variable with the mean of one and the variance of $\alpha = 2.11$ (standard error of 0.238).

Monthly Estimates from Count-Data Regressions. Figure 5 shows the estimates for the negative binomial regressions (12) run separately for each month from year 2003 through year 2008. We present four series of seventy two monthly estimates based on both the number of news articles (solid line) and the number of news tags (dashed lines) for the two samples, the sample of all CRSP firms (in dark blue) and the Thomson Reuters universe (in light blue).

There is a clear structural break in the dynamics of the estimates in year 2005. This structural break can be attributed to the change in the corporate strategy of the Reuters firm, when upon the requests of its clients, the news provider has started to extend its coverage. To examine the difference between two periods, we discuss next the average estimates before and after June 2005.

The estimates of the intercept η are stable across time, except for a permanent jump in year 2005. For the news articles and the sample of all firms, the average estimates change from 1.87 to 2.11, implying the increase from 6.50 to 8.22 news articles per month, on average, for the benchmark stock. For the news tags and the sample of all firms, the estimated intercepts increase from 2.53 to 2.76, implying the increase from 12.50 to 15.74 news tags per month. The difference between the intercepts based on the news articles and news tags does not vary throughout the sample, suggesting that news articles are typically tagged with two topic codes.

Before the structural break in 2005, the average estimated intercepts of 2.00 and 2.64 are higher for the Thomson Reuters universe than the corresponding estimates of 1.87 and 2.53 for the sample of all stocks. Afterwards, the average estimates of 2.11 and 2.76 for the Thomson Reuters universe are almost identical to the corresponding estimates 2.14 and 2.79 for the sample of all stocks. This reveals that the difference between coverage of the two samples has largely disappeared after 2005.

The estimates of the slope γ fluctuates around 2/3 as predicted by the invariance hypothesis, but these estimates also exhibit different behavior before and after the structural break. Before 2005, the average estimate for the news articles is equal to 0.78 for the sample of all firms and 0.71 for the Thomson Reuters universe. The average estimate for the news tags is equal to 0.82 for the sample of all firms and 0.73 for the Thomson Reuters universe. These estimates are slightly higher than the corresponding estimates for news tags. These patterns would be observed if news articles about bigger companies are usually tagged with more topic codes. Note that all estimates are slightly higher than predicted 2/3, possibly because the Reuters firm used to underreport information about small firms prior to year 2005. In the second half of the sample, the average estimated exponents of 0.78, 0.71, 0.82, and 0.73

decrease to 0.61, 0.60, 0.63, and 0.61, respectively, and become much more similar to each other. These lower-than-predicted sensitivities might be explained by increasing propensity of the news provider to send out articles about small firms to meet the announced goal of a global coverage, even though these articles may have not much of information content.

The monthly estimates of the over-dispersion parameter α fluctuates a lot before the structural break and remain somewhat stable and low in values afterwards. Before 2005, the average estimates of the four discussed series are 2.96, 4.55, 2.11, and 3.18, respectively. After 2005, these estimates are 1.39, 2.18, 1.28, and 1.99, respectively. Since all estimates of α are above zero, this indicates that the data is over-dispersed relative to the Poisson model and the negative binomial model therefore provides a better description of the data. Note also that the average estimates of 4.55 and 3.18 for the news tags are larger than the average estimates of 2.96 and 2.11 for the news articles before 2005 as well as afterwards (2.18 and 1.99 versus 1.39 and 1.28). As mentioned, the excessive over-dispersion of the news tags data can be explained by the tendency of the news provider to assign at least two topic codes to each news article.

The Robustness Check: Separate Coefficients for Price, Volume, and Volatility. Table 6 reports the Fama-MacBeth estimates from the negative binomial regressions with the arrival rate μ of news items per month modeled as,

$$\mu(W_{t,i}, V_{t,i}, P_{t,i}, \sigma_{r,t,i}) = e^{\eta + \frac{2}{3}\cdot\ln\left[\frac{W_{t,j}}{W^*}\right] + \beta_1\cdot\ln\left[\frac{V_{t,i}}{(10^6)}\right] + \beta_2\cdot\ln\left[\frac{P_{t,i}}{(40)}\right] + \beta_3\cdot\ln\left[\frac{\sigma_{r,t,i}}{(0.02)}\right]} \cdot \tilde{G}_{t,i}(\kappa, \theta). \qquad (13)$$

The regression imposes the restriction that the coefficient $\gamma = 2/3$ as predicted by the invariance hypothesis. It then allows the coefficient on the three components of trading activity - volume, price, and volatility - to vary freely. Since the invariance hypothesis suggests that most variation in the arrival rate of information is captured by variation in the trading activity, it predicts that $\beta_1 = \beta_2 = \beta_3 = 0$. The model of invariant bet frequency predicts $\beta_1 = \beta_2 = \beta_3 = -2/3$, and the model of invariant bet size predicts $\beta_1 = \beta_2 = \beta_3 = 1/3$.

All estimates are statistically different from zero. Across four different samples, the coefficient β_1 ranges from 0.06 to 0.09, β_2 ranges from -0.32 to -0.17, and β_3 ranges from -0.84 to -0.78. This suggests that, in addition to differences in time clock reflected in differences in trading activity, other factors may influence the information flow as well. These factors might be correlated with volume, price, and volatility, but we think that the multi-collinearity between these variables may complicate interpretation of these estimates. Note also that the F-tests of 177, 500, 126, and 367 reject the invariance hypothesis for all four samples. The alternative hypothesis are, however, rejected with much bigger F values. For the model of invariant bet frequency, the F-tests are 610, 1082, 527, and 922. For the model of invariant bet size, the F-tests are 642, 2066, 465, and 1573.

Although β_1, β_2, and β_3 are statistically significant in explaining variations in the number of news items, they are less significant in terms of their economic importance. Indeed, the values of log-likelihood functions increase only modestly from -7216, -6942, -8628, and -8325 for the univariate regressions in table 4 to -7021, -6745, -8462, and -8164 for the unrestricted regressions in table 6, respectively.

The following two examples further illustrate that, while statistically significant, the addition of three extra degrees of freedom improves the predictions about arrival rates of news articles by only a small amount. Both examples use the estimates based on the news articles for the sample of all CRSP firms. They show that the predictions of restricted and unrestricted regressions are economically similar to each other for both inactive and active stocks.

For example, if we consider inactive stocks from volume group one and plug the average values of 75588, 13.6, and 0.033 for volume, price and volatility for these stocks from table 1 into formula (12) together with estimates 1.97 of η from table 5, we find that there are, on average, 0.87 news articles reported per month about inactive stocks. Similar estimations based on less restrictive formula (13) and the estimates in table 6 imply about 0.74 news articles per month. Both estimates 0.87 and 0.74 are very similar to each other. If we consider inactive stocks in volume group ten with the average values of 8344319, 55.80, and 0.023 for volume, price and volatility for these stocks in table 1, then similar calculations imply again similar values of 40.43 and 49.40 news articles per month.

The Robustness Check: A Quadratic Term. Our earlier results showed that the relation between the news arrival rate and trading activity may be convex. We examine next this convexity effect by adding a quadratic term to the negative binomial model (12). Table 7 presents the estimates of the first-order term γ_1, the second-order term γ_2, and the over-dispersion parameter α for the negative binomial model with the news arrival rate μ modeled as,

$$\mu(W_{t,i}, V_{t,i}, P_{t,i}, \sigma_{r,t,i}) = e^{\eta+\gamma_1\cdot\left(\ln\left[\frac{W_{t,j}}{W^*}\right]-\mu_{W,t}\right)+\gamma_2\cdot\left((\ln\left[\frac{W_{t,j}}{W^*}\right]-\mu_{W,t})^2-\sigma^2_{W,t}\right)}\cdot\tilde{G}_{t,i}(\alpha). \qquad (14)$$

To avoid a multi-collinearity, we demean the covariates and ensure that they are orthogonal to each others. First, the linear term has been transformed into $\ln\left[\frac{W_{t,j}}{W^*}\right]-\mu_{W,t}$ by subtracting the cross-sectional sample mean $\mu_{W,t}$ of $\ln\left[\frac{W_{t,j}}{W^*}\right]$ for each month t. Second, the quadratic term is modeled as $(\ln\left[\frac{W_{t,j}}{W^*}\right]-\mu_{W,t})^2-\sigma^2_{W,t}$, where $\sigma_{W,t}$ is the sample standard deviation of $\ln\left[\frac{W_{t,j}}{W^*}\right]$ for each month t.

In panel A, we constrain the first-order effect γ_1 to be equal to 2/3 and then estimate the second-order effect γ_2. For both news articles and news tags, the estimated second-order coefficients γ_2 are equal to 0.03 for the sample of all firms and 0.04 for the sample of the Thomson-Reuters firms. Since their standard errors are roughly equal to only one-tenth of the estimates themselves, these estimates are statistically significant. The statistical significance of the quadratic term is consistent with non-linear patterns for the averages in figure 2.

In panel B, we estimate simultaneously both the first-order effect γ_1 and the quadratic effect γ_2. The estimates of 0.61, 0.58, 0.65, and 0.63 for γ_1 for four samples are slightly lower than the estimates of 0.68, 0.65, 0.71, and 0.66 for γ in table 4 for the negative binomial models where no quadratic term added. Although the inclusion of a quadratic term reduces these estimates, they are still close to 2/3 as predicted by the invariance hypothesis. The estimates of γ_2 are equal to about 0.04 for all four samples. The log-likelihoods of -7088, -6804, -8512, and -8204 do not improve substantially comparing to the log-likelihoods of

−7170, −6900, −8584, and −8289 for the unrestricted negative binomial model in table 4 and comparing to the log-likelihoods of −7216, −6942, −8628, and −8325 for the restricted negative binomial model in table 5.

The Robustness Check: Market Capitalization, B/M ratio, and Past Returns. We next examine how information flow depends on other stock characteristics such as the market capitalization, the book-to-market ratio, and the past return. Table 8 reports the Fama-MacBeth estimates from the monthly negative binomial regressions with the arrival rate μ of news items per month modeled as,

$$\mu(W_{t,i}, M_{t,i}, B_{t,i}/M_{t,i}, R_{t,i}) = e^{\eta + \frac{2}{3}\cdot\ln\left[\frac{W_{t,j}}{W^*}\right] + \beta_4\cdot\ln\left[M_{t,i}\right] + \beta_5\cdot\ln\left[B_{t,i}/M_{t,i}\right] + \beta_6\cdot\ln\left[R_{t,i}\right]} \cdot \tilde{G}_{t,i}(\alpha), \quad (15)$$

where $M_{t,i}$ is the market capitalization of stock i in month previous to month t, $B_{t,i}/M_{t,i}$ is its book-to-market ratio during that month, and $R_{t,i}$ is its return over previous year. The regression imposes the restriction that $\gamma = 2/3$ as predicted by the invariance hypothesis. Since the hypothesis suggests that most variation in arrival rates of information should be related to variation in the trading activity, it predicts $\beta_4 = \beta_5 = \beta_6 = 0$.

The estimates of β_4, β_5 and β_6 are statistically significant, and their values are stable across the four samples. The positive estimates of β_4 range from 0.12 to 0.17, suggesting that large stocks have "too many" news articles and small stocks have "too few" news articles relative to predictions of the invariance hypothesis. Stories about large companies are interesting to a broader audience, and therefore they may be more likely to be picked up by the news provider.

The positive coefficient of β_5 ranges from 0.24 to 0.28, implying that value stocks have "too many" news articles and growth stocks have "too few" articles. Even though growth stocks have certainly higher recognition among readers so that journalists may tend to focus more on these stocks, the number of news articles can still be too low to catch up with the increasing popularity of these stocks among traders and therefore their higher-than-normal trading activity. Large number of news articles about value stocks can be also due to the big number of articles about bankruptcy and insolvency.

The negative estimates of β_6 vary from -0.62 to -0.59 across four samples. These estimates suggest that firms with low returns have "too many" news articles and firms with high returns have "too few" articles. One possible interpretation is that there is a sluggish adjustment of information flow to changes in trading activity. For example, if stock price goes down, then the trading activity decreases, but the slowly adjusting number of news articles reported by journalists may appear to be "too high" for a new level of trading activity.

Market capitalization, B/M ratio, and past returns seem to be economically more important factors for explaining the information flow than the individual components of trading activity such as volume, price, and volatility. Indeed, the log-likelihoods of −5174, −5005, −6233, and −6048 for model (15) in table 8 are much bigger than the log-likelihoods of −7170, −6900, −8584, and −8289 for the negative binomial model in table 4. Our results suggest that further investigation of these patterns is warranted.

Count Regressions For Different News Types. So far, our tests have suggested that the invariance hypothesis provides a good explanation for the arrival rates of news events,

but it does not imply that the number of news items of different types are related to trading activity with exponent 2/3 as well. There are many examples that would contradict this literal interpretation of the invariance hypothesis. For instance, there are certainly no news events about debt markets (tag 'DBT') for firms with no debt. There are no news about bankruptcies (tag 'BKRT') for firms with healthy future prospects. There are almost no major breaking news (tag 'NEWS') about small companies, but there are many of them about large companies. Furthermore, news reports concerning earnings and dividends (tags 'RES' and 'DIV') are expected to be released at a regular schedule, regardless of firm's trading activity.

To explore our sample, we nevertheless estimate the negative binomial model (12) for the news items marked by the news provider as belonging to different news categories. Table 9 reports estimates η, γ and α for ten different categories. The first nine categories correspond to the nine most frequent topic codes in the sample - 'STX', 'RES', 'MRG', 'RESF', 'NEWS', 'CORA', 'DBT', 'RCH', 'HOT'. Their descriptions are listed in table 2. The last category "others" aggregates the remaining topic tags. The estimates are reported in ten separate lines. Each line shows the estimates of η, γ and α with their standard errors, the F-tests with p-values for three hypotheses $\gamma = 2/3$, $\gamma = 0$, and $\gamma = 1$ as well as the log-likelihood functions.

For different topic codes, the estimate of γ ranges from 0.60 to 1.23, being the lowest for "corporate results" (tag 'RES') and the highest for "major breaking news" (tag 'NEWS'). The "corporate results" include all corporate financial results, tabular and textual reports, dividends, accounts, and annual reports. Regardless of trading activity, some of these releases are usually reported at a regular calendar-time basis, for instance, once per year or per quarter. This regularity may push the exponent γ downwards to zero. In contrast, the "major breaking news category" include the stories that are expected to dominate the financial and general headlines of the worlds major newspapers, Web sites, television and radio networks. They are dominated by news articles about large firms. This may push the exponent γ upwards. The estimate of 0.69 for "forecasting of corporate financial results" (tag 'RESF') is the closest to 2/3 among all ten estimates.

6 Conclusions

We use the news data from Thomson Reuters to examine how the information flow varies across stocks and across time. Our empirical tests show that the arrival of news articles can be modeled as the negative binomial model with the stochastic Poisson arrival rate being a particular function of trading activity such that a one-percent increase in trading activity leads to two-third of one percent increase in the expected arrival rate.

This specification comes naturally from the invariance hypothesis, the main idea of which is the invariance of trading games: Trading games are the same across stocks, except for the speed with which they are being played. Our paper provides an empirical evidence supporting the conjecture that not only the actual trading processes unfold in a trading-game time clock but the information flow conforms to the same time clock as well. The conjecture is necessary to make the invariance hypothesis internally consistent.

We study one particular source of information, namely, the news articles distributed by

a news service providers to its clients. There are, however, other sources of information available in financial markets. Studying the variation in the information flow from other sources (e.g., changes in analysts' earnings forecasts and releases of 8-K filings, 10-K filings, and 10-Q filings) as well as data from other available data sets (e.g., the Dow Jones News Archive) are interesting topics for the future research. The invariance hypothesis also has implications for the flow of "hard" information and "soft" information, as examined in Engelberg (2008).

In this paper, we focus on the number of news items and put aside the discussion about news "size" by assuming that all news articles and tags are equally important. Some articles, however, are more important than others. Their important might be related to its length or the measures of how significantly articles differ from previous articles based on some language processing tools such as, for example, tools described in Hanley and Hoberg (2010). Studying variations in different aspects of the information flow is an interesting topic for the future research as well.

References

Antweiler, Werner, and Murray Z. Frank, Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards, *Journal of Finance* 59(3), 1259-1294,

Berry, Thomas D., and Keith M. Howe, 1994, Public information arrival, *Journal of Finance* 49, 1331–1346.

Chae, Joon, 2005, Trading volume, information asymmetry, and timing information, *Journal of Finance* 60, 413–442

Chan, Wesley S., 2003, Stock Price Reaction to News and No-news: Drift and Reversal after Headlines, *Journal of Financial Economics* 70, 223–260.

Clark, Peter, 1973, "A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices," *Econometrica*, 41, 135–155.

Engelberg, Joseph, 2008, Costly Information Processing: Evidence from Earnings Announcements, *University of North Carolina working paper.*

Green, T. Clifton, 2004, Economic News and the Impact of Trading on Bond Prices, *Journal of Finance* 59, 1201–1233.

Hanley, Kathleen, and Gerard Hoberg, 2010, The information content of IPO prospectuses, *Review Financial Studies*, 23 (7), 2821–2864.

Hasbrouck, Joel, 1999, "Trading Fast and Slow: Security Market Events in Real Time," *Working Paper.*

Kyle, Albert S., and Anna A. Obizhaeva, 2011a, "Market microstructure invariants: Theory and Implications of Calibration," *Working Paper*, University of Maryland.

Kyle, Albert S., and Anna A. Obizhaeva, 2011b, "Market microstructure invariants: Empirical Evidence from Portfolio Transitions," *Working Paper*, University of Maryland.

Kyle, Albert S., Anna A. Obizhaeva, and Tugkan Tuzun, 2011, "Trading Game Invariance in the TAQ Dataset," *University of Maryland, Working paper.*

Mandelbrot, Benoît, and Howard M. Taylor, 1967, "On the Distribution of Stock Price Differences," *Operations Research* 15(6), 1057–1062.

Meschke Felix and Y. Han Kim, 2010, CEO Interviews on CNBC, *Working Paper.*

Mitchell, Mark L., and J. Harold Mulherin, 1994, The impact of public information on the stock market, *Journal of Finance* 49, 923–950.

Sinha, Nitish R., 2011, Underreaction to News in the US Stock Market, *Working Paper.*

Tetlock, Paul C., 2010, Does Public Financial News Resolve Asymmetric Information?, Review of Financial Studies 23, 3520–3557.

Tetlock, Paul C., Maytal Saar-Tsechansky, and Sofus Macskassy, 2008, More Than Words: Quantifying Language to Measure Firms Fundamentals, *Journal of Finance* 63, 1437–1467.

Tetlock, Paul C., 2007, Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *Journal of Finance* 62, 1139–1168.

Tetlock, Paul C., All the News Thats Fit to Reprint: Do Investors React to Stale Information? *Working paper.*

Table 1: Descriptive Statistics

Volume Groups:	All	1	2	3	4	5	6	7	8	9	10
	Panel A: The Descriptive Statistics for the Sample of Stocks.										
Avg. Volume ($1000)	21,931	1,028	8,671	16,634	26,692	38,218	50,409	67,315	95,086	154,837	465,613
Volatility	0.031	0.033	0.027	0.026	0.025	0.024	0.024	0.024	0.023	0.023	0.023
Avg. Price	21.1	13.6	27.1	30.8	33.7	38.1	40.9	41.7	45.9	49.2	55.8
	Panel B: The Descriptive Statistics for the Sample of Thomson Reuters News.										
Avg. # of articles/month	4.24	0.58	2.13	3.36	5.16	7.18	9.37	11.78	15.12	26.74	82.86
Med. # of articles/month	0	0	1	2	3	4	5	7	10	16	46
Max. # of articles/month	3344	143	183	242	221	198	367	259	817	1,789	3,344
Min. # of articles/month	0	0	0	0	0	0	0	0	0	0	0
Avg. # of tags/month	9	1	4	6	9	13	17	21	27	47	145
Med. # of tags/month	1	0	2	3	5	6	9	13	17	28	84
Max. # of tags/month	7679	310	579	569	423	306	986	484	1407	3370	7,679
Min. # of tags/month	0	0	0	0	0	0	0	0	0	0	0
No news articles/month	58%	73%	45%	35%	27%	22%	17%	13%	10%	7%	5%
# Obs. (CRSP firms-months)	340,505	222,543	41,719	17,620	16,070	7,622	7,171	6,947	6,829	6,841	7,143
# Obs. (TR firms-months)	275,059	166,679	37,170	15,916	14,715	7,072	6,730	6,598	6,583	6,649	6,947

Table provides a descriptive statistics for our sample. Panel A reports the average dollar trading volume per day, the standard deviation of daily returns, the average price, and the trading activity of stocks in our sample. Panel B reports the average, the median, the minimum and the maximum numbers of news articles per month; the average, the median, the minimum and the maximum numbers of news tags per month, the fraction of stocks without any news articles during a given month. The table reports also the number of all observations, stock-month pairs, in our sample of all CRSP firms from January 2003 through December 2008 as well as in the sample of firms covered by Thomson-Reuters firm. Statistics is reported for the total sample and for the ten volume groups. The volume groups are based on the average dollar trading volume with thresholds corresponding to 30th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, and 95th percentiles of the dollar volume for common NYSE-listed stocks. Volume group 1 (group 10) has stocks with the lowest (highest) trading volume.

Table 2: The List of Topic Codes.

TOPIC	Description	# news tags	% of all
STX	Regulations, additions and deletions from indices, new listings, delistings.	508,430	14.79%
RES	All corporate financial results, tabular and textual reports, dividends, annual reports.	475,536	13.84%
MRG	Changes of ownership including mergers and acquisitions	402,443	11.71%
RESF	Forecasting of corporate financial results, reports.	333,921	9.72%
NEWS	Major breaking news.	322,301	9.38%
CORA	Corporate analysis.	228,267	6.64%
DBT	All debt market news.	222,902	6.49%
RCH	All news about broker research and recommendations.	165,252	4.81%
HOT	News about stocks on the move.	152,253	4.43%
INV	All news about the process of investing on the part of individuals.	131,537	3.83%
REGS	Regulatory issues	101,693	2.96%
PRO	People in the news, biographies, profiles.	77,476	2.25%
MNGIS	Management issues/policy.	68,819	2.00%
AAA	All news about credit ratings.	51,690	1.50%
IPO	Initial public offerings.	30,073	0.88%
PRESS	Press digests.	29,795	0.87%
DIV	Dividends forecasts, declarations, and payments.	28,424	0.83%
JUDIC	Stories about judicial processes, court cases and decisions.	26,609	0.77%
WIN	Reuters exclusive news.	17,829	0.52%
EXCA	Exchange activities.	15,061	0.44%
FED	Federal Reserve Board activities and news.	12,843	0.37%
ECI	News, forecasts or analysis of economic indicators.	11,379	0.33%
BKRT	Stories on bankruptcies and insolvencies.	11,166	0.32%
RSUM	Stories from Reuters summits.	10,243	0.30%
FES	Editorial special, analysis and future stories.	267	0.01%
ERR	Errors.	204	0.01%
CFIN	Corporate finance.	143	0.00%
INSI	Stories about technical analysis of markets.	80	0.00%
CDM	Credit market news.	38	0.00%
TRN	Translated news.	29	0.00%
CONV	Convertible bonds news.	24	0.00%
NEWR	Original corporate news releases	1	0.00%
			100.00%

Table describes a listing of topic codes in the sample. The topic code tag, its brief description, the number of news articles tagged with the particular topic code and the percentage of these tags in the total sample are reported.

Table 3: The OLS Estimates for The Average Number of News Items.

	News Articles		News Tags	
	CRSP	Thomson-Reuters	CRSP	Thomson-Reuters
η	2.32	2.41	3.02	3.12
	(0.270)	(0.215)	(0.270)	(0.216)
γ	0.70	0.65	0.75	0.70
	(0.104)	(0.086)	(0.085)	(0.075)
	Model of Trading Game Invariance : $\gamma = 2/3$			
F-Test	0.08	0.03	0.79	0.13
p-val	0.774	0.862	0.382	0.719
	Model of Invariant Bet Frequency : $\gamma = 0$			
F-Test	45.55	57.39	76.15	85.68
p-val	0.000	0.000	0.000	0.000
	Model of Invariant Bet Size: $\gamma = 1$			
F-Test	8.36	15.94	8.84	16.13
p-val	0.007	0.000	0.006	0.000
Avg. R^2	0.912	0.893	0.917	0.896
# Obs	30	30	30	30
# Months	72	72	72	72

Tables shows the estimates for the regression:

$$\ln \bar{N}^*_{t,i} = \eta + \gamma \ln \big[\frac{\bar{W}_{t,i}}{W^*}\big] + \tilde{\epsilon}_{t,i}.$$

For each month, stocks are sorted into the thirty groups based on the trading activity, such that these groups have the same total number of news. Each observation corresponds to the pair of month t and group i. The variable $\bar{N}^*_{t,i}$ is equal to the average number of news about stocks in group i, arrived during month t and adjusted for the within-group variation in the trading activity for that observations. The variable $\bar{W}_{t,i}$ is the average trading activity of stocks in group i, with the trading activity being the product of the average daily dollar volume and the standard deviation of daily returns. The scaling constant $W_* = (40)(10^6)(0.02)$ corresponds to the trading activity of the benchmark stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. The Newey-West standard errors computed with the three lags from the Fama-MacBeth regressions are in parentheses. The F-statistics and p-values are calculated from the Fama-MacBeth regressions with the Newey-West correction for three different models. The sample of all firms and the sample of firms covered by the Thomson-Reuters company are considered. The estimates for the number of news articles and news tags are presented separately. The sample ranges from January 2003 to December 2008.

Table 4: The Count Regression Estimates for The Number of News Items.

	News Articles				News Tags			
	CRSP		Thomson-Reuters		CRSP		Thomson-Reuters	
	Pois	NB	Pois	NB	Pois	NB	Pois	NB
η	2.11	2.01	2.19	2.08	2.78	2.66	2.85	2.73
	(0.044)	(0.036)	(0.037)	(0.028)	(0.049)	(0.037)	(0.044)	(0.030)
γ	0.81	0.68	0.78	0.65	0.86	0.71	0.84	0.66
	(0.007)	(0.024)	(0.007)	(0.018)	(0.008)	(0.025)	(0.010)	(0.019)
α		2.05		1.63		3.17		2.49
		(0.218)		(0.120)		(0.325)		(0.170)
	Model of Trading Game Invariance : $\gamma = 2/3$							
F-Test	4,078	0	282	2	566	2	286	0
p-val	0.000	0.532	0.000	0.189	0.000	0.165	0.000	0.732
	Model of Invariant Bet Frequency : $\gamma = 0$							
F-Test	14095	835	13,296	1,366	11,793	772	7,270	1,273
p-val	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Model of Invariant Bet Size: $\gamma = 1$							
F-Test	803	177	1008	407	324	134	281	327
p-val	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
$\log(L)$	-16,590	-7,170	-15,722	-6,900	-33,249	-8,584	-31,570	-8,289

Tables shows the estimates for the count regressions. For the Poisson regression, the arrival rate of news items $\mu_{t,i}$ for stock i and month t is modeled as,

$$\mu_{t,i} = e^{\eta+\gamma\cdot\ln\left[\frac{W_{t,i}}{W^*}\right]}.$$

For the Negative Binomial regression, the arrival rate of news $\mu_{t,i}$ for stock i and month t is modeled as,

$$\mu_{t,i} = e^{\eta+\gamma\cdot\ln\left[\frac{W_{t,i}}{W^*}\right]} \cdot \tilde{G}_{t,i}(\alpha),$$

where the Gamma variable $\tilde{G}_{t,i}$ has the mean equal to one and the variance equal to α. The trading activity $W_{t,i}$ is the product of the average daily dollar volume and the average standard deviation of daily returns. The scaling constant $W_* = (40)(10^6)(0.02)$ corresponds to the trading activity of the benchmark stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. The Newey-West standard errors computed with the three lags from the Fama-MacBeth regressions are in parentheses. The F-statistics and p-values are calculated from the Fama-MacBeth regressions with the Newey-West correction for three different models. The sample of all firms and the sample of firms covered by the Thomson-Reuters company are considered. The estimates for the number of news articles and news tags are presented separately. The sample ranges from January 2003 to December 2008.

Table 5: The Model Estimation.

	News Articles		News Tags	
	CRSP	Thomson-Reuters	CSRP	Thomson-Reuters
η	1.97	2.11	2.58	2.75
	(0.068)	(0.043)	(0.079)	(0.050)
α	2.11	1.65	3.30	2.54
	(0.238)	(0.126)	(0.350)	(0.177)
$\log(L)$	-7,216	-6,942	-8,628	-8,325

Tables shows the estimates for the Negative Binomial regression with the arrival rate of news items $\mu_{t,i}$ for stock i and month t being modeled as,

$$\mu_{t,i} = e^{\eta + 2/3 \ln\left[\frac{W_{t,i}}{W^*}\right]} \cdot \tilde{G}_{t,i}(\alpha),$$

where the Gamma variable $\tilde{G}_{t,i}$ has the mean equal to one and the variance equal to α. The trading activity $W_{t,i}$ is the product of the average daily dollar volume and the average standard deviation of daily returns. The scaling constant $W_* = (40)(10^6)(0.02)$ corresponds to the trading activity of the benchmark stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. The Newey-West standard errors computed with the three lags from the Fama-MacBeth regressions are in parentheses. The F-statistics and p-values are calculated from the Fama-MacBeth regressions with the Newey-West correction for three different models. The sample of all firms and the sample of firms covered by the Thomson-Reuters company are considered. The estimates for the number of news articles and news tags are presented separately. The sample ranges from January 2003 to December 2008.

Table 6: The Count Regression: A More General Specification.

	News Articles		News Tags	
	CRSP	Thomson-Reuters	CSRP	Thomson-Reuters
η	2.19	2.14	2.91	2.84
	(0.058)	(0.050)	(0.070)	(0.059)
β_1	0.08	0.06	0.09	0.07
	(0.019)	(0.015)	(0.021)	(0.017)
β_2	-0.22	-0.32	-0.17	-0.28
	(0.032)	(0.017)	(0.034)	0.018
β_3	-0.83	-0.84	-0.78	-0.81
	(0.061)	(0.058)	(0.061)	0.059
	Model of Trading Game Invariance: $\beta_1 = \beta_2 = \beta_3 = 0$			
F-Test	177	500	126	367
p-val	0.000	0.000	0.000	0.000
	Model of Invariant Bet Frequency: $\beta_1 = \beta_2 = \beta_3 = -2/3$			
F-Test	610	1,082	527	922
p-val	0.000	0.000	0.000	0.000
	Model of Invariant Bet Size: $\beta_1 = \beta_2 = \beta_3 = 1/3$			
F-Test	642	2,066	465	1,573
p-val	0.000	0.000	0.000	0.000
$\log(L)$	-7,021	-6,745	-8,462	-8,164

Tables shows the estimates for the Negative Binomial regression with the arrival rate of news items $\mu_{t,i}$ for stock i and month t being modeled as,

$$\mu_{t,i} = e^{\eta + 2/3 \ln\left[\frac{W_{t,i}}{W^*}\right] + \beta_1 \ln\left[\frac{V_{t,i}}{10^6}\right] + \beta_2 \ln\left[\frac{P_{t,i}}{40}\right] + \beta_3 \ln\left[\frac{\sigma_{t,i}}{0.02}\right]} \cdot \tilde{G}_{t,i}(\alpha),$$

where the Gamma variable $\tilde{G}_{t,i}$ has the mean equal to one and the variance equal to α. The trading activity $W_{t,i}$ is the product of the average daily dollar volume $V_{t,i} \cdot P_{t,i}$ and the average standard deviation of daily returns $\sigma_{t,i}$. The scaling constant $W_* = (40)(10^6)(0.02)$ corresponds to the trading activity of the benchmark stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. The Newey-West standard errors computed with the three lags from the Fama-MacBeth regressions are in parentheses. The F-statistics and p-values are calculated from the Fama-MacBeth regressions with the Newey-West correction for three different models. The sample of all firms and the sample of firms covered by the Thomson-Reuters company are considered. The estimates for the number of news articles and news tags are presented separately. The sample ranges from January 2003 to December 2008.

Table 7: The Count Regression: A Quadratic Specification.

	News Articles		News Tags	
	CRSP	Thomson-Reuters	CRSP	Thomson-Reuters
	Panel A: Restricted Model, $\gamma_1 = 2/3$.			
η	-0.23	0.24	0.39	0.87
	(0.065)	(0.042)	(0.077)	(0.047)
γ_2	0.03	0.04	0.03	0.04
	(0.004)	(0.003)	(0.003)	(0.003)
α	2.01	1.58	3.15	2.42
	(0.223)	(0.121)	(0.337)	(0.170)
$\log(L)$	-7,176	-6,886	-8,579	-8,257
	Panel B: Unrestricted Model.			
η	-0.15	0.32	0.39	0.88
	(0.105)	(0.055)	(0.106)	(0.056)
γ_1	0.61	0.58	0.65	0.63
	(0.029)	(0.018)	(0.030)	(0.019)
γ_2	0.04	0.04	0.04	0.04
	(0.003)	(0.002)	(0.003)	(0.002)
α	1.95	1.52	3.05	2.36
	(0.226)	(0.125)	(0.335)	(0.176)
$\log(L)$	-7,088	-6,804	-8,512	-8,204
	Model of Trading Game Invariance: $\gamma_1 = 2/3, \gamma_2 = 0$			
F-Test	66	279	59	271
p-val	0.000	0.000	0.000	0.000

Tables shows the estimates for the Negative Binomial regression with the arrival rate of news items $\mu_{t,i}$ for stock i and month t being modeled as,

$$\mu(W_{t,i}, V_{t,i}, P_{t,i}, \sigma_{r,t,i}) = e^{\eta+\gamma_1 \cdot \left(\ln\left[\frac{W_{t,j}}{W^*}\right] - \mu_W \right) + \gamma_2 \cdot \left((\ln\left[\frac{W_{t,j}}{W^*}\right] - \mu_W)^2 - \sigma_W^2 \right)} \cdot \tilde{G}_{t,i}(\alpha),$$

where the Gamma variable $\tilde{G}_{t,i}$ has the mean equal to one and the variance equal to α. The constants μ_W is the sample mean of $\ln\left[\frac{W}{W^*}\right]$ and σ_W is the sample standard deviation of $\ln\left[\frac{W}{W^*}\right]$. The trading activity $W_{t,i}$ is the product of the average daily dollar volume $V_{t,i} \cdot P_{t,i}$ and the average standard deviation of daily returns $\sigma_{t,i}$. The scaling constant $W_* = (40)(10^6)(0.02)$ corresponds to the trading activity of the benchmark stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. The Newey-West standard errors computed with the three lags from the Fama-MacBeth regressions are in parentheses. The F-statistics and p-values are calculated from the Fama-MacBeth regressions with the Newey-West correction for restrictions $\gamma_1 = 2/3, \gamma_2 = 0$. The sample of all firms and the sample of firms covered by the Thomson-Reuters company are considered. The estimates for the number of news articles and news tags are presented separately. The sample ranges from January 2003 to December 2008.

Table 8: The Count Regression: Market Capitalization, B/M, and Past Returns.

	News Articles		News Tags	
	CRSP	Thomson-Reuters	CRSP	Thomson-Reuters
η	1.58	0.23	0.08	0.83
	(0.159)	(0.332)	(0.476)	(0.344)
β_4	0.16	0.12	0.17	0.12
	(0.028)	(0.022)	(0.031)	(0.023)
β_5	0.26	0.28	0.24	0.25
	(0.025)	(0.026)	(0.021)	(0.023)
β_6	-0.61	-0.62	-0.59	-0.61
	(0.028)	(0.027)	(0.028)	(0.027)
	Model of Trading Game Invariance: $\beta_4 = \beta_5 = \beta_6 = 0$			
F-Test	245	265	281	243
p-val	0.000	0.000	0.000	0.000
$\log(L)$	-5,174	-5,005	-6,233	-6,048

Tables shows the estimates for the Negative Binomial regression with the arrival rate of news items $\mu_{t,i}$ for stock i and month t being modeled as,

$$\mu_{t,i} = e^{\eta + 2/3 \ln\left[\frac{W_{t,i}}{W^*}\right] + \beta_1 \ln\left[M_{t,i}\right] + \beta_2 \ln\left[B/M_{t,i}\right] + \beta_3 \ln\left[R_{t,i}\right]} \cdot \tilde{G}_{t,i}(\alpha),$$

where the Gamma variable $\tilde{G}_{t,i}$ has the mean equal to one and the variance equal to α. $M_{t,i}$ is the market capitalization of stock i in month t, $B_{t,i}$ is the book value of equity of stock i in month t, $R_{t,i}$ is the past return of stock i in month t. The trading activity $W_{t,i}$ is the product of the average daily dollar volume $V_{t,i} \cdot P_{t,i}$ and the average standard deviation of daily returns $\sigma_{t,i}$. The scaling constant $W_* = (40)(10^6)(0.02)$ corresponds to the trading activity of the benchmark stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. The Newey-West standard errors computed with the three lags from the Fama-MacBeth regressions are in parentheses. The F-statistics and p-values are calculated from the Fama-MacBeth regressions with the Newey-West correction for the hypothesis that market capitalization, B/M ratios, and past returns do not have additional explanatory power. The sample of all firms and the sample of firms covered by the Thomson-Reuters company are considered. The estimates for the number of news articles and news tags are presented separately. The sample ranges from January 2003 to December 2008.

Table 9: The Count Regression Estimates for Different Types of News Tags.

Tags	%	Estimates			F-Tests (p-val)			log(L)
		η	γ	α	$\gamma = 2/3$	γ=0	γ=1	
STX	14.79%	0.46	0.98	4.94	254.01	2565.71	1.35	-3,038.4
		(0.142)	(0.019)	(0.520)	(0.000)	(0.000)	(0.250)	
RES	13.84%	0.89	0.60	2.84	6.38	420.18	192.27	-4,673.9
		(0.050)	(0.029)	(0.229)	(0.014)	(0.000)	(0.000)	
MRG	11.71%	0.40	0.81	7.98	15.58	518.33	28.33	-2,712.2
		(0.036)	(0.036)	(0.681)	(0.000)	(0.000)	(0.000)	
RESF	9.72%	0.60	0.69	3.15	0.18	354.45	74.75	-3,725.3
		(0.048)	(0.036)	(0.235)	(0.676)	(0.000)	(0.000)	
NEWS	9.38%	0.01	1.23	9.50	792.22	3,842.12	131.79	-1,814.8
		(0.109)	(0.020)	(0.302)	(0.000)	(0.000)	(0.000)	
CORA	6.64%	-3.25	1.08	38.78	62.45	438.39	2.20	-1,886.6
		(1.332)	(0.051)	(22.38)	(0.000)	(0.000)	(0.143)	
DBT	6.49%	-0.03	0.86	7.22	120.63	2,406.01	60.33	-2,110.9
		(0.050)	(0.018)	(0.248)	(0.000)	(0.000)	(0.000)	
RCH	4.81%	-0.37	0.74	3.67	6.53	658.36	77.87	-2,325.6
		(0.153)	(0.029)	(0.537)	(0.013)	(0.000)	(0.000)	
HOT	4.43%	-0.26	0.90	5.46	123.86	1,865.08	21.72	-1,967.7
		(0.068)	(0.021)	(0.185)	(0.000)	(0.000)	(0.000)	
Others	18.19%	0.84	0.72	5.41	3.68	783.35	119.29	-3,853.8
		(0.043)	(0.026)	(0.318)	(0.059)	(0.000)	(0.000)	

Tables shows the estimates of the intercept η, the slope γ, and the dispersion α from the Negative Binomial regressions for the number of news tags, with the arrival rate of news tags $\mu_{t,i}$ for stock i and month t being modeled as,

$$\mu_{t,i} = e^{\eta+\gamma\cdot\ln\left[\frac{W_{t,i}}{W^*}\right]} \cdot \tilde{G}_{t,i}(\alpha),$$

where the Gamma variable $\tilde{G}_{t,i}$ has the mean equal to one and the variance equal to α. The trading activity $W_{t,i}$ is the product of the average daily dollar volume and the average standard deviation of daily returns. The scaling constant $W_* = (40)(10^6)(0.02)$ corresponds to the trading activity of the benchmark stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. The analysis is implemented separately for the nine most frequent types of new tags (RES, STX, MRG, RESF, NEWS, CORA, DBT, RCH, HOT) as well as the remaining news tags aggregated in the line "Others." The Newey-West standard errors computed with the three lags from the Fama-MacBeth regressions are in parentheses. The F-statistics and p-values (in parentheses) are calculated from the Fama-MacBeth regressions with the Newey-West correction for three different models. The percentage of news tags in each news category is shown in percents. The logarithm of likelihood function is in the last column. The sample of all firms is considered. The sample ranges from January 2003 to December 2008.

Figure 1: The Historical Distributions of The Number of News Items.

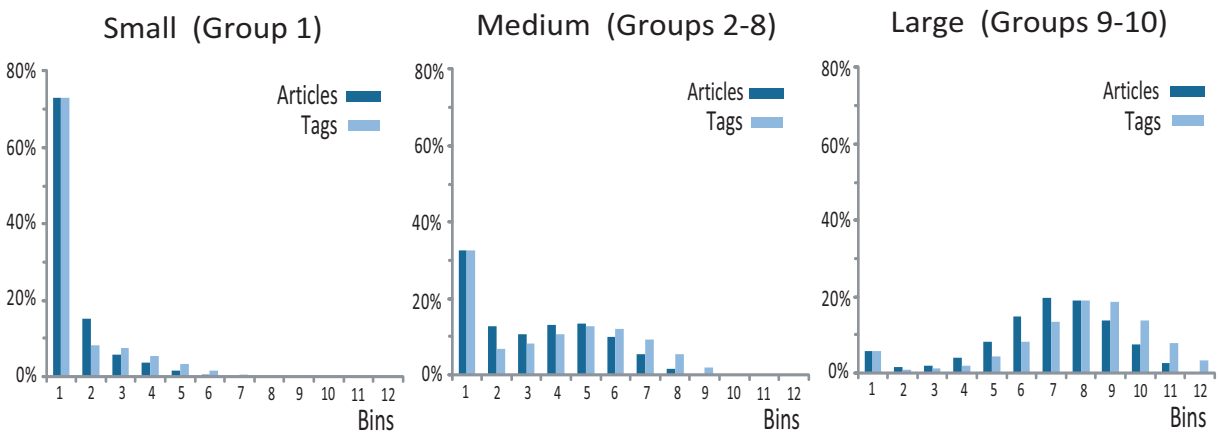

Figure shows the historical distributions of the number of news items N per month. The twelve bins have observations with $0, 1, 2, 3-4, 5-8, 9-16, 17-32, 33-64, 65-128, 129-256, 257-512, 513-1024$ news items per month, respectively; most of them have upper cutoffs of the form 2^i news items per month. The distributions are averaged across stocks. There are subplots for the small stocks from volume group 1, the medium stocks from volume group 2 through 8, and the large stocks from volume groups 9 and 10. Volume groups are based on average dollar trading volume with thresholds corresponding to 30th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, and 95th percentiles of the dollar volume for common NYSE-listed stocks. Volume group 1 (group 10) has stocks with the lowest (highest) trading volume. The distribution of the number of news articles is marked in dark blue color. The distribution of the number of news tags is marked in dark blue color. The sample covers all firms including those not covered by the Thomson-Reuters dataset. The sample ranges from January 2003 to December 2008.

Figure 2: The Number of News Items Across Trading Activity Groups.

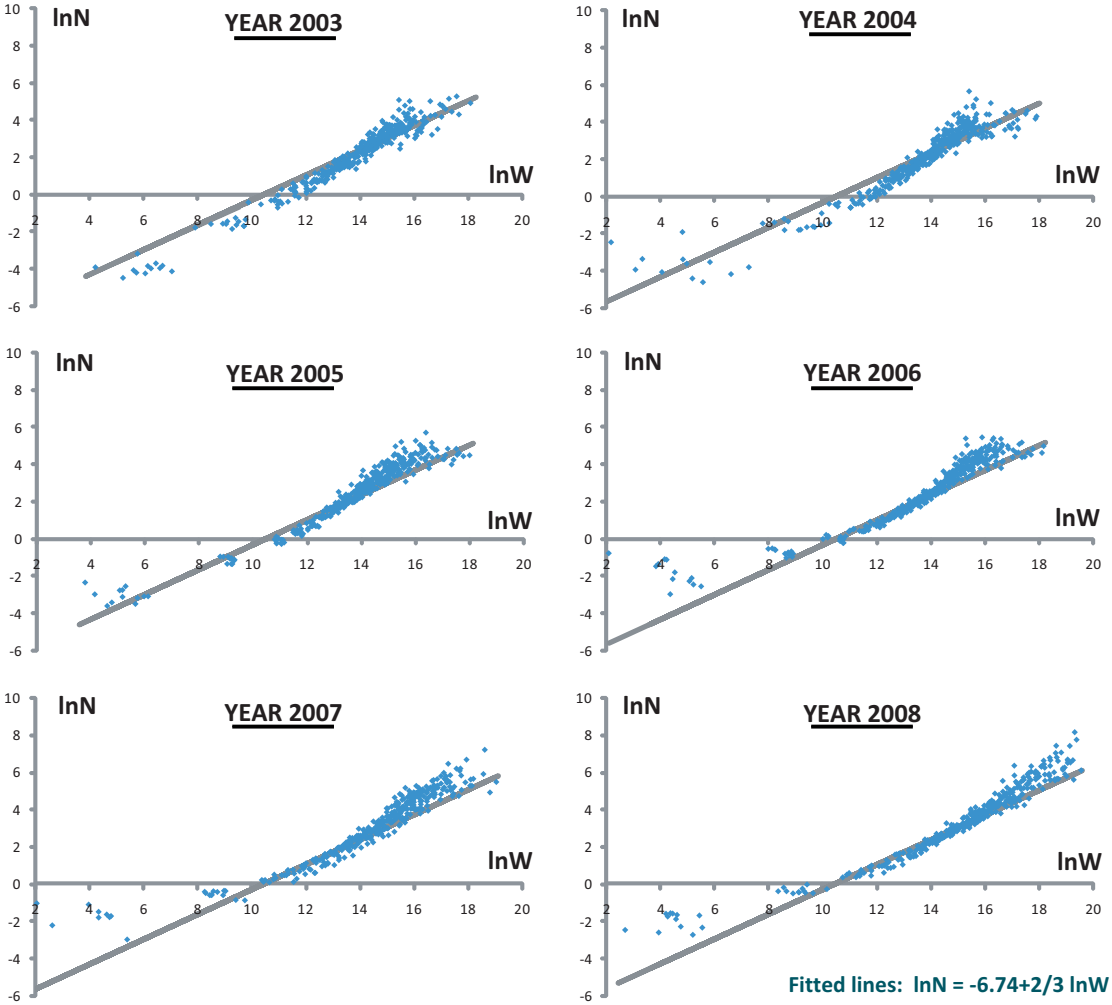

Figure shows the average of the logarithm of the number of news articles released per month for thirty groups based on the trading activity. For each month, stocks are sorted into thirty W-groups such that these groups have the same total number of news articles. The variable N^* is equal to the number N of news articles arrived during the month and adjusted for the within-group variation in the trading activity. The trading activity W is calculated as the product of the monthly dollar volume and returns standard deviation. For each group and each month, the average number N^* of news articles and the average measure of trading activity W are plotted, separately for each of the six years from 2003 through 2008. The same fitted line $\ln N^* = -6.74 + 2/3 \times \ln W$ is superimposed on each subplot, its intercept -6.74 estimated from the sample of all observations. The sample covers all firms including those not covered by the Thomson-Reuters dataset.

Figure 3: The Residuals from Poisson and Negative Binomial Specifications.

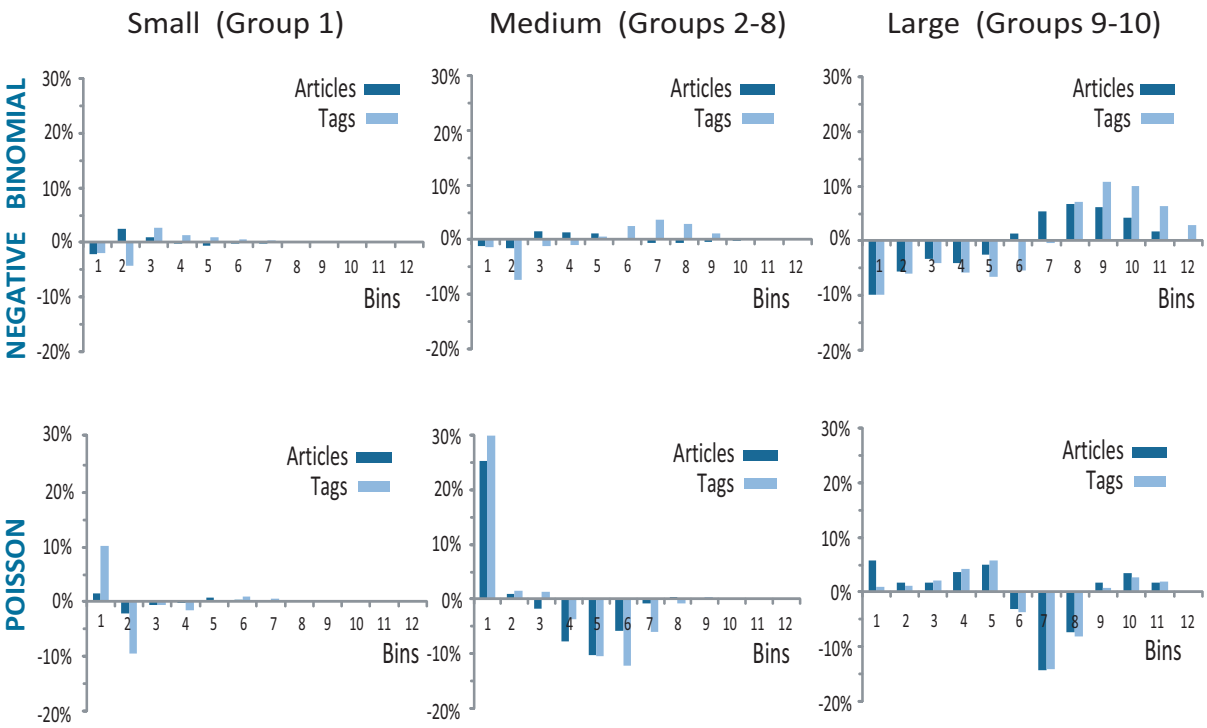

Figure shows the difference between the historical distribution and the fitted distribution of the number of news items N per month. Two specifications are used for the fitted distributions: the Poisson model and the Negative binomial model. The estimates of their parameters are taken from table 4. The twelve bins have observations with $0, 1, 2, 3-4, 5-8, 9-16, 17-32, 33-64, 65-128, 129-256, 257-512, 513-1024$ news items per month, respectively; most of them have upper cutoffs of the form 2^i news items per month. The distributions are averaged across stocks. There are subplots for the small stocks from volume group 1, the medium stocks from volume group 2 through 8, and the large stocks from volume groups 9 and 10. Volume groups are based on average dollar trading volume with thresholds corresponding to 30th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, and 95th percentiles of the dollar volume for common NYSE-listed stocks. Volume group 1 (group 10) has stocks with the lowest (highest) trading volume. The difference between historical and estimated distributions based on the number of news articles is marked in dark blue color. The difference between historical and estimated distributions based on the number of news tags is marked in dark blue color. The sample of all firms is considered. The sample ranges from January 2003 to December 2008.

Figure 4: The Residuals from The Three Models.

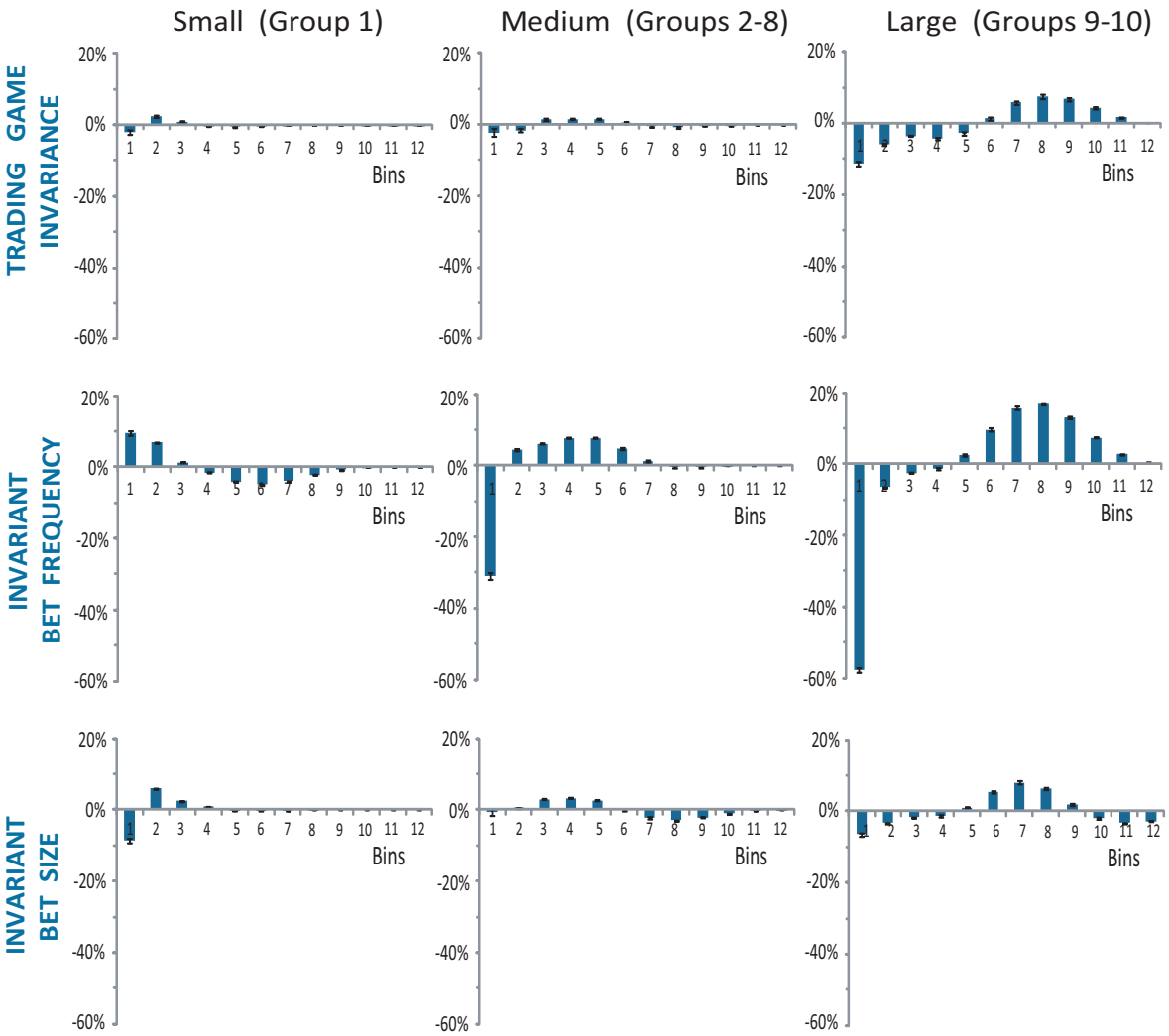

Figure shows the difference between the historical distribution and the fitted distribution of the number of news items articles N per month for the three models. The fitted distribution is based on the estimates for the Negative binomial specification. In calibrating the model, the parameter γ is restricted to be "2/3" for the model of trading game invariance, "0" for the model of invariant bet frequency, and "1" for the model of invariant bet size. The twelve bins have observations with $0, 1, 2, 3-4, 5-8, 9-16, 17-32, 33-64, 65-128, 129-256, 257-512, 513-1024$ news items per month, respectively; most of them have upper cutoffs of the form 2^i news items per month. The distributions are averaged across stocks. There are subplots for the small stocks from volume group 1, the medium stocks from volume group 2 through 8, and the large stocks from volume groups 9 and 10. Volume groups are based on average dollar trading volume with thresholds corresponding to 30th, 50th, 60th, 70th, 75th, 80th, 85th, 90th, and 95th percentiles of the dollar volume for common NYSE-listed stocks. Volume group 1 (group 10) has stocks with the lowest (highest) trading volume. The standard errors are calculated using a bootstrap[39] The sample of all firms is considered. The sample ranges from January 2003 to December 2008.

Figure 5: The Estimates from Count Regressions from January 2003 to December 2008.

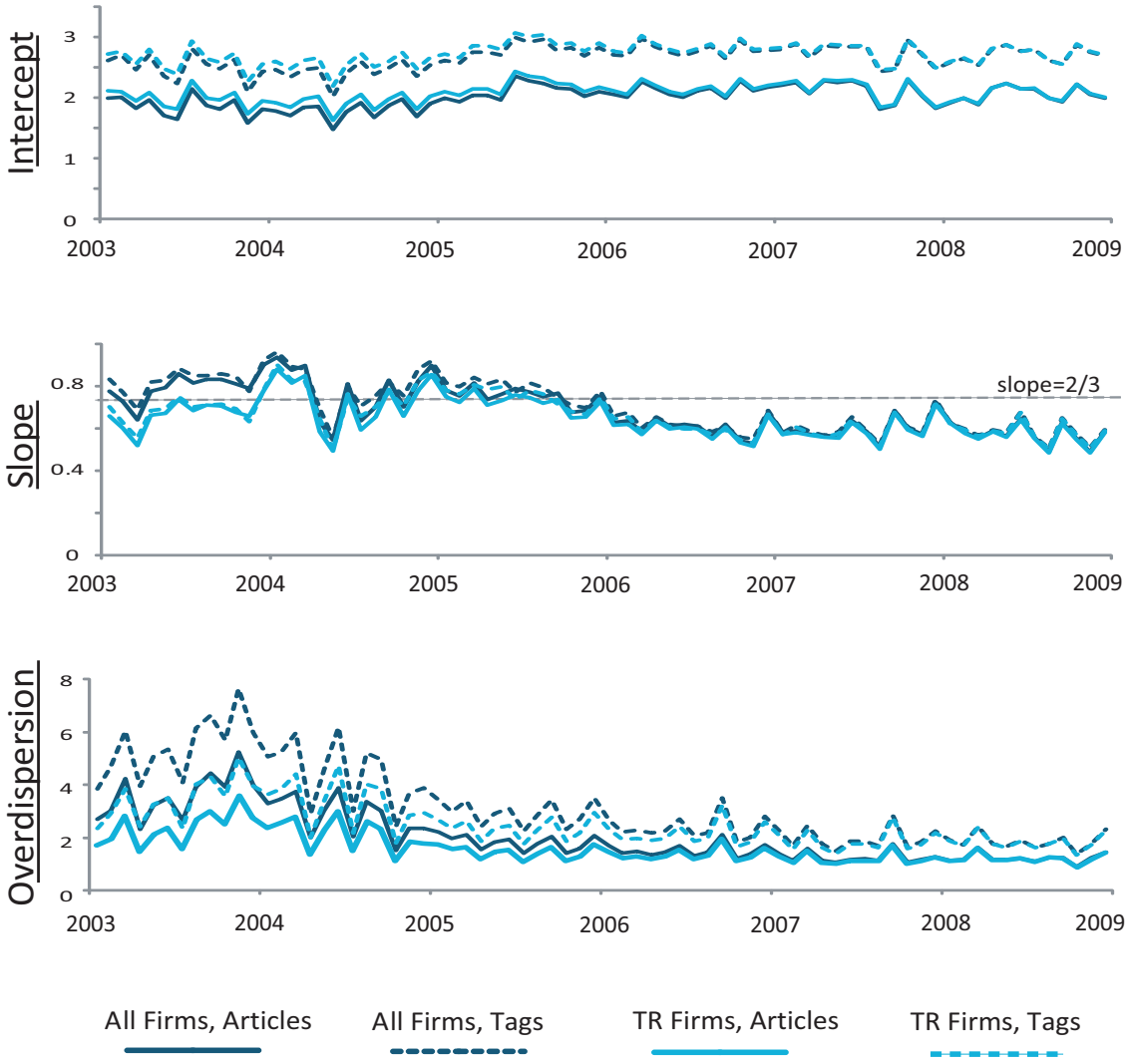

Figure shows the estimates of the intercept η, the slope γ, and the overdispersion parameter α from the negative binomial regression,with the arrival rate of news items $\mu_{t,i}$ for stock i and month t being modeled as,

$$\mu_{t,i} = e^{\eta+\gamma\cdot\ln\left[\frac{W_{t,i}}{W^*}\right]} \cdot \tilde{G}_{t,i},$$

where the Gamma variable $\tilde{G}_{t,i}$ has the mean equal to one and the variance equal to α. The trading activity $W_{t,i}$ is the product of the average daily dollar volume and the average standard deviation of daily returns. The scaling constant $W_* = (40)(10^6)(0.02)$ corresponds to the trading activity of the benchmark stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. The estimates are plotted for each of sixty months between 2003 and 2008. The estimates are provided for four samples: the sample of news articles about all firms, the sample of news articles about firms covered by the Thomson-Reuters company, the sample of news tags about all firms, and the sample of news tags about firms covered by Thomson-Reuters.