News Arrival and Trading Game Invariance

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
Robert H. Smith School of Business
University of Maryland
stuzun@rhsmith.umd.edu

Preliminary Version

First Draft: August 5, 2010; This Draft: March 14, 2011

Abstract

Our paper examines how the number of news articles varies across stocks and across time, using the news data from Thomson Reuters firm for the period from 2003 through 2008. This variation is largely explained by the variation in the level of trading activity, defined as a product of dollar trading volume and volatility. When the trading activity increases by one percent, the number of news articles increases by a two-third of one percent. This paper thus provides the evidence important for making the model of market microstructure invariance internally consistent: Not only the trading process unfolds with a trading-game time clock, but the information flow conforms to the same time clock as well.

1 Introduction

The information flow varies across stocks and across time. It is not unusual that over the same period, numerous news articles are released about large firms like IBM and hardly any articles about small firms like Xilinx. The model of market microstructure invariance, developed in Kyle and Obizhaeva (2010), predicts a specific quantitative relationship between the information flow and stock characteristics. We study this prediction by examining the frequency of the news articles provided by Thomson-Reuters firm for its clients.

In Kyle and Obizhaeva (2010), traders are playing trading games. The deep parameters of these games are the same, up to some Modigliani-Miller transformations, except for the speed with which the games are being played. The time clock runs faster for active stocks and slower for inactive stocks. This idea leads to a number of specific predictions about how various trading variables such as order sizes, order arrival rate, and trading costs vary across stocks with different characteristics. The flow of information sets a context in which trading processes evolve, representing another side of the trading game. It is natural therefore to assume that the information flow unfolds with the same time clock as the time clock according to which trading games themselves are being played. This conjecture leads to a specific prediction about the relationship between the information flow and stock characteristics.

In particular, traders are assumed to place "bets" - statistically independent decisions to trade specific quantities - at a rate proportional to which the time clock ticks. The bet size is defined as the product of the dollar order size and the standard deviation of daily returns; it represents the risk transferred during a transaction. "Trading activity" is defined as the product of the daily dollar volume and the standard deviation of daily returns; it represents the total risk transferred during all transactions executed over a day. The trading activity and the information flow are not affected by the Modigliani-Miller transformations such as changes in leverage and stock splits. In contrast, they are affected by speeding up or slowing down trading games. As the time clock speeds up, the rate of information flow increases proportionately, but the trading activity increases even faster because of the interaction of two effects: "volume effect" and "volatility effect." The volume increases proportionately and the volatility, being the square root of the variance, increases half as fast. Putting these two observations together, it is easy to draw an important conclusion: When trading activity increases by one percent, the rate of information flow speeds up by a two-third of one percent.

To convert this theoretical prediction about an individual stock into empirically testable hypotheses about the cross-section of stocks, we make two empirical assumptions:

- Information Flow Invariance: (Public and private) information arrives at a rate proportional to the rate at which the time clock ticks, with a proportionality constant being the same across stocks and across time for the same stock.

- News Article Invariance: News articles arrive at a rate proportional to the rate at which information arrives, with a proportionality constant being the same across stocks and across time for the same stock.

The proportionality constants are examples of the market microstructure invariants. These empirical hypothesis are parallel to the hypotheses of Trading Game Invariance, Market

Impact Invariance, and Bid-Ask Spread Invariance, discussed in Kyle and Obizhaeva (2010). These hypotheses together with the model of trading game invariance predict that when the trading activity increases by one percent, the number of articles increases by a two-third of one percent.

We also consider two alternative models: the model of invariant bet frequency and the model of invariant bet size. These models do not have a natural concept of a time clock, making it unclear what their predictions about the arrival rate of news articles are. We make, however, assumptions about the information flow in both models, which are consistent with their general spirit. In the first alternative model, we assume that the same number of articles about firms arrive over a given period of time regardless of the level of their trading activity. In the second alternative model, we assume that the number of articles about firms is proportional to the number of bets placed. According to alternative models, the number of article is therefore either constant across stocks or increases proportionately with the trading activity.

The predictions of all three models are conveniently nested into one specification. We test them using the news data from Thomson Reuters firm for the period from 2003 through 2008. We run a number of empirical tests based on the log-linear regressions and the count data regressions with the arrival rate of news articles specified either as a Poisson process or as a negative binomial process. Our tests provide a strong evidence in favor of the model of trading game invariance. For the sample of all news and the negative binomial specification, for example, the power coefficient is estimated to be 0.68 with the standard errors of 0.024. The two alternative models, which predict the power coefficients of 0 and 1, are soundly rejected.

We find that the arrival of news articles is best described by the negative binomial model with the arrival rate μ being a function of the trading activity W,

$$\mu(W) = e^{\eta} \cdot \left(\frac{W}{W^{*}}\right)^{2/3} \cdot \tilde{G}, \tag{1}$$

where a constant $\eta = 1.97$ with the standard error of 0.068, and the Gamma variable $\tilde{G}$ has the mean of one and the variance of $\alpha = 2.11$ with the standard error of 0.238. The scaling constant $W_{*} = 40 \cdot 10^{6} \cdot 0.02$ 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 stock will be in the bottom of S&P 500. For the benchmark stock, there are, on average, about seven ($= e^{1.97}$) news articles released per month.

Since the over-dispersion parameter is statistically different from one (the level corresponding to the Poisson model), we conclude that the negative binomial model describes the arrival of the news articles much better than the Poisson model. Indeed, there is much more variation in the number of news articles across stocks than we would observe if the simple Poisson process with a non-stochastic arrival rate were the true model. The intuition is that small stocks either have one news article released per month or no news articles at all, although their true expected arrival rate may be a non-integer number between zero and one. This discreteness introduces additional variation in the data.

The negative binomial specification 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, (2) an additional component of variation in the stochastic Poisson arrival rate associated with otherwise unmodeled features, which are being captured by the Gamma distribution, and (3) the random variation in the actual number of Poisson events given the Poisson arrival rate determined by the particular level of the trading activity and the realization of a Gamma random variable. In our tests, we find that the variation unexplained by the model of market microstructure invariance can be related to the differences in the market capitalization, the book-to-market ratios, the past returns, and the square term of the trading activity.

Our estimates are quite stable across all seventy two months in the sample. There is, however, a structural break in the mid of the year 2005. In some sense, the model of market microstructure invariance works better before the structural break: The estimated coefficient is close to 2/3 before the change and slightly lower after it. The structural break corresponds to the changes in strategies of the news provider in response to demands of its clients requesting to broaden the coverage. Subsequently, the number of news articles has increased and the sensitivity of their arrival rate to the trading activity has slightly decreased. This would be observed if the news provider tend to send out news articles about small firms, even though these articles may have not much of the information content.

Some news articles contain more useful information than others. If a news article has several news stories, it is tagged with several news topics in the database. For example, if a news article talks about the downgrade of a firm's debt and the worsened forecasts of its earnings, it will be tagged twice. When news items are counted by news tags rather than news articles, we find the sensitivity coefficient of 0.71 with the standard error of 0.025. This estimate is slightly higher than the predicted 2/3. We conjecture that the difference arises because most news articles are linked to at least two news topics. The distribution of news tags, conditional on a news article reported, has more variation comparing to a Poisson model, because there will be too many cases of two or more news events and not enough cases of one news event.

For different news topics, the sensitivity coefficient ranges from 0.60 to 1.23, being the lowest for "corporate results" category and the highest for "major breaking news" category. The "corporate results" include all corporate financial results, tabular and textual reports, dividends, accounts, and annual reports. Some of these releases are reported at a regularly basis, regardless of the level of trading activity, thus pushing the sensitivity coefficient towards zero relative to the predicted 2/3. In contrast, stories carrying the "major breaking news category" code would be expected to dominate the financial and general headlines of the worlds major newspapers, Web sites, television and radio networks. They will be disproportionately dominated by news articles about large firms, being of interest to a wide audience, thus pushing the sensitivity coefficient upwards relative to the predicted 2/3.

Several papers have tested the predictions of the model of market microstructure invariance with regard to the trading data. For example, Kyle and Obizhaeva (2010) document a supportive evidence for its predictions concerning the distribution of order sizes, the price impact and the spread using the sample of portfolio transitions. Kyle, Obizhaeva and Tuzun (2010) implement tests based on the transactions in Trades and Quotes dataset. This paper suggests that not only the trading process unfolds in a trading-game time but the information flow conforms to the same time clock as well. This finding provides a natural answer on the fundamental question about the role of time in financial markets, discussed in the work of Mandelbrot and Taylor (1967), Clarke (1973), and Hasbrouck (1999). It is also important for making the model of market microstructure invariance internally consistent.

Earlier papers, Berry and Howe (1994) and Mitchell and Mulherin (1994), studied the relationship between the public information flow measured by the number of news releases and the market activity for the aggregate market. They suggest a small positive relationship between public information and trading volume as well as an insignificant relationship between public information and price volatility. In contrast, our paper shows a strong cross-sectional relationship between these variables, possibly because we combine the volume and the volatility into one measure of the trading activity.

Recently, a growing literature has been devoted to studying how different measures of trading activity, such as volume, volatility and returns, change around various news events for individual stocks. The examples include the analysis of the stock messages on the Internet boards in Antweiler and Frank (2004), the economic news announcements in Green (2005), the CEO interviews on CNBC in Mescke (2004), the information in the WSJ column in Tetlock (2007), the corporate announcements in Chae (2005) as well as the data in the Dow Jones news archives in Chan (2003), Tetlock, Saar-Tsechansky, and Macskassy (2008), and Tetlock (2010). These papers discuss interesting patterns of a general form in the news data. In contrast, our paper examines a specific quantitative prediction for the relationship between the number of news articles and the trading activity.

The remainder of the paper states the implications of the market microstructure invariance 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 Market Microstructure Invariance To News Data

In Kyle and Obizhaeva (2010), traders are thought as playing trading games. They arrive to the market and place "bets," defined as statistically independent decisions to trade specific quantities. The trading games have to satisfy two irrelevance principles: the Modigliani-Miller irrelevance and the Time-Clock irrelevance.

The Modigliani-Miller irrelevance refers to the idea that the fundamentals of the games are not affected by stock splits or changes in leverage. The Time-Clock irrelevance refers to the idea that the fundamentals of the games do not change when the time clock is sped up or slowed down.

How do these transformations affect the information structure of the game? Splits and changes in leverage should not influence the information structure of the game, since traders can always undo these changes on their own accounts. Speeding up the clock, however, do change the amount of information that arrives in a given time period. Suppose that μ^* articles about a firm arrive per calendar day. Consider changing a time clock so that one day is changed into H days. If H $>$ 1, the time clock has been slowed down; if H $<$ 1, the time clock has been sped up. It is reasonable to assume that the same amount of information and news articles arrives now during H days as before during one calendar day. This implies that the arrival rate of news articles μ per calendar day changes by a factor 1/H,

$$\mu = \mu^* \cdot H^{-1}. \tag{2}$$

In some sense, if we think about a natural "trading game" time, then the amount of information that arrives per clock tick remains unaffected by irrelevance transformations. The information flow is one of the deep parameters of the trading game.

We can convert these theoretical irrelevance principles into empirically testable hypotheses by making two empirical hypotheses:

- Information Flow Invariance: (Public and private) information arrives at a rate proportional to the rate at which the time clock ticks, with a proportionality constant being the same across stocks and across time for the same stock.

- News Article Invariance: News articles arrive at a rate proportional to the rate at which information arrives, with a proportionality constant being the same across stocks and across time for the same stock.

The proportionality constants are examples of the market microstructure invariants. These hypothesis are parallel to the hypotheses of Trading Game Invariance, Market Impact Invariance, and Bid-Ask Spread Invariance, proposed in Kyle and Obizhaeva (2010).

At first glance, it may appear unlikely that the amount of public news is effectively proportional to the amount of private information, but there are good reasons to believe so. First, private information may arise due to the manner in which public information is processed. Second, news reporters may write articles about the same firms for which traders are starting to acquire private information.

The speed of the trading game H is unobservable. To formulate the testable implications, we use the intuition from Kyle and Obizhaeva (2010) saying that this speed is related to the measure of "trading activity" W, defined as the product of the dollar trading volume and volatility, which represents the risk transferred through trading during a calendar day,

$$W = V \cdot P \cdot \sigma_r, \tag{3}$$

where V is the daily trading volume (in shares), P is the stock price, and σ_r is the standard deviation of daily returns. Since all three variables are observable, the trading activity W is easy to measure.

The trading activity is unaffected by the Modigliani-Miller transformations but it is affected by the time-clock transformation. Suppose a stock has the trading activity W^*. Consider changing a time clock so that one hour is changed into H hours. Speeding up the time clock (H<1) affects trading activity in two ways. First, there is the "volume effect" - the number of bets per day and therefore the dollar volume increase proportionately with 1/H. Second, there is the "volatility effect" - returns variance increases proportionately with 1/H, so the volatility (the square root of variance) increases proportionately with $1/H^{1/2}$, making each bet riskier. Since the trading activity is defined as the product of dollar volume and volatility, then combining both effects, the trading activity W has to be related to H as follows,

$$W = W^* \cdot H^{-2/3}. \tag{4}$$

This equation shows that the ratio of trading activities reveals how much the time clock of one security is ticking faster than the time clock of another,

$$H = \left(\frac{W}{W^*}\right)^{-2/3}. \tag{5}$$

Plugging (5) into (2), we obtain the relationship between the arrival rates of news articles μ and the measures of trading activity W,

$$\mu = \mu^* \cdot \left(\frac{W}{W^*}\right)^{2/3}. \tag{6}$$

Equation (6) identifies the main relationship we test in this paper. The equation states that the elasticity of the arrival rate of news articles with respect to trading activity W is two-thirds: A one percent increase in trading activity will be accompanied by a two-thirds of one percent increase in the arrival rate of news articles.

Let us consider a simple numerical example. Suppose that the time is sped up and $H = 1/2$. The information flow speeds up: The analysts faster type their reports, the journalists publish more articles, the news service providers release more news items, and more news messages appear on the screens of traders. The same amount of information that used to arrive in a day now comes in half a day. The number of news articles released per day increases 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 $\sqrt{2}$. The trading activity, being the product of dollar volume and volatility, increases by a factor of $2^{\frac{3}{2}}$. The trading activity and the number of news therefore satisfy (6). Note that the particular dollar volume factor of 2 and the standard deviation factor of $\sqrt{2}$ are irrelevant, because both are affected by the Modigliani-Miller transformations; the economically important number is their product $2^{\frac{3}{2}}$. This is why all predictions are stated in terms of trading activity.

In the Model of Trading Game Invariance, the time clock is given by the rate at which trading games are being played. This rate naturally determines not only the rate at which bets arrive to the market place but also the rate at which information flow unfolds, including the arrival of news articles.

Alternative Models. Kyle and Obizhaeva (2010) consider two alternative models: Model of Invariant Bet Frequency and Model of Invariant Bet Size. These models do not have a well-defined concept of a time clock, making unclear what predictions they deliver about the arrival rate of news articles. We contemplate the assumptions about the information flow in both models, which are consistent with their general flavor and allow us to generate testable implications.

The model of Invariant Bet Frequency assumes that the variation in trading activity comes entirely from variation in bet sizes, while the number of bets over a calendar day remains invariant across stocks. In a spirit of this model, we assume that the number of news article arrived per calendar day is invariant across stocks as well. Bets are larger for more active stocks, because the news articles about them have more valuable information, allowing traders to place larger bets upon reading these articles. This assumption implies a testable prediction concerning how the number of news article should varies with trading activity,

$$\mu = \mu^* \cdot \left(\frac{W}{W^*}\right)^0. \tag{7}$$

The model of Invariant Bet Size assumes that the variation in trading activity comes entirely from variation in the number of bets placed over a calendar 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 article varies across stocks proportionally to the number of bets. Each news article leads to a certain number of bets, similar in their size. This assumption implies a testable prediction concerning how the number of news article should varies with trading activity,

$$\mu = \mu^* \cdot \left(\frac{W}{W^*}\right)^1. \tag{8}$$

Of course, since there is no well-defined concept of a time clock in these alternative models, we choose our assumptions about the information flow in a somewhat ad hoc way, as many other assumptions could have been made as well. The assumptions are ultimately related to the assumptions about how information is being processed institutionally.

Indeed, in the first alternative model, we assume that there is the same number of news articles per day across stocks. The model could, however, have a different interpretation. For example, actively traded stocks are usually traded by large financial institutions, which often internalize their decision process: The news from several firm's units, which deal with different aspects of the market, are collectively analyzed and result in only one trading decision. Thus, while having the same number of bets placed over a calendar day, stocks may have different number of news articles issued about them during the same period.

Or, in the second alternative model, we assume that the number of news articles vary across stocks. This model, however, could incorporate the idea that many traders are active in actively traded stocks. These traders tend to disagree with each other about how to interpret news articles and place independent bets based on their own views. Thus, stocks may differ in the number of independent bets actually placed into the marketplace, even if the same number of news articles arrive per calendar day.

All three models imply specific relations between the number of news articles μ on one side and the measure of trading activity W on the other side. They can be conveniently nested into one specification,

$$\mu = \mu^* \cdot \left(\frac{W}{W^*}\right)^\gamma. \tag{9}$$

The three models differ only in their predictions about a coefficient γ. The model of trading game invariance predicts that $\gamma = 2/3$, the model of invariant bet frequency predicts that $\gamma = 0$, and the model of invariant bet size predicts that $\gamma = 1$. We discuss next how to test which of the three models more accurately describes the news data, with the model of trading game invariance being, not surprisingly, our preferable choice.

3 Data

The news data, collected in NewsScope dataset, is provided by the Thomson Reuters firm. The sample covers all news articles sent by the news service provider to its clients from

January 2003 through December 2008. Each news items has the following fields: the time stamp, the ticker of a company, the relevance that measures how substantive the news item is for the company, the sentiment that indicates a prevailing tone of the news item, the probabilities of the news item being positive, negative, or neutral providing a more granular sentiment, the news item type (alert, article, update, or correction), the headline, the linked counts showing the number of repetitions over different time periods, and the topic code describing what the news item is about. One news item can be linked to several firms and tagged with multiple topic codes. The news dateset is matched with the CRSP data on daily returns, prices, and daily volume for common stocks traded on the NYSE, the Amex and the NASDAQ.

The rate at which information arrives to the marketplace is hard to quantify. We want to use the number of news articles shown on the screens of traders as a proxy for the arrival rate of the information about firms. We first need to identify new information about each firm. Thomson Reuters often sends a one-line alert about an important news article before the news article appears; we omit all alert messages from our analysis. We also exclude updates and corrections, since they usually do not carry new information but rather make some parts of initial article more specific. The linked counts may also indicate the novelty of a news item by showing how many times similar news article has appeared in the sample before; we exclude news items linked to more than two previous articles.

One news item can be associated with several firms. Of course, news items are not relevant for each of mentioned firms. For example, large companies are often mentioned in the news articles about small companies, just in a context of a general description of an industry in which both of them operate. These news articles are obviously irrelevant for the actual information flow about large companies. Thomson Reuters provides a relevance parameter associated with each pair of a news item and a firm. This parameter ranges from zero to one, with one indicating that the news item is highly relevant for a particular firm. We include the news item in the sample only if its relevance for a given firm is greater than 0.35.

One news item can have information on the multiple dimensions of a firm. If such a news item is counted only once, we will potentially underestimate the amount of actual information it reveals. The same news item can be tagged by Thomson Reuters to several topic codes, for example, it can mention an earnings announcement, a earnings forecast, and a merger announcement. We therefore 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, i.e., we count each news tag separately. The news item in the example above will be counted as one observation in the first sample and three observations in the second sample. We exclude news tags about an industry or about the firm, when it is not matched to a specific ticker symbol.

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, all new listings, delistings and suspensions; it is tagged to 15% of news article. The topic code 'RES' indicates all corporate financial results, tabular and textual reports, dividends, annual and quarterly reports; it is tagged to 14% of news articles. The topic code 'MRG' indicates mergers and acquisitions; it is tagged to 12% of news articles. Most topic codes indicate economic news. For example, the topic code 'DBT' tags news articles related to debt market, 'RESF' tags news indicates results of corporate financial results, 'CORA' and 'RCH' tag analysis of a company by a journalist and a broker, respectively. Other topic tags indicate behavior news. For example, the topic code 'HOT' tags news articles about stocks that are on move and the topic code 'NEWS' tags news articles that are likely to lead to television or radio bulletins or to make the front pages of major international newspapers and web sites.

We construct two proxies for the arrival rate of information μ. We calculate the number of news articles and the number of news tags about a given firm released in a given month. In total, there are about 1.4 million news articles in the database and about 3.4 million news tags mentioning a particular topic code.

We also consider two samples of observations. One sample consists of firms covered by Thomson Reuters from the instance we observe the first news article about a firm. If the arrival rate of information is so slow that a firm does not have any news in some of subsequent months, we count the number of news articles and news tags as zeros. Of course, the Thomson Reuters's decision to cover particular firms can be endogenous, i.e., the small firms with a few news articles can be left out of the sample. This can introduce a selection bias into our estimates, because the small firms will appear to have "too many" news articles. To deal with this concern, we also implement our tests on the sample of all firms recorded in the CRSP from 2003 through 2008.

We have 275,059 firm-month observations in our sample of news articles, resulting from at least one match between a firm and a news article topic code. The observations are spread over 72 months. The coverage has increased over time and converged to almost 100% by year 2006, as the firm has responded to the requests of its clients demanding 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.

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 daily 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 NYSE with CRSP share codes of 10 and 11. Group 1 contains stocks in the bottom 30^{th} percentile by dollar trading volume. Group 10 contains stocks in the top 5^{th} percentile. It 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 the most 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 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 the trading activity, being a product of trading volume and volatility, varies by a factor of 315 between inactive and active securities.

Panel B of Table 1 reports the statistics for the number of news articles in 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 model of market microstructure invariance, 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 can be explained by 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. This significant variation reveals that releases of news articles about a given firm tend to cluster in time. For example, inactive stocks get no attention during most months, but when something happens - e.g., 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. The estimation procedures have to be adjusted for a excessive variation in the news arrival rate due to the news clustering. We will see, for instance, that a negative binomial model provides a better fit for the news data than a Poisson model.

Similar conclusions can be drawn from the statistics on the fraction of firms with no news articles during a given month. For high-volume stocks, only 5% of firms do not have any news articles during a given month; note that 2.5% of firm-month pairs, (7.143-6,947) out of 7,143 pairs, are not covered by Thomson-Reuters at all, with the other 2.5% of firms having no news articles reported. For low-volume stocks, 73% of firms do not have any news articles during a given month; note that 25% of firm-month pairs, (222,543-166,679) out of 222,543 pairs, are not covered by Thomson-Reuters at all, with almost half of the covered firms having no news articles reported. For the aggregate sample, about 58% of firms have no news articles.

The data clearly exhibits over-dispersion relative to a Poisson model, with too many zeros in the sample. Indeed, if a Poisson model were a correct model, then the fraction of firms with no news could be calculated as $e^{-\mu}$, where μ is the average number of news per month, reported 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 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 the selection of a good model for the news arrival rate, suggesting that a negative binomial model, which allows to correct for over-dispersion, can be a better choice than a Poisson model.

Each news articles can be tagged with several news topics. In the table, all statistics for the news tags is about twice bigger than statistics for the news articles. This implies that one news articles is usually tagged to two news topics. 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 may follow a distribution that is different from a Poisson. Indeed, the number of observations with only one news tag per month will be very small, since usually there will be either no news articles about a given firm at all or there will be one news articles with two news tags attached.

The 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. All observations are splitted into the twelve news 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. The distributions are constructed either based on the number of news articles per months in dark blue or based on the number of news tags per month in light blue. All observations are pooled together.

The figure shows that the distributions are very different across active and inactive stocks. Among inactive stocks in the lowest volume group, 73% have no news articles, 17% are mentioned in one article, and 6% are mentioned in two articles. Among active stocks in the two highest volume groups, 6% have no news articles, 1% are mentioned in one news articles, 1% are mentioned in two articles, and the remaining observations are spread over higher news bins. The density is the biggest in the news bin "seven" indicating that actively traded stocks are typically mentioned in 17 to 32 news articles per month. The distribution based on the news tags is similar but shifted to the right. Also, the densities of the number of news articles and news tags are, by definition, identical in the no-news bin.

In this paper, we examine whether the model of market microstructure invariance can explain the cross-sectional differences in the distribution of the number of news articles, shown in figure 1.

4 Estimation Procedures

We test the three models by analyzing the relationship between the trading activity W and the arrival rate of information μ, proxied either by the number of news articles about a stock or the number of linked news tags in a given month. 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 a stock's average daily dollar volume and volatility, calculated using the CRSP data. The variable $N_{t,i}$ is a count variable calculated using the news data; it may have zero values.

We implement three estimation approaches: a log-linear model, a Poisson model, and a negative binomial model.

- **Log-linear model:** The simplest approach is to estimate the log-linear model for the average number of news per stock with the trading activity being an explanatory variable. The problem is that for many observations, the number of news is equal to zero, since many firms often do not generate any news, thus making the logarithm of the number of news being infinite. To avoid this problem, we look at 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 articles. We then calculate the average number of news and the average trading activity per firm in each group. 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 $W_{t,j}$ in each group j and month t,

$$\ln \bar{N}^{*}_{t,j} = \eta + \gamma \cdot \ln \big[\frac{\bar{W}_{t,j}}{W^{*}}\big] + \epsilon_{t,j}, \tag{10}$$

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. We explain below why we need to adjust the average number of news articles $\bar{N}_{t,j}$ for the within-group variation in the trading activity and how to do this adjustment.

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

  $$f(N_{t,i}|W_{t,i}) = \frac{e^{-\mu(W_{t,i})} \times \mu(W_{t,i})^{N_{t,i}}}{N_{t,i}!}, \tag{11}$$

  where the arrival rate of news items $\mu_{t,i}$ is a non-linear function of the trading activity $W_{t,i}$,

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

  The Poisson model assumes that the arrival rate is a non-stochastic function of the trading activity, i.e., all variation in arrival rates occurs within the context of the Poisson distribution. From the properties of the Poisson distribution, we know that $\mu(W_{t,i}) = \mathrm{V}(N_{t,i}|W_{t,i}) = \mathrm{E}(N_{t,i}|W_{t,i})$. The Poisson model assumes that stocks with the same level of trading activity have the same expected number of news items $\mu(W)$ and the same variance equal to $\mu(W)$. Our discussion of the descriptive statistics suggests that these assumptions may be too restrictive, because the news data seem to exhibit over-dispersion, with the variance of the arrival rate 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. We model this variation with a continuous mixture of the Poisson distributions where the mixing distribution of the Poisson rate is modeled as the gamma distribution,

  $$\mu(W_{t,i}) = e^{\eta + \gamma \cdot \ln \left[\frac{\bar{W}_{t,j}}{W^{*}}\right]} \cdot \tilde{G}_{t,i}(\alpha). \tag{13}$$

  The Gamma variable $\tilde{G}_{t,i}$ has the mean of $\kappa \cdot \theta$ and the variance of $\kappa \cdot \theta^{2}$. It makes sense to impose the restrictions $\kappa = 1/\alpha$ and $\theta = \alpha$, thus restricting the mean of the Gamma variable to be equal to one and its variance to be equal to α, with $\alpha = k\theta^{2} = \theta$. The model parameter η therefore identifies the same mean as the mean in the Poisson model above. 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 with $\alpha = 0$. 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, varying across stocks too much to be explained by a simple Poisson model, then the negative binomial specification will be a reasonable model to use for describing the news data.

The log-linear model with data in bins 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 zero news items in the same month. 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, which are being captured by the Gamma distribution, and (3) the random variation in the actual number of Poisson events given the Poisson arrival rate determined by the particular level of the trading activity and the realization of a Gamma random variable. For negative binomial specification, the Poisson arrival rate varies randomly according to the realization from the Gamma distribution, even if two firms have with the same trading activity. When we restrict the over-dispersion parameter α to be equal to zero, we get the Poisson specification that does not have the second source of uncertainty: The Poisson arrival rate is a non-stochastic function of the trading activity.

In all specifications, we scale the explanatory variable so that a constant term e^{η} quantifies the average number of news arrived monthly about the benchmark stock with the trading activity $W_* = (40)(10^6)(0.02)$. The benchmark stock is selected arbitrarily as a stock with price \$40 per share, trading volume of one million shares per day, and volatility of 0.02. This stock would be at the bottom of S&P 100.

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

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

with $\tilde{Z}_{t,i}$ being 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 get,

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

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

This equation suggests that we can not simply regress $\ln E\bar{N}_{t,j}$ on $\ln \bar{W}_{t,j}$, but we need to adjust the average number of news items for the potential variation in the trading activity within a group, captured by the last term. The adjustment term is always positive. Moreover, since the adjustment term can be more significant for groups with lower trading activity, this can introduce the bias into our estimates. Rather, we calculate the adjusted average number of news $\bar{N}^*_{t,j}$ for group j and month t as,

$$\ln \bar{N}^*_{t,j} = \ln \bar{N}_{t,j} - \ln\Big(\frac{1}{M_{t,j}}\sum_{i=1}^{M_{t,j}} e^{2/3(\ln W_{t,i} - \ln \bar{W}_{t,j})}\Big), \tag{14}$$

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 for the averages. It is not necessary to implement this adjustment for the count data regressions, for which we use the actual news data, not the averages.

We implement the empirical tests of the three models by estimating a coefficient γ and testing whether $\gamma = 2/3$ as predicted by the model of trading game invariance, $\gamma = 0$ 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 in SAS for each of 72 months and then averaging the estimates across months. We also correct the standard errors using the Newey-West procedure with three lags. Since this approach does not require to specify a particular form of interdependencies between residuals, it is a reasonable estimation methodology which we implement in our tests.

5 Results

In this section, we discuss the results of the tests described in the previous section. We report results based on the log-linear models with data sorted into bins as well as results based on the count models.

5.1 Log-Linear Models with Data in Bins

Each month, as discussed, we sort all stocks based on their trading activity into 30 groups in an ascending order, 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, $\ln \bar{N}^*_{t,j}$, 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. Six subplots contain observations for each of six years from year 2003 through year 2008. For each month, there are 30 points for 30 groups. For each group, there are twelve points in a year. Each subplot therefore has $30 \cdot 12$ points. The observations from the lowest group form a distinctive set of twelve points on the left side of each plot. As the trading activity increases, all monthly observations form tighter groups on the plots. This justifies our intuition that the within-group variation in the trading activity is the biggest for the lowest group. For the convenience of comparison, the same fitted line with a slope fixed at 2/3 and the intercept -6.74, estimated over the entire sample, is superimposed on each plot.

The scatter plots shows that the data exhibit patterns similar to those predicted by the model of trading game invariance. The observations pile up around the fitted line with a slope of 2/3. 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 very inactive stocks might be explained by the strategy of the news services provider to expand the coverage and cover all firms in the economy, the strategy which became especially important after year 2005. Consistently with this explanation, the observations are closer to the fitted lines before year 2005. "Too many" news articles for very active stocks might be explained by a large number of news article simply referring to that stocks as "hot stocks," rather than carrying new information content. The "smile" suggests that, in addition to a linear term of the trading activity in our log-linear regressions, its quadratic term may add some explanatory power as well.

Table 3 shows the estimates of the intercept η and the slope γ from the regression equation (10). The table presents four different sets of estimates. We consider the sample or all firms and the sample of firms covered by Thomson Reuters. For each sample, we estimate two regressions using two different proxies for the arrival rate of information, the number of news articles and the number of news tags.

In the four versions of the regression, the coefficient γ ranges from 0.65 to 0.75, being economically close to 2/3 predicted by the model of trading game invariance and economically very different from 0 and 1 predicted by the alternative models. The values of F-tests for the hypothesis $\gamma = 2/3$ varies from 0.03 to 0.79, indicating that the hypothesis $\gamma = 2/3$ can not be rejected, while F-tests reject the alternative models. The intercept η is lower for the sample of all firms than for the sample of firms covered by Thomson Reuters, since these two samples differ only by a set of firms with no news articles reported. Also, the intercept for the number of news tags is bigger than the intercept for the number of news articles, since there are more news tags than news articles, as each news article can correspond to several news tags but not vice versa.

The average R^2 ranges from 0.893 to 0.917. While this indicates that the linear specification explains most of the variation in the average number of news items across bins, the "smile" in figure 2 suggests that much of the unexplained variance could be captured by including a quadratic term in the regression.

5.2 Count Data Models

The count data models are a natural choice for modeling the arrival rate of news items. Table 4 reports the estimates of the intercept η and the slope γ for the count data specification (12) and (13). The estimates are reported for the sample of all firms and firms covered by Thomson Reuters as well as for the number of news articles and the number of news tags.

The estimates based on the number of news articles for the sample of all firms, our main sample, are in columns one and two of table 4. There are four important facts to note about these estimates. First, the negative binomial model estimate of $\gamma = 0.68$ is almost identical to the estimate from the 30-bin model in table 3. This result is consistent with the intuition that the log-linear 30-bin model is an approximately unbiased way to model how news arrival rates vary with the trading activity. Second, the Poisson estimate of $\gamma = 0.81$ is much larger than the negative binomial estimate of $\gamma = 0.68$. This result indicates that the Poisson model produces biased estimates of the mean arrival rate of news events when expected arrival rates are not constant in a true model. Third, the Newey-West standard error for γ is 0.024. This standard error is sufficiently large that the hypothesis $\gamma = 2/3$ is not rejected. It is, however, sufficiently small that the hypotheses of the alternative models, $\gamma = 0$ and $\gamma = 1$, are soundly rejected. Fourth, the Newey-West standard error of α or 0.218 is almost ten times smaller than the coefficient estimate $\alpha = 2.05$ itself, indicating a strong statistical support for the negative binomial model over the Poisson model.

In table 4, columns three and four present the results of the Poisson and negative binomial models when the subset of firms is restricted to those covered by Thomson Reuters firm. The results have a flavor very similar to columns one and two. The estimate $\gamma = 0.65$ from the negative binomial model is again almost identical to the coefficient estimate in the 30-bin model from table 3. The Poisson estimate of $\gamma = 0.86$ is much higher than the estimate of the negative binomial model, indicating a bias in the Poisson model. The standard errors are still sufficiently large that the hypothesis $\gamma = \frac{2}{3}$ is not rejected, but sufficiently small that the alternative models are soundly rejected. The standard error of 0.120 on the coefficient estimate $\alpha = 1.63$ is so small that the model provides a strong statistical support for the negative binomial model over the Poisson model.

The last four columns in table 4 provide the estimates for both the Poisson and negative binomial models when the news tags are counted, i.e., news items are counted by the number of tagged topic codes, both for the dataset including "all firms" and the dataset including only firms in the "Thomson Reuters" universe. The coefficient estimates of the negative binomial, $\gamma = 0.71$ for "all firms" and $\gamma = 0.66$ for "Thomson Reuters" firms, are lower than the corresponding estimates of $\gamma = 0.75$ and $\gamma = 0.70$ for the 30-bin model in table 3. We conjecture that this difference arises because, conditional on a news item occurring, the count of topics linked to that news item may follow a distribution that is different from a Poisson distribution.

For example, consider firms for which the Poisson arrival rate is very low. Such firms are likely to receive zero news articles in a month, but occasionally they receive one news and only rarely more than one. 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 as 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. To examine how well the models fit historical data, 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. To investigate in more detail how well models fit the news data across stocks, the three subplots depict the results for three sub-samples: the inactive stocks from the volume group one, the medium stocks from the volume groups two through eight, and the active stocks from the volume groups nine and ten. The charts for the number of news articles for a firm in a month are in dark color and the charts for the number of news tags are in light color.

The bottom panels show the difference between the empirical frequencies and the fitted frequencies from the Poisson model. The positive values in the first no-news bin indicate that the news data have the excessive number of no-news observations comparing to the Poisson model. The news data is over-dispersed relative to that model. The top panels show differences in the empirical frequencies and frequencies from the calibrated negative binomial model. The residuals in the top panels are clearly much smaller than residuals in the bottom panels, suggesting that the negative binomial model is a better fit. We thus implement most of our subsequent tests only for the negative binomial model.

The negative binomial model explains well the news data for most stocks, except for those in the two highest volume groups. Intuitively, our estimation procedures tend to fit the models to match observations for stocks in the lower volume groups, since these groups have a bigger number of observations. For active stocks in volume groups nine and ten, the calibrated negative binomial and the Poisson models fit the news data less accurately.

The above results for news articles are similar to the results for news tags. Note that the negative binomial model overestimates the number of observations with one news tag in a month, as indicated by its negative residual in the second bin. It seems that there are too few cases of one news event, since most news articles have more than one topic code assigned to them. Even though the negative binomial model accurately describes the data on the number of news articles, the data on the number of news tags have a more complicated distribution.

A Comparison of Three Models. Figure 4 shows how accurately the empirical data is described by each of the three models. We fix the parameter γ at $2/3$ for the model of trading game invariance, 0 for the model of invariant bet frequency, and 1 for the model of invariant bet size. We then estimate the parameters η and α in the negative binomial model with the number of news articles (13). The figure shows the difference between the empirical distributions and the fitted distributions for the twelve news bins, with the standard errors calculated using the bootstrap procedure.

The magnitudes of residuals indicate that the model of trading game invariance describes the empirical data reasonably well, and also better than the two alternatives. For example, the model of invariant bet frequency assumes that the same number of news items is released per month about both inactive and active stocks. For inactive stocks, this model underestimates the fraction of months with only few news items, as indicated by positive residuals in the lower news bins, and overestimates the fraction of months with many news items, as indicated by negative residuals in the higher bins. For active stocks, in contrast, this model significantly overestimates the fraction of months with few news items and underestimates the fraction of months with many news items, as shown by large negative values in the lower bins and large positive values in the higher bins.

The second alternative model certainly explains the data better than the first alternative model, but worse than our preferable model, if we look at the panels for stocks with low and medium trading volume. For stocks with high trading volume, however, the residuals of two models are somewhat comparable. This is consistent with the convex patterns in figure 2, where the information flow accelerating with the trading activity in the high-volume bins faster than predicted by the model of trading game invariance.

The Model Estimation. Table 5 provides the estimates for the negative binomial model (13) with the coefficient γ fixed at 2/3. The estimates of the intercept η and the over-dispersion α are reported for the sample of all firms and firms covered by Thomson Reuters with information flow proxied by the number of news articles and news tags. Fixing $\gamma = 2/3$ only slightly reduces the log-likelihood function comparing to the unrestricted specification in table 4. For our main sample of news articles and all firms, for example, the log-likelihood decreases from -7,170 to -7,216. We interpret these result as implying that the negative binomial model with the expected arrival rate modeled according to the model of trading game invariance provides a good description for the news data.

In our tests for the number of news articles, the intercept $\eta = 1.97$ with the standard error of 0.068 implies that, on average, seven news articles are released per month about a benchmark stock. The over-dispersion $\alpha = 2.11$ with the standard error 0.238 indicates a more variation in the arrival rate of news articles comparing to the Poisson model.

The Dynamics of Estimates from Count Regressions. Figure 5 shows the estimates from the negative binomial specification (13) for each month from year 2003 through year 2008. As before, we consider the sample of all firms and the Thomson Reuters universe. We also obtain monthly estimates for the number of news articles and the number of news tags. The figure reveals an interesting dynamics. There is a clear structural break in year 2005, which can be attributed to the change in the corporate strategy of the news provider, when Thomson Reuters has increased its coverage upon the requests of its clients

The estimates of the intercept η quantifies the average number of news items in a month for a benchmark stock. It is relatively stable except for a permanent jump in year 2005. The estimate is slightly higher for the Thomson Reuters sample than for the sample of all firms, but this small difference disappeared after the year 2005 when these two samples became almost identical. The average number of news articles for a benchmark stock is about 7.77

articles per month ($= e^{2.05}$). The average number of news tags is about 14.88 news tags per month ($= e^{2.70}$). This shows that each news article is usually tagged to two topic codes, consistently with our discussion above.

The estimate of the slope γ quantifies the sensitivity of the number of news items to the trading activity. It fluctuates around 2/3 during all six years, as predicted by the model of trading game invariance. All four estimates of γ are identical after the year 2005 but slightly lower than 2/3. This lower sensitivity might be explained by the recent tendency of the news provider to sent out more articles about small firms to meet its goals of the global coverage, even though these articles may have not much of the information content. In some sense, the model of trading game invariance better explains the data before the structural break.

The estimate of the over-dispersion parameter α shows whether the Poisson model is a reasonable approximation or the more general negative binomial model has to be used. The estimates of α is above zero, indicating the over-dispersion and suggesting that the negative binomial model is a better description of the data than the restrictive Poisson model.

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

$$\mu(W_{t,i}, V_{t,i}, P_{t,i}, \sigma_{r,t,i}) = e^{\eta+\frac{2}{3}\cdot\ln\left[\frac{W_{t,j}}{W^*}\right]+\beta_1\cdot\ln\left[\frac{V_{t,i}}{(10^6)}\right]+\beta_2\cdot\ln\left[\frac{P_{t,i}}{(40)}\right]+\beta_3\cdot\ln\left[\frac{\sigma_{r,t,i}}{(0.02)}\right]} \cdot \tilde{G}_{t,i}(\kappa, \theta). \tag{15}$$

This regression imposes the restriction that the coefficient $\gamma = 2/3$ as predicted by the model of trading game invariance. It then allows the coefficient on the three components of $W_{t,i}$ to vary freely. Since the model of trading game invariance suggests that all variation in the arrival rate of information will be 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 $\beta_1 = \beta_2 = \beta_3 = 1/3$.

The estimates are statistically different from zero. The F-tests reject the model of trading game invariance, but they reject alternative hypotheses with much bigger F values. This suggests that other factors influence the information flow in addition to the trading activity. These factors might be correlated with volume, price, and volatility, but the multicollinearity between these variables put a hurdle in interpreting our results. Although the three variables are statistically significant in explaining the variation in the number of news items, the increase in the value of the log-likelihood function relative to the univariate regressions in table 4 is modest.

The Robustness Check: A Quadratic Term. In Panel A of table 7, we estimate a convexity effect of the trading activity on the news arrival rate by adding a quadratic term to the negative binomial model of table 4. The table presents the estimates for the count regression models with the news arrival rate modeled as,

$$\mu(W_{t,i}, V_{t,i}, P_{t,i}, \sigma_{r,t,i}) = e^{\eta+\gamma_1\cdot\left(\ln\left[\frac{W_{t,j}}{W^*}\right]-\mu_W\right)+\gamma_2\cdot\left((\ln\left[\frac{W_{t,j}}{W^*}\right]-\mu_W)^2-\sigma_W^2\right)} \cdot \tilde{G}_{t,i}(\alpha). \tag{16}$$

Here, μ_W is the sample mean of $\ln\left[\frac{W}{W^*}\right]$ and σ_W is the sample standard deviation of $\ln\left[\frac{W}{W^*}\right]$. To avoid a multi-collinearity, the level of trading activity is captured by $\ln\left[\frac{W}{W^*}\right] - \mu_W$ and the quadratic term is modeled as $(\ln\left[\frac{W}{W^*}\right] - mu_W)^2 - \sigma_W^2$. This procedure ensures that all covariates are orthogonal to each others. We constrain the first-order effect γ_1 to be equal to 2/3 and then estimate the second-order effect γ_2. The coefficient γ_2 is between 0.03 and 0.04, being statistically significant since its standard error is equal to roughly one-tenth of the estimate itself. The significance of the quadratic term is consistent with the non-linear patterns in figure 2.

In Panel B of table 7, we estimate simultaneously both the quadratic effect γ_2 and the first-order effect γ_1 of the trading activity on the news arrival rate in equation (16). The estimate of γ_2 is equal to 0.04. The estimates of γ_1 ranges from 0.58 to 0.65. If the news articles are used as proxies for information flow, for example, the estimate for γ_1 is 0.61; if the new tags are used as proxies, then the estimate for γ_1 is 0.65. The inclusion of the quadratic term reduces the estimate of γ_1 comparing to the estimates reported earlier, but this coefficient is still close to 2/3 predicted by the model of trading game invariance. The log-likelihood functions also do not improve substantially comparing to models in table 4 with no quadratic term.

The Robustness Check: Market Capitalization, B/M ratio, and Past Returns. Several other factors may potentially affect how many news items are published about a firm. Large firms are likely to appear more frequently in the news articles than would have been explained by their levels of trading activity. Indeed, large firms are usually used in the news articles just as placeholders for smaller firms. For example, a story about a small technology firm often mentions other technology heavyweights like Intel, Apple, and Microsoft, but the news story does not carry any new information about these large firms. Growth stocks may be mentioned in news stories more often as well, because they have a higher recognition among the readers, and journalists certainly cater to their interests. Finally, if the number of articles adjusts to change in the level of trading activity with a lag, then it will appear that too many news items are being released about firms with negative past performance and therefore slowed down trading activity.

We next examine the dependence of information flow on the market capitalization, the book-to-market ratio, and the firm's past return. Table 8 reports the Fama-MacBeth estimates from the monthly negative binomial regressions with the arrival rate of news items modeled as,

$$\mu(W_{t,i}, M_{t,i}, B_{t,i}/M_{t,i}, R_{t,i}) = e^{\eta + \frac{2}{3}\cdot\ln\left[\frac{W_{t,j}}{W^*}\right] + \beta_4\cdot\ln\left[M_{t,i}\right] + \beta_5\cdot\ln\left[B_{t,i}/M_{t,i}\right] + \beta_6\cdot\ln\left[R_{t,i}\right]} \cdot \tilde{G}_{t,i}(\alpha), \quad (17)$$

where $M_{t,i}$ is the previous month's market capitalization of stock i in month t, $B_{t,i}/M_{t,i}$ is its book-to-market ratio, and $R_{t,i}$ is its return over the last year. This regression imposes the restriction that the coefficient $\gamma = 2/3$ as predicted by the model of trading game invariance. Since the model of trading game invariance suggests that all variation in arrival rate of information will be captured by variation in the trading activity, it predicts that $\beta_4 = \beta_5 = \beta_6 = 0$.

The estimates are statistically different from zero. In a line with our intuition, the coefficient β_4 is positive and β_6 is negative. The coefficient β_5 is, however, positive implying that an increase in book-to-market ratio leads to more news articles, potentially due to the articles about bankruptcy and insolvency.

The F-tests reject the model of trading game invariance. The increase in the value of the log-likelihood functions relative to the univariate regressions in table 4 is more significant than for table 6. This indicates that market capitalization, B/M ratio, and past returns may be economically important factors for explaining the rate of information flow.

Count Regressions For Different News Types. Our tests suggest that the model of trading game invariance describes the data on the number of news items reasonably well. We examine next whether it explains the arrival rate of different types of news events, indicated in our sample by various topic codes.

The model of trading game invariance should not be taken too literally: It does not imply that the number of news items of a particular type should be related to the trading activity with the coefficient 2/3, similarly to the number of news items in aggregate. For example, we expect that there are no news events about debt markets, tagged in our sample with 'DBT', for firms with no debt. There are no news about bankruptcies, tagged with 'BKRT', for firms with healthy future prospects. There are almost no major breaking news, tagged with 'NEWS', for small companies but there are many of these news events for large companies. Finally, news reports concerning earnings and dividends, tagged in our sample with 'RES' and 'DIV', will be released regularly over a year for all firms, regardless of their trading activity.

We estimate the negative binomial regression (13) based on the number of news tags for various topic codes. Table 9 reports the estimate for the nine most common topic codes with the rest of the tags aggregated in the last category "Others." Each line in table 9 contains the results for a particular topic code: the estimates with the standard errors, the F-tests with p-values for the three hypotheses $\gamma = \frac{2}{3}$, $\gamma = 0$, and $\gamma = 1$, as well as the log-likelihood functions.

For different news topics, the coefficient γ ranges from 0.60 to 1.23, being the lowest for a category "corporate results" and the highest for a category "major breaking news." The "corporate results" include all corporate financial results, tabular and textual reports, dividends, accounts, and annual reports. Some of these releases are reported at a regularly basis, regardless of the level of trading activity, thus pushing the sensitivity coefficient towards zero relative to the predicted 2/3. In contrast, stories carrying the "major breaking news category" code would be expected to dominate the financial and general headlines of the worlds major newspapers, Web sites, television and radio networks. They will be disproportionately dominated by news articles about large firms, being of interest to a wide audience, thus pushing the sensitivity coefficient upwards relative to the predicted 2/3.

6 Conclusions

We use the news data from Thomson Reuters firm to examine how the information flow varies across stocks and across time for the same stock. We study the arrival rate of the news articles which can be thought as a proxy for the arrival rate of public information.

Our empirical tests show that the arrival of news articles can be modeled as the negative binomial model with the stochastic Poisson arrival rate determined by the level of trading activity. In particular, when the trading activity increases by one percent, the arrival rate of news articles increases by two-third of one percent.

This specification comes naturally from the model of trading game invariance, the main idea of which is the invariance of trading games: Trading games are the same across stocks varying only in the speed with which they are being played. Our paper provides an empirical evidence in favor of 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. This conjecture is necessary to make the model of trading game invariance 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 10-K 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 model of market microstructure invariance has also 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, but not on their "size". We put aside the discussion about news size by assuming that all news articles, or news tags, are equally important. In reality, some articles are more important than others. The importance of article might be related to its length or the measures of how significantly articles differ from previous ones based on some language processing tools such as described in Hanley and Hoberg (2010). Studying variations in different aspects of the information flow is the 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, 2010, “Market Microstructure Invariants,” *University of Maryland, Working paper.*

Kyle, Albert S., Anna A. Obizhaeva, and Tugkan Tuzun, 2010, “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.

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

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.

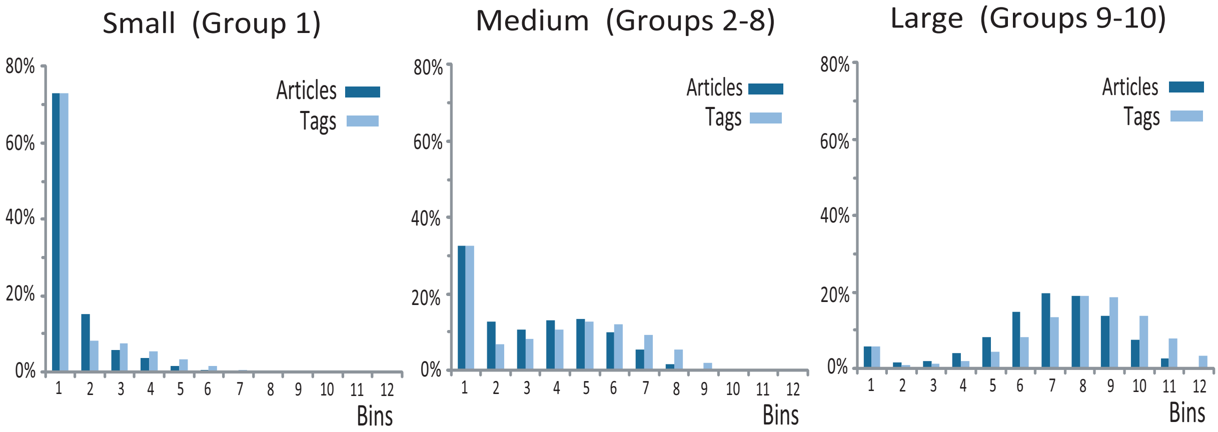

Figure shows the historical distributions of the number of news items N per month. The twelve bins have observations with $0, 1, 2, 3-4, 5-8, 9-16, 17-32, 33-64, 65-128, 129-256, 257-512, 513-1024$ news items per month, respectively; most of them have upper cutoffs of the form 2^i news items per month. 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.

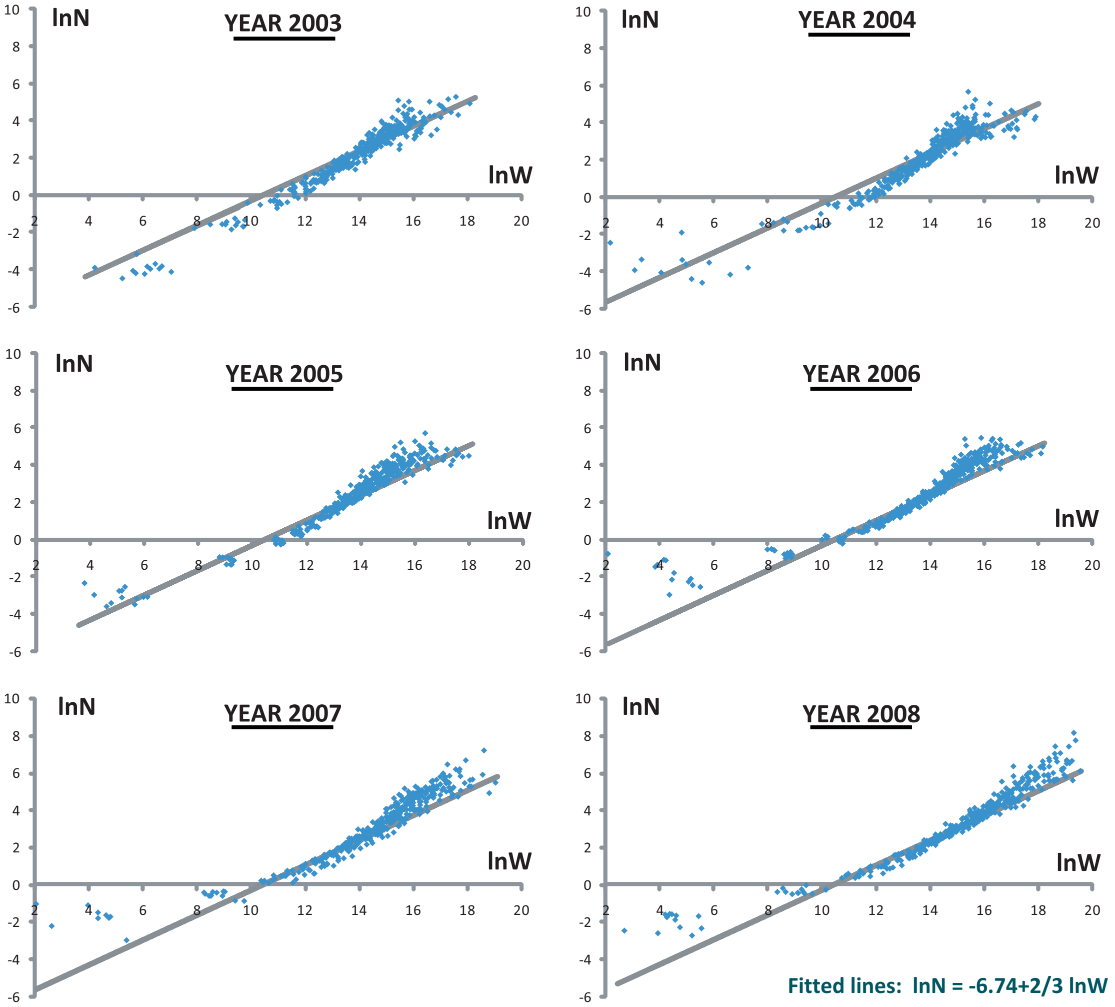

Figure shows the average of the logarithm of the number of news articles released per month for thirty groups based on the trading activity. For each month, stocks are sorted into thirty W-groups such that these groups have the same total number of news articles. The variable N^* is equal to the number 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.

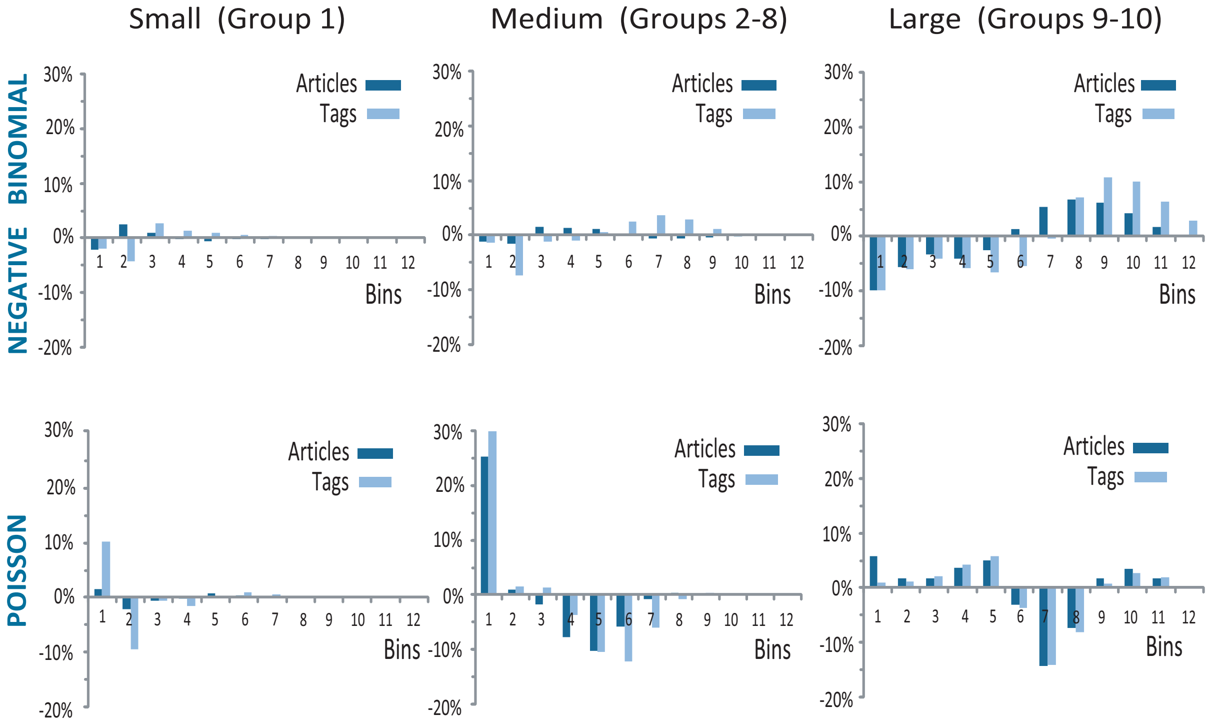

Figure shows the difference between the historical distribution and the fitted distribution of the number of news items N per month. Two specifications are used for the fitted distributions: the Poisson model and the Negative binomial model. The estimates of their parameters are taken from table 4. The 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.

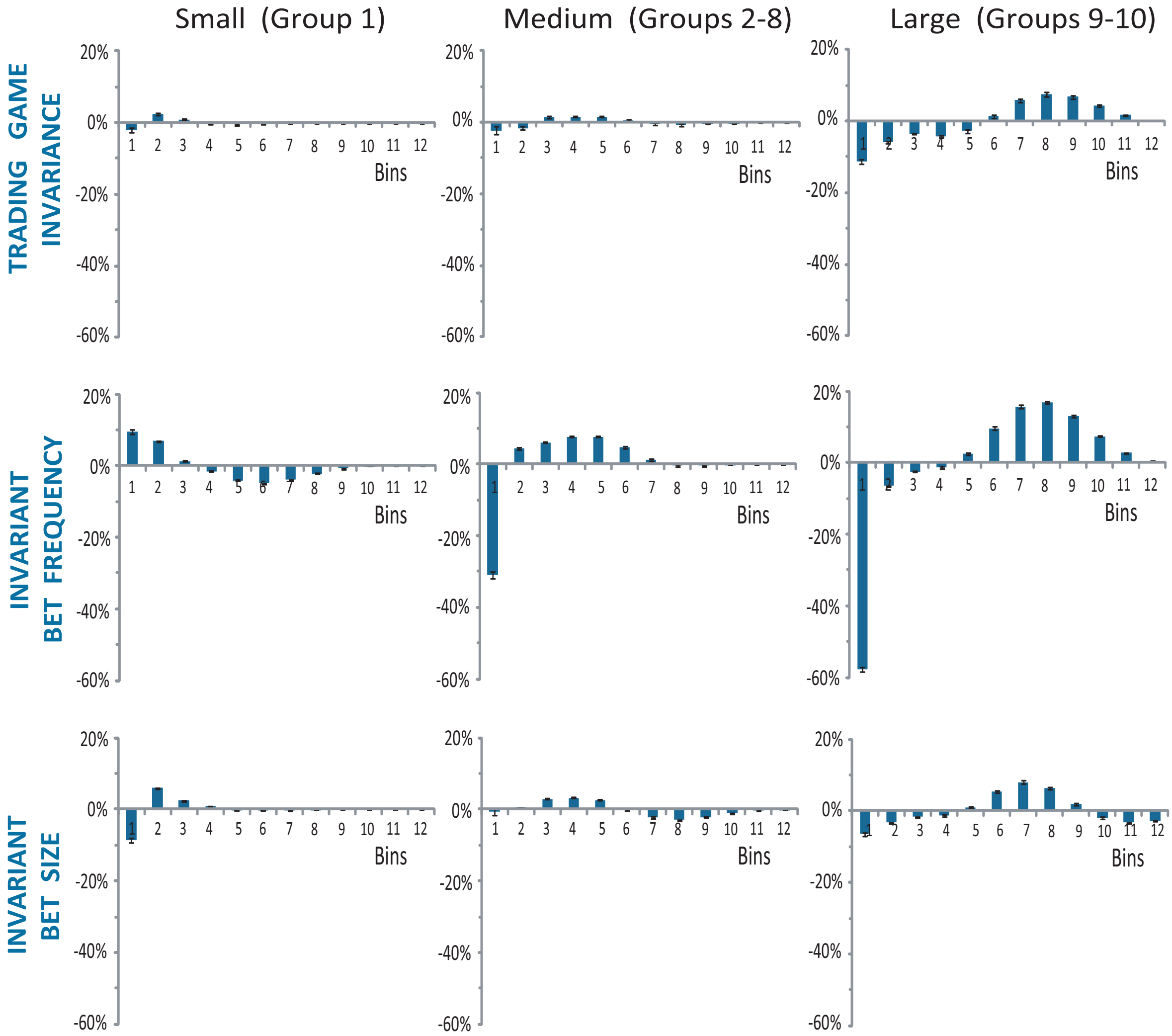

Figure shows the difference between the historical distribution and the fitted distribution of the number of news items articles N per month for the three models. The fitted distribution is based on the estimates for the Negative binomial specification. In calibrating the model, the parameter γ is restricted 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. 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.

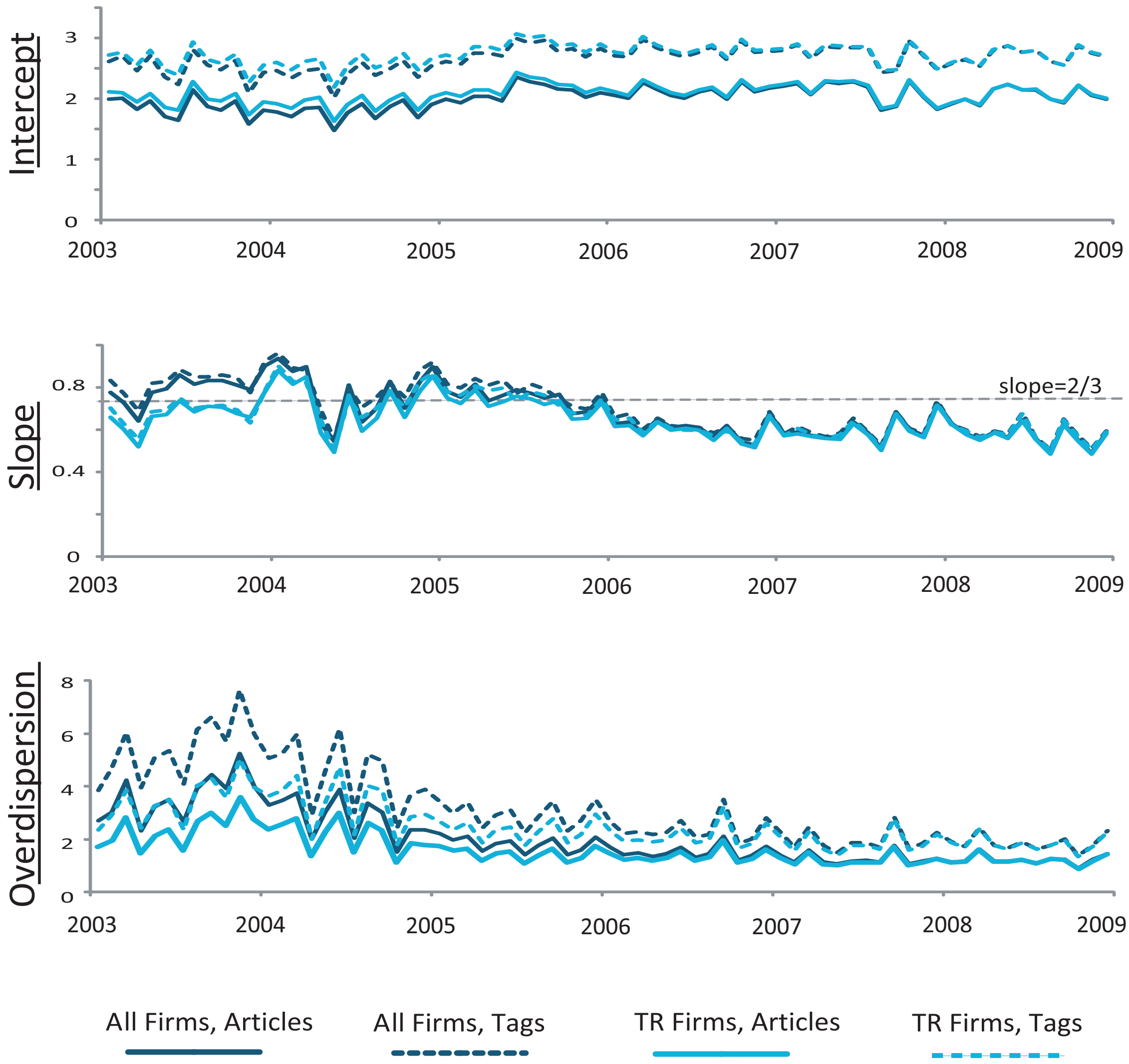

Figure shows the estimates of the intercept η, the slope γ, and the overdispersion parameter α from the negative binomial regression,with the arrival rate of news items $\mu_{t,i}$ for stock i and month t being modeled as,

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

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.