Board of Governors of the Federal Reserve System

International Finance Discussion Papers

ISSN 1073-2500 (Print)
ISSN 2767-4509 (Online)

Number 1372

March 2023

Effects of Information Overload on Financial Markets: How Much Is Too Much?

Alejandro Bernales, Marcela Valenzuela, and Ilknur Zer

Please cite this paper as:
Bernales, Alejandro, Marcela Valenzuela, and Ilknur Zer (2023). "Effects of Information Overload on Financial Markets: How Much Is Too Much?," International Finance Discussion Papers 1372. Washington: Board of Governors of the Federal Reserve System, https://doi.org/10.17016/IFDP.2023.1372.

Effects of Information Overload on Financial Markets: How Much Is Too Much?*

Alejandro Bernales Marcela Valenzuela Ilknur Zer

March 2023

Abstract

Motivated by cognitive theories verifying that investors have limited capacity to process information, we study the effects of information overload on stock market dynamics. We construct an information overload index using textual analysis tools on daily data from The New York Times since 1885. We structure our empirical analysis around a discrete-time learning model, which links information overload with asset prices and trading volume when investors are attention constrained. We find that our index is associated with lower trading volume and predicts higher market returns for up to 18 months, even after controlling for standard predictors and other news-based measures. Information overload also affects the cross-section of stock returns: Investors require higher risk premia to hold small, high beta, high volatile, and unprofitable stocks. Such findings are consistent with theories emphasizing that information overload increases information and estimation risk and deteriorates investors' decision accuracy amid their limited attention.

Keywords: Limited attention, dispersion, sentiment, predicting returns, behavioral biases
JEL classification: G40, G41, G12, G14

*Alejandro Bernales is at Universidad de Chile (Facultad de Economa y Negocios, Departamento de Administracin). Marcela Valenzuela is at the School of Management, Pontificia Universidad Católica de Chile. Ilknur Zer is at the Federal Reserve Board. We thank seminar participants at the Federal Reserve Board. We appreciate Sergen Akarsu and Omur Suer for sharing their Google Search Volume Index data. Valenzuela acknowledges the support of Fondecyt Project No. 1190477 and the support of Instituto Milenio ICM IS130002. Bernales acknowledges the funds provided for the Fondecyt Project No. 1190162. The support of the Economic and Social Research Council (ESRC) in funding the SRC is gratefully acknowledged [grant number ES/R009724/1]. The views in this paper are solely those of the authors and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or of any other person associated with the Federal Reserve System.

1 Introduction

The traditional asset pricing theory assumes that prices instantly incorporate all available information. However, the explosion of information is a defining feature of the current media landscape, and we are often plagued with excess. Thus, in reality, processing the excess information requires sufficient attention from investors, and "a wealth of information creates a poverty of attention," as quoted by Nobel laureate Herber Simon. Given that investors have limited processing capacity (Kahneman, 1973; Johnston and Pashler, 1998), does the information overload affect their investment decisions?

In this paper, we construct a historical news-based index of information load using textual analyses applied to two million articles published in *The New York Times* since 1885. We then use these novel data to study the time-series and cross-sectional predictive ability of information overload over returns. We structure our empirical analysis around a discrete-time learning model, which provides a motivating framework for empirical analysis by linking information overload, expected asset prices, and trading volume.

Why does information overload affect stock market returns? We argue that following periods of information overload, risk-averse investors require a higher risk premium to hold the asset because of increased *information risk* and *estimation risk*.[1] First, the investors' attention capacity correlates negatively with the "excess" information they receive, compromising their processing ability and thus increasing information asymmetry. High information asymmetry, in turn, increases the investors' information risk, as in Easley and O'hara (2004). Second, when information is excessive and dispersed, extracted information is less precise than otherwise, as some useful information will not be analyzed due to limited attention. Consequently, estimated parameters of future asset value are more likely to be wrong, increasing investors' uncertainty and estimation risk.[2]

[1] See for example, Coles and Loewenstein (1988); Coles et al. (1995); Leuz and Verrecchia (2000); Francis et al. (2005); Bawden and Robinson (2009) that document the relationship between the information, estimation risks and the information load.

[2] Alternatively, information overload could lead to positive future asset returns through its effects on speculative demand. As information overload deteriorates investors' attention, investors could eventually reduce their speculative demand in such periods given that they cannot actively trade a stock if they are not paying attention to it. A reduction in demand, in turn, reduces trading volume (Hirshleifer et al., 2009; Hou et al., 2009), puts downward pressure on stock prices, and leads to positive future market returns.

We further hypothesize that information overload does not necessarily affect all stock returns uniformly and thus, has cross-sectional effects. In the category learning model of Peng and Xiong (2006), cognitively constrained investors allocate a higher capacity of attention to a certain group of stocks (such as large stocks or less risky, bond-like stocks). Moreover, small, unprofitable, high beta stocks are more difficult to arbitrage (e.g., Asquith et al., 2005; Hong and Sraer, 2016; Baker and Wurgler, 2006). Hence, we argue that, price reactions driven by information overload would be exacerbated for stocks with weaker arbitrage forces and those that are subject to less attention, making them even more difficult to value and further increasing estimation and information risk. The speculative demand for such stocks should be particularly reduced, translating into higher future returns in periods of information overload.

Our first task is then to quantify *information load*. Based on the extant literature from various disciplines, including organization science, business psychology, accounting, and marketing, we consider the *quantity* and *inconsistency* of information (see Eppler and Mengis, 2004; Roetzel, 2019, for a literature survey). This literature argues that when the quantity of news increases, investors need to follow and process more information, becoming cognitively overloaded. Similarly, information quality, in addition to quantity, should be considered (Bawden and Robinson, 2009). Even if investors do not receive too much information, but if the information they receive is inconsistent or dispersed, then the investor's information processing capacity might be exhausted well before the full quantity of information is used.

To quantify *quantity* and *inconsistency* of information and construct the information load (*InfLoad*) index, we first scan the full content of *The New York Times* daily print edition from January 1, 1885, to December 31, 2018.[3] Second, we apply the logistic regression machine learning tool to distinguish business-related news from other news (such as sports and weather). Third, we adopt the word clouds of Calomiris and Mamaysky (2019) to identify the news related to financial markets only. We employ human readings on randomly selected articles to verify our

[3]One can argue that excess information flow from mass and social media is one of the most salient features of the modern information age. Yet, the phenomenon is not confined to the modern world (Roetzel, 2019; Gleick, 2011; Blair, 2012). For instance, according to Blair (2012), even in the 13th century, information overload was present in the form of "the multitude of books, the shortness of time, and the slipperiness of memory."

procedure. Finally, we define the *InfLoad* index as the average of the quantity and inconsistency components of information flow in a given day. To get the quantity component, we count the number of articles. To measure the inconsistency component, we calculate the standard deviation of the tone of articles published in a day, where the tone is classified via the Loughran and McDonald (2011) dictionary.

After quantifying the information load, our second task is to validate the *InfLoad* index. To this end, we first show that it captures expected trends in the information flow, such as the introduction of the internet or other news outlets and major economic events or stress periods. Second, *InfLoad* is significantly correlated with various proxies of investor attention and asymmetric information, including the Google Search Volume Index, analysts' forecast dispersion, bid-ask spreads, and stock market volatility. Moreover, excess information flow increases the co-movement between stocks. Overall, these exercises bolster our confidence that our index accurately measures the information load investors face in a given period.

We then introduce a discrete-time learning model as a motivating framework for our empirical analysis. The model includes a single asset and two investors who extract information from the news and update their beliefs on the asset value according to Bayes' rule. Importantly, the investors are subject to a limited attention capacity, which we model as a nonlinear inverse function of information load. By doing so, we allow agents' decision quality to improve with new information, but only up to a point. When their capacity to process information is full, the additional information received diminishes the decision quality, in line with the predictions of early cognitive scientists, such as Schroder et al. (1967); Miller (1956); Simon and Newell (1971). Thus, investors' attention span is especially decreased in periods of information overload, which restricts them from interpreting the news perfectly, reducing the precision of their extracted information and implying heterogeneous posterior beliefs.[4] Heterogeneous beliefs, in turn, affect expected returns and trading volume.

To test the predictive ability of information overload over stock market returns, we quantify information *overload* as the proportion of days in which *InfLoad* is above three different historical thresholds in a given month with increasing extremity:

[4]For example, even if attention-constrained investors read the same newspaper at the same time, they can only imperfectly analyze new information by paying attention to a random fraction of the news or possibly misinterpreting it.

historical mean, a one standard deviation band, and a two standard deviation band. Such thresholds help us quantify the economic effects of information overload on returns depending on how "excessive" the flow is.[5]

We first find that, when the information load is "moderate" (when it is above the historical mean), higher news flows increase trading volume, likely due to increased investor attention, as in the model of Barber and Odean (2008). Investors buy stocks when their attention span increases, creating a temporary positive price pressure, but it does not significantly affect market returns.

Second, consistent with the model prediction, "high" information load (when it is one standard deviation above the historical mean) is followed by high excess stock market returns. In this case, information overload rather puts constraints on investors' processing capacity, suggesting increased information and estimation risk as well as decreased investor attention.

Third, when information load is "excessive" (two standard deviations above the mean), the effect of extraneous news events becomes economically and statistically stronger. Excessive information leads to higher future returns. The effect is economically meaningful: A one standard deviation increase in excess information load is associated with a 35 basis point increase in monthly market returns. The predictive power of information overload over returns is persistent and reverses in 18 months. Moreover, in periods of excess information load, investors are distracted more and pay less attention to stock, and thus the trading volume decreases significantly.

Fourth, information overload contains additional explanatory power on market returns beyond the market sentiment (SENT) measure of Tetlock (2007) and Garcia (2013) and other news-based measures, namely the Economic Policy Uncertainty index (EPU) from Baker et al. (2016), the Geopolitical Risk Index (GPR) from Caldara and Iacoviello (2018), and the News-implied Volatility Index (NVIX) from Manela and Moreira (2017). In regressions where we include all of the news-based

[5]One can think of these historical thresholds as the representative agent's processing capacity limits or the "usual" level of information load. The existing literature provides at least two possible overarching explanations for this turning point (see Roetzel, 2019, for a survey). First, the decision-maker may stop acquiring information when her processing capacity is reached. Second, she has limited resources (e.g., time or budget), preventing her from using the available information efficiently. Thus, using historical thresholds, we implicitly assume that investors can learn to digest more or less information as they become accustomed to it. Similarly, they can adjust their resources based on the usual information flow.

measures together, SENT, NVIX, and GPR do not have a statistically significant explanatory power over market returns, and EPU is significant at a 10% level, whereas information overload is still a significant predictor of market returns at a 5% level. These results are not surprising given that both EPU and GPR are shown to be useful predictors of macroeconomic series rather than financial ones. Shapley value analysis reveals that information overload alone contributes about 12 to the variation in market returns, and the other news-based measures (SENT, EPU, GPR, and NVIX) together explain about 35% of the variation, with the contribution of EPU being the highest.

Fifth, *InfLoad* has out-of-sample predictive ability over one-month-ahead market returns. It delivers positive out-of-sample R^2s and improves the forecast accuracy compared with (1) the historical mean, (2) a baseline model that excludes the information overload, and (3) a baseline model obtained through dynamically selecting variables by running the least absolute shrinkage and selection operator (LASSO) regressions.

Finally, we test the effects of information overload in the cross section of stock returns. To this end, we consider four long-short portfolios: big minus small stocks, high beta minus low beta stocks, high volatility minus low volatility stocks, and high operating profit minus low operating profit stocks. We find that investors require higher risk premiums to hold small, high beta, highly volatile, and unprofitable stocks. A one standard deviation increase in information overload translates into 45 bps higher future returns for small stocks than big ones, 89 bps higher future returns for more volatile stocks, 76 bps higher future returns for high beta stocks, and 48 bps higher future returns for unprofitable stocks than profitable ones.

Overall, we find that information overload has both time-series and cross-sectional predictive ability on stock market returns. Our findings challenge the traditional asset pricing theories that the level of information does not affect returns, as prices incorporate all available information instantly. It is worth emphasizing that our results do not necessarily imply a behavioral bias. Price reactions driven by information overload are different from the behavioral bias reaction of Tetlock (2007) and Garcia (2013), which argue that investment decisions and price changes reflect market sentiment. Similarly, we depart from the behavioral model of Hong and Stein (2003), in which investors have different opinions driven by overconfidence.

Instead, we argue that information load affects returns through the constraints in investors' information processing capabilities and it may provide a potential explanation for the gradual information flow mechanism of Hong and Stein (1999).

In this study, we contribute to various strands of the extant literature. First, there is a vast literature across various disciplines, including organization science, accounting, and marketing that studies the effects of information load on agents' decision quality (Eppler and Mengis, 2004; Edmunds and Morris, 2000; Grisé and Gallupe, 1999; Loughran and McDonald, 2014). We apply this field to finance by studying the effects of information overload on stock market returns and trading volume.

Second, we contribute to the literature that focuses on investors' limited attention and its links to stock market dynamics (see, for example, DellaVigna and Pollet, 2009; Hirshleifer et al., 2009; Da et al., 2011). In this paper, we construct a historical index that quantifies the information flow agents face in a given day and show that it is a predictor of stock returns. Various exercises suggest that our index can be used as a valid proxy for investor attention.

Third, our paper is related to the literature on the role and content of the media and its impact on investor behavior. Tetlock (2007) and Garcia (2013) provide evidence that news sentiment can predict movements in stock market activity. More recently, Calomiris and Mamaysky (2019) study the effect of news flows on risk and return predictability in a cross-country setting. In this paper, we provide supporting evidence on the significant role of the media in stock markets similar to the aforementioned papers. We find that not only market sentiment but also the excessive amount of information predict market returns.

Finally, we contribute to the growing literature that relies on news and textual-analysis tools to construct several indexes to proxy financial and economic series, including Baker et al. (2016); Manela and Moreira (2017); Caldara and Iacoviello (2018). In this paper, we construct a news-based historical index that measures the level of excess information agents face. It is correlated by up to 30% with the other news-based measures of uncertainty and risk, yet it contains relevant information that is not captured by the others.

2 Construction and validation of the information load index

2.1 Measuring information load

We construct the information load index in four steps. First, we scan the *full* content of daily *The New York Times* newspapers from January 1, 1885, to December 31, 2018. We obtain titles, keywords, and the lead paragraph of each article published. Given the extensive data and the fact that the main message and tonality must be set out in the first paragraph as noted in the *The New York Times* writing practices, we obtain the lead paragraph as in Chan (2003) as opposed to the full article.[6]

Our second task is to distinguish the financial and economic news ("business news") from other news (i.e., sports, weather, etc.). Post-1981, the name of the corresponding section of the article is provided, which enables us to identify the business news. For the pre-1981 period, we classify each article using the Logistic Regression machine learning tool.[7] Specifically, we train the tool on the articles with section names, so that it can learn from the already tagged news how to classify the rest of the articles. We end up with a total of 2,199,210 business news articles.

Third, within the business news, we focus on news related to financial markets only. To create this sample, we adopt the word clouds of Calomiris and Mamaysky (2019). The authors employ the Louvain method (Blondel et al., 2008), which assigns salient words to mutually exclusive topic areas based on word co-occurrence and divides words into topic groups. They provide word clouds for the five topic groups: markets, governments, commodities, corporate governance and structure, and the extension of credit. We adopt the word clouds of the topic *markets*.

Clearly, an article can contain news on more than one topic. Thus, following Calomiris and Mamaysky (2019), for a given article j published in day d, we

[6] See `https://archive.nytimes.com/www.nytimes.com/learning/general/weblines/411.html`

[7] We apply Neural Networks, Gradient Method, and Logistic Regression algorithms. After training, testing, and validating algorithms, we conclude that the Logistic Regression has the best performance, with an accuracy of 92.5%, and thus we use the optimized parameters from the Logistic Regression model to classify the news.

assign a weight corresponding to the topic markets (*Mkt*), as follows:

$$w_{Mkt,j,d} = \frac{C_{Mkt,j,d}}{C_{j,d}}, \tag{1}$$

where $C_{Mkt,j,d}$ and $C_{j,d}$ are the number of words associated with topic markets and the total number of words appearing in the lead paragraph of article j in day d, respectively. Thus, for each article, we calculate the relative frequency of words that correspond to markets.

Finally, we construct our information load index by considering two components: the *quantity* and *inconsistency* of information following the extant literature from various disciplines (see Eppler and Mengis, 2004; Roetzel, 2019, for a literature survey.). When the *quantity* of information becomes excessive, the decision accuracy declines as investors need to follow and process more information. Not only the quantity, but also the content of the news matters. *Inconsistent* news can plague an investor's limited information processing capacity, as it could lead to more confusion in interpreting the available information. After reading such news, investors can fail to recall and respond appropriately to the information provided (Hirst and Hopkins, 1998). Thus, we define the information load index ($InfLoad_d$) as the daily average of these (scaled) components:[8]

$$InfLoad_d = avg(Q_d, INC_d), \tag{2}$$

where Q_d and INC_d measure the quantity and the inconsistency of the news flow in day d, respectively.

We measure the quantity by counting the number of articles related to financial markets published on day d. In other words, we sum the weights associated with market news, $w_{Mkt,j,d}$, introduced in (1), across all articles:

$$Q_d = \sum_j w_{Mkt,j,d}. \tag{3}$$

To quantify inconsistency, we estimate the tone of each article on a given day by calculating the ratio of positive words in excess of the negative ones, divided by

[8]In Section 4.5, we present the results when we employ a principal component analysis or when we consider the quantity and inconsistency components separately.

the total number of words. Then, we calculate the standard deviation of article tones published on the same day. That is, for article j and day d:

$$INC_d = \sigma_{Tone_{j,d}}, \tag{4}$$

with the tone of each article obtained as:

$$Tone_{j,d} = \frac{n_{pos,j,d} - n_{neg,j,d}}{n_{j,d}} \times w_{Mkt,j,d}, \tag{5}$$

where, $n_{pos,j,d}$, $n_{neg,j,d}$, and $n_{j,d}$ are the number of positive, negative, and total words in the lead paragraphs of financial market news articles, respectively. We determine the positive and negative words using the financial dictionary developed by Loughran and McDonald (2011), while carefully addressing the double negation problem.

2.2 Validation of the information load index

In this section, we present two exercises in which we aim to validate our proposed *InfLoad* index introduced in (2). First, we show that it varies over time to capture changing trends in the media and to follow major economic events. Second, it is significantly correlated with various proxies of investor attention, asymmetric information, indicators of financial uncertainty, and stress, including news-based ones. Overall, these exercises bolster our confidence in our measure *InfLoad* as an accurate measure of information load.

2.2.1 Plausibility of quantifying historical news flows

To be a relevant measure of information load, *InfLoad* should change over time to capture expected trends in the information flow. Figure 1 plots the information load measure, along with the quantity and inconsistency components averaged across days in a given year since 1885. *InfLoad*, Q, and INC are all positively correlated with each other and follow an inverted-V trend overall: increasing early in the sample and decreasing post-1990s.

Such trends are expected in the information flow throughout the sample period. In the end, the print edition of *The New York Times* has changed significantly

over time. For instance, a published newspaper in the 1900s had about the same number of articles as today, whereas a newspaper in the 1950s had about five times as many articles as today. In addition, post-1990s the introduction of the internet and other news outlets such as Bloomberg or social media changed the use of newspapers. Yet big spikes are observed even in recent periods, such as during the 2008 Global Financial Crisis.

Moreover, we see that the information load increases during stress periods, on average, which is not surprising given the increased news flow during periods of uncertainty. The largest spikes coincide with the well-known episodes of stress in the financial markets, including the Great Depression, the two world wars, and the Global Financial Crisis.

Not only the quantity of news but also the consistency of the news follow a similar pattern. One interesting period is the great moderation–the mid-1980s until the Global Financial Crisis. Macroeconomic volatility in the United States (similar to other developed economies) was significantly low during this period, possibly driving a consistent information flow, followed by a spike post-2007.

2.2.2 Correlations with related measures

In Table 1, we present contemporaneous Pearson correlation coefficients of the quantity of information (Q), the tone dispersion (INC), and the information load index (*InfLoad*), with proxies for (1) investor attention, (2) asymmetric information, (3) investor uncertainty, (4) financial uncertainty, and, finally, (5) other news-based measures proposed in the literature.

First, we expect information load to be negatively correlated with investor attention. We do not have direct measures of investor attention. However, following Da et al. (2011); Akarsu and Süer (2021), we first use the abnormal search frequency in the Google Search Volume Index–ASVI as a proxy for investor attention.[9] Second, we use the marketwide excess stock correlation as another proxy

[9]ASVI is calculated as the excess search volume in a month relative to the previous quarter:

$$\text{ASVI}_t = \log(\text{SVI}_t) - \log[Median(\text{SVI}_{t-1}, \text{SVI}_{t-2}, \text{SVI}_{t-3})], \tag{6}$$

where SVI_t is the average of 579 U.S. firms' weekly Google Search Volume Indexes in a given month t.

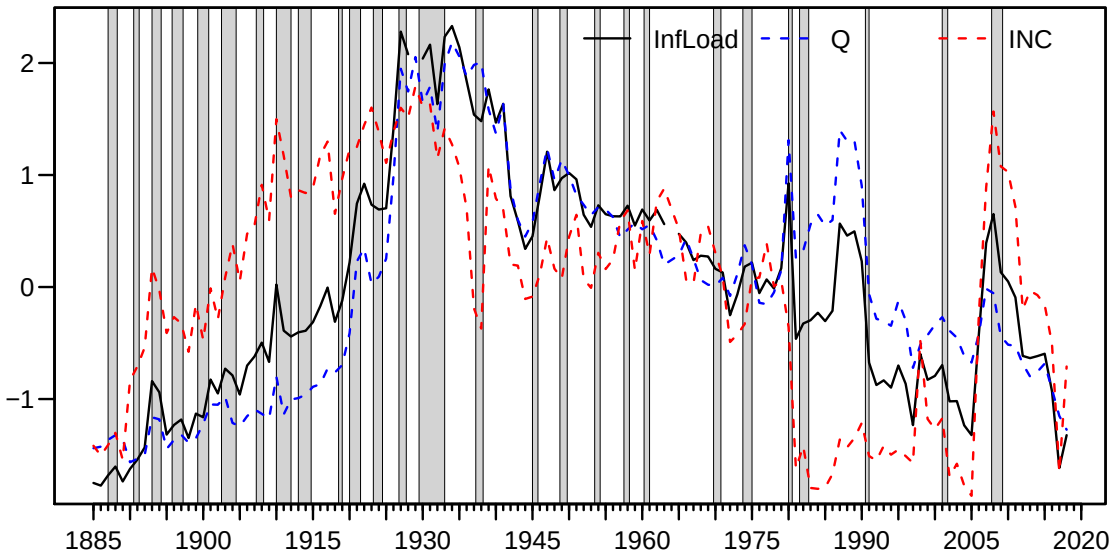


Figure 1: Information Load

The figure presents the annual averages of the daily information load measure along with its quantity (Q) and inconsistency (INC) components introduced in (2), (3), and (4), respectively. Quantity is the sum of the weights related to financial markets across all articles published on a given day. Inconsistency is the standard deviation of the tone of the articles published in a day, where the tone is classified using the Loughran and McDonald (2011) dictionary. *InfLoad* is the average of the (scaled) quantity and inconsistency components. NBER recession periods are marked in gray. The sample period is 1885:Q1–2018:Q4. Data are obtained from the printed edition of *The New York Times*, ProQuest, TDM Studio. New York Times Historical Newspapers.

for investor inattention (CORR).[10] We use this measure because limited investor attention leads investors to process more marketwide and sectorwide information than firm-specific information, increasing correlations among the stocks, in line with the category-learning behavior of Peng and Xiong (2006). Table 1 rows 1 and 2 show that periods of high information load coincide with decreased Google search volumes and increased market-wide correlation, both suggesting that the excess news flow may increase confusion among investors, leading to more passive portfolio trading.

Second, high information load should be associated with higher information risk and asymmetry, as it compromises investors' ability to process information, especially for less-informed investors (see, for example, Leuz and Verrecchia, 2000; Bawden and Robinson, 2009; Muslu et al., 2015). To proxy the information risk,

[10]To obtain CORR, we use the Fama-French 25 equally-weighted portfolios formed on size and book-to-market. For each month, we calculate the sample correlation average between all portfolio pairs and calculate its difference from the past 24 months average.

we rely on two limit order book metrics at stock level obtained from the monthly Center for Research in Security Prices, CRSP 1925 US Indices Database, Wharton Research Data Services and aggregated across firms (equally-weighted).[11] We show that a higher information load is associated with a higher bid-ask spread (SPR). We then consider the dispersion of bid and ask prices by calculating the distance between the highest ask and lowest bid prices (DIST). High information load is significantly and negatively associated with the dispersion of the bid-ask prices, suggesting that excess information exacerbates information asymmetries and investor confusion.

Third, when the information is excessive and dispersed, estimated parameters of future returns or cash flows are more likely to be wrong, increasing estimation risk and in turn, investor uncertainty (see, for example, Coles and Loewenstein, 1988; Coles et al., 1995). To proxy the estimation risk, we first use stock market volatility, calculated as the standard deviation of daily market returns for a given month. Row 4 shows that high information load is associated with higher dispersion in trading prices. Second, we construct two measures to quantify the analysts' estimated dispersion using the Institutional Brokers' Estimate System (Refinitiv, IBES North American Summary & Detail Estimates, Level 2, Current & History Data, Adjusted and Unadjusted) summary database. Accordingly, we first calculate the standard deviation of the analysis earnings-per-share (EPS) forecasts for a given stock and time period (DISP1). Second, we calculate the average value of the absolute deviation of the highest EPS forecast from the actual value and the lowest EPS forecast from the actual value (DISP2). We then calculate the cross-sectional averages for both measures. Rows 5 and 6 show that both analysts' dispersion measures increase with higher information load, consistent with higher estimation risk in the periods of excess information.

Fourth, information load is expected to increase in times of increased financial or economic uncertainty and stress. We find that information load increases with the CBOE Volatility Index (VIX) and Bekaert et. al (2019)'s uncertainty (BEX) measure. When we look at the information load components, we see that increased uncertainty is related more to the dispersion in tone (INC) than to the *quantity* of news. The correlations range from 20% to 55%, suggesting that excessive infor-

[11] We calculate the equally-weighted averages instead of value-weighted ones because we expect cross-sectional differences for the big and small stocks as discussed in Section 4.4.

mation load and economic uncertainty are distinct phenomena and that the effects of news dispersion are more likely to operate through investor disagreement than uncertainty. Moreover, in periods of increased financial stress, we observe a boost in both the quantity and the inconsistency of the news (rows 9 and 10).

Finally, we compare the *InfLoad* index with other news-based uncertainty measures: EPU, GPR, and NVIX. The Pearson correlation coefficients range from 0.05 to 0.30, implying that the *InfLoad* measure is related to the aforementioned news-based measures, yet it also captures important information that is not reflected in those measures.

3 Information overload and financial market dynamics

3.1 Motivating framework

As a road map to our empirical analysis, we consider a discrete-time learning model. Our model incorporates attention capacity as a function of the information load. In periods of excessive information load, it is more difficult for investors to process relevant information, making them reach their maximum attention capacity quicker, and reducing their decision accuracy rapidly—an argument with a theoretical basis that is from psychologists and cognitive scientists such as Schroder et al. (1967), Miller (1956), and Simon and Newell (1971).

In our model, there are two types of investors, A and B, who are rational and subject to the same attention capacity, denoted by κ. There are two periods ($t = 0, 1$) where risk-averse investors trade a single risky asset or a portfolio of assets at $t = 0$ with a payoff of ν at $t = 1$. There are M outstanding shares that can be traded, and each investor is born with $M/2$ shares. For simplicity, we assume the risk-free rate is zero.

3.1.1 The learning process

At $t = 0$, both investor types A and B start with the same prior beliefs on the true value of asset ν, with $\nu \sim \mathcal{N}(\bar{\nu}, 1/\tau_0)$, where τ_0 is the precision of their prior

beliefs on the asset value.

The investors receive news (or a load of information, denoted as *InfLoad*) at the same time through the same sources that they have access to (e.g., newspapers, analyst reports, and other media). Denote Inf^A and Inf^B as the information extracted from the news at $t = 0$ by the investor types A and B, respectively:

$$Inf^A = \nu + \varepsilon^A, \qquad \varepsilon^A \sim \mathcal{N}(0, 1/\tau_\varepsilon), \tag{7}$$

$$Inf^B = \nu + \varepsilon^B, \qquad \varepsilon^B \sim \mathcal{N}(0, 1/\tau_\varepsilon). \tag{8}$$

Although the investors are subject to the same attention capacity, they differ in their interpretation of the news ($Inf^A \neq Inf^B$) because their attention capacity is *finite*. For example, even if attention-constrained investors read the same newspaper at the same time, they can only imperfectly analyze new information by paying attention to a random fraction of the news or possibly misinterpreting it. Hence, even under perfect information, noise in the learning process is introduced through limited capacity and information load. The precision of the extracted information (τ_ε) depends on the investors' attention capacity and is assumed to be the same for both investors for simplicity.

Using the standard Bayesian updating process, the following lemma characterizes the investors' posterior beliefs.

Lemma 1. *The posterior beliefs of the investor types A and B at $t = 0$ are normally distributed, and denoted by $\nu^A|Inf^A \sim \mathcal{N}(\hat{\nu}^A, 1/\tau)$ and $\nu^B|Inf^B \sim \mathcal{N}(\hat{\nu}^B, 1/\tau)$, respectively. The precision of posterior beliefs (τ) is given by:*

$$\tau = \tau_0 + \tau_\varepsilon, \tag{9}$$

while the expected values of their posterior beliefs are:

$$\hat{\nu}^A = \frac{\tau_0}{\tau}\bar{\nu} + \frac{\tau_\varepsilon}{\tau} Inf^A \tag{10}$$

$$\hat{\nu}^B = \frac{\tau_0}{\tau}\bar{\nu} + \frac{\tau_\varepsilon}{\tau} Inf^B. \tag{11}$$

Proof. See the appendix. □

3.1.2 Quantifying the precision of the posterior beliefs

To quantify the precision of the posterior beliefs τ, we use the concept of entropy (H) from the information theory. Let I^A and I^B be the reduction in entropy in the posterior beliefs in relation to prior beliefs of the asset value for investors A and B, respectively:[12]

$$I^A = I^B = H(\textit{prior beliefs}) - H(\textit{posterior beliefs}) = \frac{1}{2}\ln\frac{1/\tau_0}{1/\tau}. \tag{13}$$

Following Sims (2003) and Peng and Xiong (2006), we assume I^A and I^B are linear and positive functions of the investors' attention capacity (κ):

$$I^A = I^B = \frac{1}{2}\kappa. \tag{14}$$

Using (9), (13), and (14), we have

$$\tau_\varepsilon = \tau_0(e^\kappa - 1), \tag{15}$$

$$\tau = \tau_0 e^\kappa. \tag{16}$$

We assume κ has the following form:

$$\kappa(\textit{InfLoad}) = \phi\left(1 - \frac{(\textit{InfLoad} - q)^2}{\textit{InfLoad}^2 + q^2}\right), \tag{17}$$

where ϕ is the maximum attention capacity of investors and q is a given threshold for the level of information (*InfLoad*) that can be processed by investors. Thus, in line with the predictions of early cognitive theories, (17) allows attention capacity to alter with *Infload* in a nonlinear way. In the model of Schroder et al. (1967), for example, the task performance of a decision-maker initially improves as more information is received. But, when the amount of information reaches the threshold, the additional information diminishes the quality of the decision-making.

[12] In a nutshell, the entropy of a random variable x measures its uncertainty. Assuming $x \sim \mathcal{N}(\mu, \sigma)$, it is defined as:

$$H(x) = \frac{1}{2}\ln(\sigma) + \frac{1}{2}\ln(2\pi e), \tag{12}$$

Using (15), (16), and (17), the precision of the extracted information (τ_ε) and the precision of the investors' posterior beliefs (τ) are:

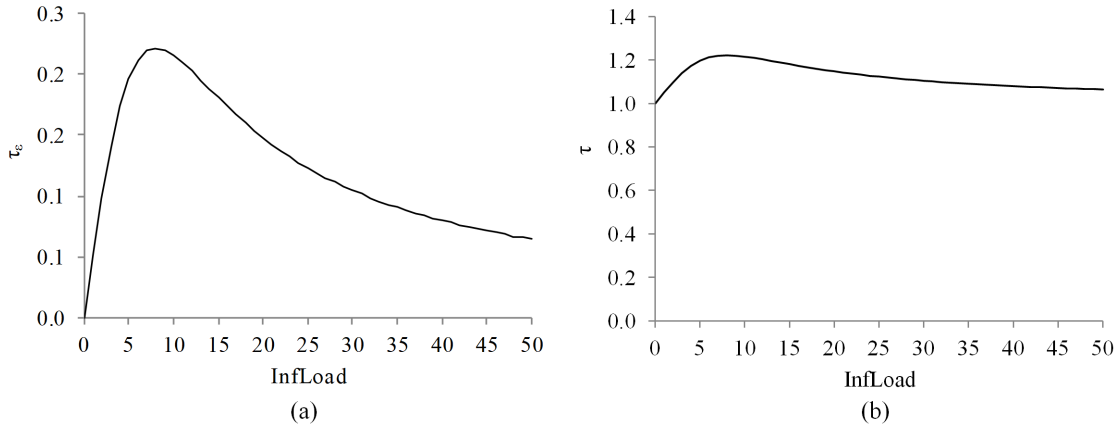
$$\tau_\varepsilon = \tau_0 \left(e^{\phi\left(1-\frac{(InfLoad-q)^2}{InfLoad^2+q^2}\right)} - 1 \right), \tag{18}$$

$$\tau = \tau_0 e^{\phi\left(1-\frac{(InfLoad-q)^2}{InfLoad^2+q^2}\right)}. \tag{19}$$

Figure 2 illustrates the relation between information load and τ_ε and τ. First, at a given threshold q, when *InfLoad*=0 (i.e., there is no news), investors cannot extract any information and do not need to use any capacity ($\kappa = 0$). Thus, the precision of the posterior beliefs is equal to the precision of prior beliefs ($\tau = \tau_0$) and $\hat{\nu}^A = \hat{\nu}^B = \bar{\nu}$ in Lemma 1.

Figure 2: Effect of information load on the precisions of the extracted information and posterior beliefs

Panels (a) and (b) show the effect of information load (Infload) on the precisions of the extracted information (18) and posterior beliefs on the true value of the asset (19). We assume $q = 8$, $\phi = 0.2$ and $\tau_0 = 1$.



Second, on the other extreme, if *InfLoad*$\to \infty$ (i.e., there is an excessive information load), the extracted information is very imprecise (i.e., $\tau_\varepsilon \to 0$), making $\kappa = 0$. Thus, the precision of the posterior beliefs is equal to the precision of prior beliefs ($\tau = \tau_0$) and $\hat{\nu}^A = \hat{\nu}^B = \bar{\nu}$.

Finally, for the rest of the cases, the precision of the beliefs is higher than zero. When *InfLoad* $= q$, investors allocate their maximum attention, i.e., $\kappa = \phi$ and hence, the information extracted has the highest precision. This precision however, is not perfect because ϕ is finite. Consequently, $\tau_\varepsilon = \tau_0(e^\phi - 1)$ and $\tau = \tau_0 e^\phi$. Then,

$\hat{\nu}^i = (1/e^\phi)\bar{\nu} + ((e^\phi - 1)/e^\phi)Inf^i$ for $i = A, B$.

3.1.3 Asset returns and trading volume

At $t = 0$, investors maximize their constant absolute risk aversion (CARA) utility functions based on their beliefs about the asset value. Because investors have a CARA utility function and all stochastic variables are normally distributed, the investors' optimization reduces to the usual mean-variance problem (see Ingersoll, 1987; Eeckhoudt et al., 2011):

$$\max_{y_i} E_t[\eta_i] - \frac{\theta}{2} Var[\eta_i], \qquad i = A, B, \tag{20}$$

with

$$\eta_A = \left(\nu^A | Inf^A - p\right) y_A \quad \text{and} \quad \eta_B = \left(\nu^B | Inf^B - p\right) y_B \quad \text{for } t = 0, \tag{21}$$

subject to the market clearing condition:

$$y_A + y_B = M, \tag{22}$$

where θ is the level of risk aversion, p is the traded asset price at $t = 0$, and y_A and y_B are the number of shares that investors A and B hold, respectively. $\nu^A | Inf^A$ and $\nu^B | Inf^B$ are the investors' posterior beliefs about the value of the asset at $t = 1$ in Lemma 1. Proposition 1 and Lemma 2 follows from solving the investors' optimization problem outlined in (20)—(22):

Proposition 1. *Investors' asset holdings and the equilibrium asset price at $t = 0$ are given by:*

$$y_A = \frac{M}{2} + \frac{\tau}{2\theta}\left(\hat{\nu}^A - \hat{\nu}^B\right), \tag{23}$$

$$y_B = \frac{M}{2} + \frac{\tau}{2\theta}\left(\hat{\nu}^B - \hat{\nu}^A\right) \tag{24}$$

$$p = \frac{\hat{\nu}^A + \hat{\nu}^B}{2} - \frac{M\theta}{2\tau}. \tag{25}$$

Proof: See the appendix.

Lemma 2. *The expected trading volume and asset price change are:*

$$E[volume] = E\left[\frac{1}{2}\left|\frac{M}{2} - y_A\right| + \frac{1}{2}\left|\frac{M}{2} - y_B\right|\right] \tag{26}$$

$$= \frac{1}{\theta}\sqrt{\frac{\tau_\varepsilon}{\pi}} \tag{27}$$

and

$$E\left[v - p\right] = E\left[v\right] - E\left[\frac{\hat{\nu}^A + \hat{\nu}^B}{2}\right] + E\left[\frac{M\theta}{2\tau}\right] \tag{28}$$

$$= \frac{M\theta}{2\tau}.$$

Proof: See the appendix.

Thus, asset price changes are determined by the average expected value of posterior beliefs of the two investor types and the risk premium of $M\theta/2\tau$. The investors' asset holdings (y_A and y_B) and the trading volume depend on the investors' heterogeneity on the expected value of their posterior beliefs ($\hat{\nu}^A - \hat{\nu}^B$). The asset prices and trading volume are functions of information load as $\hat{\nu}$ depends on the precision of the posterior beliefs (τ) and the precision of the extracted information (τ_ε), where both vary by information load (see Lemma 1 and (18) and (19)).

Figure 3 panel (a) visualizes the association of information load with the expected trading volume. There are three possible scenarios. First, when *InfLoad* $= 0$, the investors are homogenous in their posterior beliefs because they only use their prior beliefs in the learning process (i.e., $\hat{\nu}^A = \hat{\nu}^B = \bar{\nu}$). In that scenario, there is no trade and they hold $M/2$ shares.

Second, whenever *InfLoad* > 0, there will be a trade. Investors extract different level of information from the news ($Inf^A \neq Inf^B$) because they have finite capacities and cannot process news perfectly. Particularly, when *InfLoad* $= q$, then the precision τ_ε is highest (see Figure 2). (10) and (11) imply that the investors give the highest weight to new information rather than the expected value of their prior ($\bar{\nu}$). Thus, heterogeneity ($\hat{\nu}^A - \hat{\nu}^B$) reaches its maximum, inducing the highest number of trades.

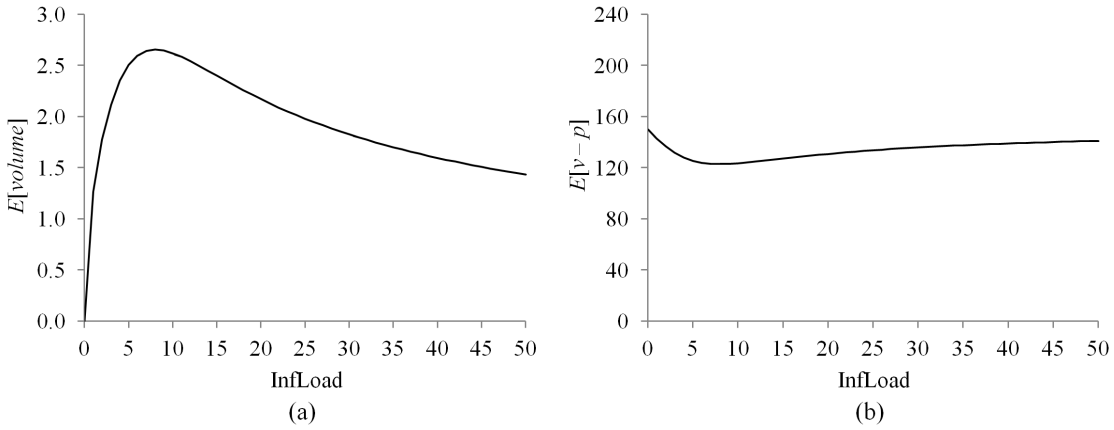
Third, when *InfLoad* $\to \infty$, the precision of the extracted information is reduced to zero ($\tau_\varepsilon \to 0$). In this case, the investors are also homogenous in their posterior

beliefs (i.e., $\hat{\nu}^A = \hat{\nu}^B = \bar{\nu}$) and again there is no trade and they hold $M/2$ asset shares.

Figure 3 panel (b) illustrates the relationship between information load and the expected price change. Asset returns decrease until they reach their lowest value when $InfLoad = q$ and increase thereafter. When the information load is equal to the threshold, τ reaches its maximum value, and investors ask for the lowest compensation to hold the asset (i.e., the investors' risk premium $M\theta/2\tau$ is at its lowest).

Figure 3: Information load, expected trading volume and expected price change

Panels (a) and (b) show the effect of information load (Infload) on expected trading volume and expected price change, respectively. Specifically, we plot the results presented in Lemma 2 using equations (26) and (28). We assume the following value for the parameters: $q = 8$, $\phi = 0.2$, $\tau_0 = 1$, $M = 3000$, and $\theta = 0.1$.



3.2 Econometric model

To examine the predictive power of information load over future monthly market returns, we rely on the following regression model:

$$rx^m_{t+h} = \alpha^h_1 + \alpha^h_2 X_t + \alpha^h_3 \text{SENT}_t + Controls + \varepsilon^h_{t+h}, \quad (29)$$

$$X_t = InfLoad_t \; V \; InfOver_t(\tau) \quad (30)$$

where rx^m_{t+h} is the cumulative market returns (scaled by the horizon h) in excess of the risk-free rate from $t+1$ to $t+h$. Market returns are based on the CRSP NYSE/AMEX/NASDAQ value-weighted portfolio in excess of the one-

month Treasury bill rates and are obtained from Kenneth French's online data library.

We first examine the effects of information load on future market returns by including $InfLoad_t$ (defined in (2)) as the main independent variable. Then, we incorporate the possible nonlinear effects of information load on market dynamics. In line with the predictions of our model, we expect such effects to be highly dependent on how "excessive" the flow is. When the information flow is "moderate," increased news flows rather increase investor attention and the precision of the information extracted from the news, leading to higher trading activity. When the information load is "high," however, it is more likely to put constraints on investors' processing capacity. Such an effect is likely to be more pronounced when the information load is "excessive," reducing the precision of the extracted information. To capture these dynamics, we define $InfOver_t(\tau)$ as the proportion of days in month t, with $InfLoad_t$ higher than the threshold τ. We consider three different thresholds with increasing extremity: historical mean, a one standard deviation band, and a two standard deviations band, all calculated using the two-week moving window sizes. The main findings are robust to choosing a different moving window size as reported in Section 4.5.

We control for market sentiment (SENT_t) as it has been shown to be a predictor of stock market activity (Tetlock, 2007; Garcia, 2013, among others). We define sentiment following Garcia (2013): For each day, we count the total number of positive and negative words as well as the total number of words in the corresponding lead paragraphs to obtain the proportion of positive and negative words. We then find the difference between those proportions.

We also control for a set of variables ($Controls$) that are shown to be significant predictors of market returns in traditional asset pricing studies. First, we consider the S&P 500 monthly dividend yield (DY_t) following Shiller (1978); Campbell (1987); Fama and French (1988), among others. We obtain data from Global Financial Data, Inc., GFDatabase (GFD). Second, we include the changes in the consumption–wealth ratio (CAY_t) of Lettau and Ludvigson (2001) to control for the effects of business cycles on the aggregate variation in stock market returns. The quarterly data are from Lettau's website. Monthly estimates are then constructed by repeating the most recently available observation. Third, we include volatility (RVOLA_t), calculated as the standard deviation of stock market returns.

We also consider the default spread (DS_t), the term spread (TS_t), and the change in short-term interest rates ($\Delta STIR_t$) following Keim and Stambaugh (1986); Campbell (1987); Fama and French (1989). DS_t is measured as the difference between the Moody's Seasoned BAA and AAA corporate bond yields and data are retrieved from FRED, Federal Reserve Bank of St. Louis. TS_t is calculated as the difference between the 10-year Treasury bond and the 3-month T-bill yields, and $\Delta STIR_t$ is the changes in the 3-month T-bill rate. We obtain interest rate data from the GFD. Finally, we include the Amihud's (2002) illiquidity measure, $ILLIQ_t$ using price and volume data from CRSP.

Furthermore, we use the contemporaneous market trading volume, obtained from GFD, as the dependent variable to see the effects of "excess" information on trading volume.

In Table 2, we present summary statistics for all of the variables. Information overload measures—where excess-news-flow days are identified using the historical mean, one standard deviation band, and two standard deviations band, respectively—are correlated positively with each other but not highly correlated with the rest of the control variables. Thus, information overload is not likely to share common information with standard predictors of stock market returns.

On average, the information load is above the historical mean of about 50% of the days in a month, reaching a maximum of almost 75% of the days in a month. The first and second-order autocorrelations for information overload measures are negligible, as opposed to DY_t, TS_t, DS_t, and CAY_t, which present first- and second-order autocorrelations of more than 0.9. The Phillip-Perron stationarity test strongly rejects the null of the unit root for all of the variables except the dividend yield (DY_t).

4 Empirical findings

In this section, we present our main findings. We start by studying the in-sample predictive power of information overload on stock market returns in a time-series setting. Then, in Section 4.2, we examine the explanatory power of the information overload index on future market returns in comparison to other text-based measures. In Section 4.3, we examine the out-of-sample predictive ability of infor-

mation overload. In Section 4.4, we present cross-sectional analysis, and finally, Section 4.5 presents the robustness analyses.

4.1 Effects of information overload on stock markets: Time-series analysis

Table 3 presents the estimated coefficients from (29) for the period spanning March 1952 to December 2018 for $h = 1$.[13] First, we include *InfLoad* as the main independent variable. Column I shows that information load is not statistically related to stock market returns. This result is expected, given our model's prediction that the information load has a nonlinear relationship with stock market returns (see Figure 3). Thus, we then incorporate such nonlinearity.

When the information load is right above the historical mean—*InfOver*(mean)—increased news flows lead to higher trading volume and no significant effect on next month's market returns (Columns I and II). Increased trading volume is consistent with the literature arguing that a higher flow of news encourages trading due to increased attention (Barber and Odean, 2008), attracting more activity from high-frequency traders (Foucault et al., 2016), or increased disagreement (Hong and Stein, 2007).

"High" information load (one standard deviation above the historical mean)—*InfOver*(1sd)—is associated with higher next-period returns (Column III). When information load is "excessive" (two standard deviations above the mean)—*InfOver*(2sd)—the predicting power of information overload over the next month's returns becomes economically and statistically stronger (Column V). A one standard deviation increase in information overload increases the market risk premium by 35 basis points. In addition, excessive information leads to lower trading volume, consistent with reduced speculative demand due to the limited attention (see, for example, Hou et al., 2009) (Column VI).

We then study the long-run predictive ability of information overload on cumula-

[13]The sample is restricted to post-1952 because of the data availability of the control variables. Because that period coincides with the stock markets becoming a central investment vehicle for the general public, we let the baseline specifications cover the post-1950s. That said, in Section 4.5 we restrict the control variables so that we can study the effects of information overload in stock market returns in a longer historical sample and observe the sensitivity of our findings in different time periods.

tive excess returns over the subsequent 24 months. We use Hodrick (1992) standard errors to address the serial correlation in the residuals induced by overlapping observations. Table 4 shows that the predictive power of *InfOver*(2sd) over returns is persistent. The economic effect is almost monotonically decreasing and vanishes after 18 months. The proportion of the variance of cumulative market returns (R^2s) is high. However, as R^2s are roughly proportional to the horizon under the null hypothesis, they should be interpreted with caution (Boudoukh et al., 2008).

4.2 Predictive ability of *InfOver* and other text-based measures over returns

We test the explanatory power of "excessive" information load—*InfOver*(2sd)—on future market returns in comparison to other text-based measures: SENT, EPU, GPR, and NVIX. To this end, we run the baseline regressions by (1) considering *InfOver* alone' (2) using control variables (*Controls*) only, (3) including *InfOver* along with the control variables, (4) including four news-based measures with controls, and (5) considering *InfOver*, SENT, EPU, GPR, and NVIX together in addition to the control variables.

Table 5 reports the estimated coefficients. *InfOver*(2sd) is a significant predictor of market returns (Columns I, III, and IX). Within the news-based measures, only the EPU index has statistically significant explanatory power (Columns IV–VII). Furthermore, Column IX shows that the SENT, NVIX, and GPR indexes do not have statistically significant explanatory power on market returns, with EPU being only 10% significant. *InfOver* is still a significant predictor of market returns, suggesting that our information load index contains additional explanatory power on market returns beyond the other news-based measures proposed in the literature.

We then calculate the contribution of each regressor to the overall R^2 (share of explained variance) using Shapley values to judge the relative importance of the variables in driving the changes in stock market returns. As expected, the standard predictors of returns (dividend yield, the consumptionwealth ratio, realized volatility, default spread, term spread, changes in interest rates, and market liquidity) together explain about half of the variation in stock market returns. Yet, information overload alone explains about 12% of the variation, whereas the other news-based measures (SENT, EPU, GPR, and NVIX) together explain 35% of the

variation, with the contribution of EPU being the highest.

4.3 Out-of-sample forecasting performance

Next, we explore the out-of-sample predictive ability of "excess" information load over one-month-ahead market returns by employing three forecasting exercises. In the first exercise (*Historical*), we regress the excess returns on *InfOver*(2sd) and compare the predicted value with the historical moving-average excess returns. Second, in the *Base* exercise, we examine whether *InfOver*(2sd) has any incremental predictive ability over the other regressors introduced in Section 3.2. Accordingly, we compare the predicted value of excess market returns obtained from (29), with a benchmark model that includes all variables but *InfOver*(2sd) in the covariates.

Finally, in the third exercise (*Lasso*), we dynamically select control variables by rolling the least absolute shrinkage and selection operator (LASSO) regressions of Tibshirani (1996) and use this "best model" as the benchmark model. LASSO regressions enable model selection (as it may set many coefficients to zero) and coefficient shrinkage (as the non-zero coefficient estimates are smaller than their OLS counterparts). We then compare the predicted values from the benchmark model with the predicted value of the model that includes *InfOver*(2sd) in addition to the selected covariates.

For each of the three exercises, we use 540 observations (45 years) as a training period, corresponding to about two-thirds of the overall sample size, and run monthly rolling regressions to calculate recursively the error terms over a testing window.

We then evaluate the out-of-sample forecasting ability of information overload by comparing the out-of-sample R^2s and the differences in the mean absolute errors (MAE) of target and benchmark models for each of the three exercises: *Historical, Base,* and *Lasso*. We test the statistical differences in MAE by employing the Diebold and Mariano (1995) test. Following Campbell and Thompson (2008) and Welch and Goyal (2008), we calculate the out-of-sample R^2 as:

$$R^2_{\text{out}} \quad = \quad 1 - \frac{\sum_{t=1}^{T_{\text{test}}-h} \varepsilon^2_{target,t+h}}{\sum_{t=1}^{T_{\text{test}}-h} \varepsilon^2_{bench,t+h}}, \tag{31}$$

where $\varepsilon_{target,t+h}$ and $\varepsilon_{bench,t+h}$ are the forecast errors of the target and benchmark models of *Historical, Base,* and *Lasso*; T_{test} is the testing window; and $h = 1$. A positive out-of-sample R^2 indicates that the predictive regression displays a lower average mean squared error than that of the benchmark model so that *InfOver* provides a relatively more accurate forecast.

Table 6 presents the results. Overall, information overload exhibits better out-of-sample performance than either the historical average market excess returns or other predictors. *InfOver* delivers positive out-of-sample R^2s and increases the forecast accuracy with significant differences in MAE in all of the three exercises. It beats the historical mean, which is a strong result because Welch and Goyal (2008) show that historical average returns have better forecasting power for market excess returns than many "popular" predictors. Second, *InfOver* has out-of-sample predictive ability over other predictors of market returns such as market sentiment, dividend yields, and consumption-over-wealth ratio. Including our index in the model increases the forecasting accuracy by about 1%. Finally, *InfOver* has incremental predictive ability even over the "best model". Note that the benchmark model in *Lasso* includes variables only with known forecasting power a priori, thus increasing the bar for our index to add any forecasting power (see, for example, Calomiris and Mamaysky, 2019, for an application).

4.4 Effects of information overload on stock markets: Cross-sectional analysis

So far we have provided evidence for the significant predictive power of information overload on stock market returns. We further hypothesize that such predictability is not necessarily uniform and that information overload has cross-sectional effects. As investors are cognitively constrained, it would be optimal for them to allocate processing capacity to only a certain group of stocks, such as larger stocks or less risky "bond-like" stocks—the so-called "category learning" model of Peng and Xiong (2006). Thus, we expect information overload to exacerbate price (under)reaction for stocks that require a higher level of attention as it aggravates investors' capacity constraints.

Moreover, there is a body of literature that shows that small, unprofitable, high beta stocks are more difficult to arbitrage. For instance, Asquith et al. (2005)

show that small-cap stocks are more likely to be short-sale constrained. Similarly, in Hong and Sraer's (2016) model, high beta stocks are more likely to experience binding short-sale constraints, and thus the negative effects of investor disagreement on future stock returns are concentrated among the high beta stocks. Hence, we test whether the stocks with weaker arbitrage forces are more underpriced and experience higher future returns in periods of information overload, as a plethora of information would make them even more difficult to value amid increased estimation and information risk.

To this end, we form four value-weighted portfolios, where the long leg contains stocks with high values of size, beta, variance, and operating profit. High and low values are based on the 20% top and 20% bottom percentiles of the corresponding firm characteristic, respectively. We obtain data from the Kenneth French data library.

Table 7 reports the estimated coefficients from (29) with the long-short portfolio returns as the dependent variables and $InfOver(2\text{sd})$ as the main independent variable. We examine different specifications, including the Carhart (1997) four-factor model, the Fama-French five-factor model (Fama and French, 2015), and the predictors introduced in Section 3.2.

Column I shows that small stocks have a significantly higher return compared to big stocks following periods of excessive information. For the variance and beta portfolios (Columns II and III), the coefficient of information overload is positive and significant, such that highly volatile and high beta stocks have a higher return than the less volatile and low beta ones. Finally, stocks with low operating profits will have a higher return the next month, compared with stocks with high operating profits (Column IV).

Irrespective of the control variables considered, we find that investors require higher risk premium to hold small, high beta, highly volatile, and unprofitable stocks. A one standard deviation increase in information overload translates into 45 bps higher future returns for small stocks than big ones, 89 bps higher future returns for more volatile stocks, 76 bps higher future returns for high beta stocks, and 48 bps higher future returns for unprofitable stocks than profitable ones.

4.5 Robustness

To test the sensitivity of our findings, we run several robustness tests. First, instead of calculating information load by calculating the average of quantity and inconsistency components as introduced in (2), we calculate it by employing a principal component analysis and include the two components separately. Second, while identifying the excessive-flow days, we calculate the threshold using two months of historical data, instead of two weeks.

Third, Garcia (2013) finds that investors' sensitivity to news is most pronounced when they are going through hard times. To test whether our results are driven only by the recession periods, we control for the NBER recessions.

Fourth, instead of calculating stock market volatility using the standard deviation of returns in (29), we calculate it using a GARCH(1,1) model.

Table 8 (Columns I to VII) presents the results. We conclude that our findings are robust to the specifications we consider and that information overload is a significant predictor of stock market excess returns. Although using the average quantity and consistency of news together predicts returns, considering these components separately does not have any predictive power.

Furthermore, we investigate the effects of information overload during different time periods. We have the information overload index from January 1885. However, stock market index data begin in July 1926, and in the baseline specifications we cover the period from March 1952 to December 2018 due to the availability of the control variables. In this subperiod robustness analysis, we keep only the variables with available data so that we can examine the predictive power of information overload on market excess returns in a historical setting. Specifically, we consider five subperiods: the early period (1926–1945), post-World War II (post-1946), and post-1960, 1980, and 1990.

Column VIII shows that, in the early period, news flows do not affect stock market returns. This result is not surprising, given that stock markets were available only to the wealthiest investors and they become a common investment alternative only after the second world war. The main results are robust and qualitatively similar starting after the WWII period (Columns IX through XI). Indeed, the economic effect gets stronger over time until the 1990s. Post-1990s (Column XIIs), the predictive power of information overload on stock market returns gets weaker in

statistical terms, but the economic effect is still meaningful. This finding is in line with the introduction of other news outlets and the massive use of social media and the internet during that period.

5 Conclusion

We construct a news-based historical information load index by considering over 2 million articles printed in *The New York Times* from January 1, 1885 to December 31, 2018. We use these novel data to study the effects of information load on stock market dynamics.

We start by verifying that our index accurately measures the load of information investors face. It captures expected changes in the information flow throughout the sample period, and it is significantly correlated with various proxies of investor attention, asymmetric information, indicators of financial uncertainty, and stress, including the news-based EPU, GPR, and NVIX.

We then show that the effects of information load on stock market dynamics are nonlinear. Moderate information flow increases the decision accuracy and attention of the investor and boosts trading volume, but it does not affect returns. Excessive news-flow periods are followed by higher market returns, as information overload puts constraints on investors' processing capacity. We argue that such results are consistent with information overload increasing information and estimation risk and deteriorating investors' decision accuracy because of their limited attention.

The effects of information overload on market returns are economically meaningful and long-lasting. Moreover, our information load index is useful in improving out-of-sample forecasts has cross-sectional predictive power over market returns.

Overall, in this paper, we provide supporting evidence on the significant role of media in stock markets. In addition to the extant literature, we show that the excess flow of information predicts stock market returns. Our findings challenge traditional asset pricing theories by noting that the flow of information affects investment decisions, plays a role in the time-series and cross-section of returns, and thus prices cannot always incorporate all information instantly.

Appendix

Proof of Lemma 1. The proof follows the standard Bayesian updating process of variables that are normally distributed with known precision (or equivalently, with known variance). See, e.g., Peng and Xiong (2006).

Proof of Proposition 1. The proof follows from obtaining the first-order condition of (20). For example, in the case of investor A:

$$\max_{y_i} E_t[\left(\nu^A | Inf^A - p\right) y_A] - \frac{\theta}{2} Var[\left(\nu^A | Inf^A - p\right) y_A],$$

with first-order condition

$$\left(\hat{\nu}^A - p\right) - \frac{\theta}{\tau} y_A = 0,$$

then

$$y_A = \left(\hat{\nu}^A - p\right) \frac{\tau}{\theta}.$$

Thus, in the case of investor B:

$$y_B = \left(\hat{\nu}^B - p\right) \frac{\tau}{\theta}.$$

By using the market clearing condition:

$$y_A + y_B = M,$$

or

$$\left(\hat{\nu}^A - p\right) \frac{\tau}{\theta} + \left(\hat{\nu}^B - p\right) \frac{\tau}{\theta} = M,$$

which means

$$p = \frac{\hat{\nu}^A + \hat{\nu}^B}{2} - \frac{M\theta}{2\tau}.$$

In the case of asset holdings of agent A:

$$y_A = \left(\hat{\nu}^A - \left(\frac{\hat{\nu}^A + \hat{\nu}^B}{2} - \frac{M\theta}{2\tau}\right)\right) \frac{\tau}{\theta},$$

or equivalently

$$y_A = \frac{M}{2} + \frac{\tau}{2\theta}\left(\hat{\nu}^A - \hat{\nu}^B\right).$$

In the same way, we can obtain asset holdings of agent B:

$$y_B = \frac{M}{2} + \frac{\tau}{2\theta}\left(\hat{\nu}^B - \hat{\nu}^A\right).$$

Q.E.D.

Proof of Lemma 2.Proof for the expected trading volume follows by using the mean of $|\varepsilon^A - \varepsilon^B|$, which is a random variable distributed from a half-normal distribution with variance $2/\tau_\varepsilon$. The mean of a half-normal distribution is $\sigma\sqrt{2/\pi}$; then the mean of $|\varepsilon^A - \varepsilon^B|$ is $\sqrt{2/\tau_\varepsilon}\sqrt{2/\pi} = 2\sqrt{1/(\tau_\varepsilon\pi)}$. Consequently, the expected trading volume is: $\tau_\varepsilon/(2\theta)(2\sqrt{1/(\tau_\varepsilon\pi)}) = (1/\theta)\sqrt{\tau_\varepsilon/\pi}$. Proof for the expected price change follows from substituting the asset prices at $t = 0$ and $t = 1$ inside the expectation operator in equation (28). Q.E.D.

References

Akarsu, S. and Ö. Süer (2021). How investor attention affects stock returns? some international evidence. *Borsa Istanbul Review*.

Amihud, Y. (2002). Illiquidity and stock returns: cross-section and time-series effects. *Journal of Financial Markets 5*, 31–56.

Asquith, P., P. A. Pathak, and J. R. Ritter (2005). Short interest, institutional ownership, and stock returns. *Journal of Financial Economics 78*(2), 243–276.

Baker, M. and J. Wurgler (2006). Investor sentiment and the cross-section of stock returns. *Journal of Finance 61*, 1645–1680.

Baker, S., N. Bloom, and S. Davis (2016). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *Quarterly Journal of Economics 131*, 1593–1636.

Barber, B. M. and T. Odean (2008). All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The review of financial studies 21*(2), 785–818.

Bawden, D. and L. Robinson (2009). The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science 35*(2), 180–191.

Blair, A. (2012). Information overload's 2,300-year-old history. Harvard business review online resources.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment 2008*(10), P10008.

Boudoukh, J., M. Richardson, and R. F. Whitelaw (2008). The myth of long-horizon predictability. *The Review of Financial Studies 21*(4), 1577–1605.

Caldara, D. and M. Iacoviello (2018). Measuring geopolitical risk. International Finance Discussion Papers, 1222.

Calomiris, C. W. and H. Mamaysky (2019). How news and its context drive risk and returns around the world. *Journal of Financial Economics 133*(2), 299–336.

Campbell, J. (1987). Stock returns and the term structure. *Journal of Financial Economics 18*, 373–399.

Campbell, J. Y. and S. Thompson (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies 21*, 1509–1531.

Carhart, M. M. (1997). On persistence in mutual fund performance. *The Journal of finance 52*(1), 57–82.

Chan, W. S. (2003). Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics 70*(2), 223–260.

Coles, J. and U. Loewenstein (1988). Equilibrium pricing and portfolio composition in the presence of uncertain parameters. *Journal of Financial Economics 22*, 279–303.

Coles, J., U. Loewenstein, and J. Suay (1995). On equilibrium pricing under parameter uncertainty. *The Journal of Financial and Quantitative Analysis 30*, 347–374.

Da, Z., J. Engelberg, and P. Gao (2011). In search of attention. *The Journal of Finance 66*(5), 1461–1499.

DellaVigna, S. and J. M. Pollet (2009). Investor inattention and friday earnings announcements. *The Journal of Finance 64*(2), 709–749.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics 13*, 253–263.

Easley, D. and M. O'hara (2004). Information and the cost of capital. *The journal of finance 59*(4), 1553–1583.

Edmunds, A. and A. Morris (2000). The problem of information overload in business organisations: a review of the literature. *International journal of information management 20*(1), 17–28.

Eeckhoudt, L., C. Gollier, and H. Schlesinger (2011). Economic and financial decisions under risk. In *Economic and Financial Decisions under Risk*. Princeton University Press.

Eppler, M. and J. Mengis (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, mis, and related disciplines. *The Information Society 20*, 325–344.

Fama, E. and K. French (1988). Dividend tields and expected stock returns. *Journal of Financial Economics 22*, 3–25.

Fama, E. and K. French (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics 25*, 23–49.

Fama, E. F. and K. R. French (2015). A five-factor asset pricing model. *Journal of financial economics 116*(1), 1–22.

Foucault, T., J. Hombert, and I. Roşu (2016). News trading and speed. *The Journal of Finance 71*(1), 335–382.

Francis, J., I. Khurana, and R. Pereira (2005). Disclosure incentives and effects on cost of capital around the world. *The Accounting Review 80*, 1125–1162.

Garcia, D. (2013). Sentiment during recessions. *Journal of Finance 68*, 1267–1300.

Gleick, J. (2011). *The information: A history, a theory, a flood*. Vintage.

Grisé, M.-L. and R. B. Gallupe (1999). Information overload: Addressing the productivity paradox in face-to-face electronic meetings. *Journal of Management Information Systems 16*(3), 157–185.

Hirshleifer, D., S. S. Lim, and S. H. Teoh (2009). Driven to distraction: Extraneous events and underreaction to earnings news. *The Journal of Finance 64*(5), 2289–2325.

Hirst, D. E. and P. E. Hopkins (1998). Comprehensive income reporting and analysts' valuation judgments. *Journal of Accounting research 36*, 47–75.

Hodrick, R. J. (1992). Dividend yields and expected stock returns: Alternative procedures for inference and measurement. *The Review of Financial Studies 5*(3), 357–386.

Hong, H. and D. A. Sraer (2016). Speculative betas. *The Journal of Finance 71*(5), 2095–2144.

Hong, H. and J. C. Stein (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of finance 54*(6), 2143–2184.

Hong, H. and J. C. Stein (2003). Differences of opinion, short-sales constraints, and market crashes. *The Review of Financial Studies 16*(2), 487–525.

Hong, H. and J. C. Stein (2007). Disagreement and the stock market. *Journal of Economic perspectives 21*, 109–128.

Hou, K., W. Xiong, and L. Peng (2009). A tale of two anomalies: The implications of investor attention for price and earnings momentum. *Available at SSRN 976394*.

Ingersoll, J. E. (1987). *Theory of Financial Decision Making*. Maryland: Rowman & Littlefield.

Johnston, J. C. and H. Pashler (1998). Attentional limitations in dual-task performance. *Attention*, 155–189.

Kahneman, D. (1973). *Attention and effort*, Volume 1063. Citeseer.

Keim, D. B. and R. F. Stambaugh (1986). Predicting returns in the stock and bond markets. *Journal of Financial Economics 17*, 357–390.

Lettau, M. and S. Ludvigson (2001). Consumption, aggregate wealth and expected stock returns. *Journal of Finance 56*, 815–849.

Leuz, C. and R. Verrecchia (2000). The economic consequences of increased disclosure. *Journal of Accounting Research 38*, 91–124.

Loughran, T. and B. McDonald (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance 66*, 35–65.

Loughran, T. and B. McDonald (2014). Measuring readability in financial disclosures. *Journal of Finance 69*, 1643–1671.

Manela, A. and A. Moreira (2017). News implied volatility and disaster concerns. *Journal of Financial Economics 123*(1), 137–162.

Miller, J. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*, 81–97.

Muslu, V., S. Radhakrishnan, K. Subramanyam, and D. Lim (2015). Disclosures and the information environment. *Management Science 61*, 931–948.

Peng, L. and W. Xiong (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics 80*(3), 563–602.

Roetzel, P. G. (2019). Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Business Research 12*, 479–522.

Schroder, H. M., M. J. Driver, and S. Streufert (1967). *Human information processingIndividuals and groups functioning in complex social situations*. New York: Holt, Rinehart, & Winston.

Shiller, R. (1978). Stock prices and social dynamics. *Brookings Papers on Economic Activity 2*, 457–510.

Simon, H. and A. Newell (1971). Human problem solving: The state of the theory in 1970. *American Psychologist 26*, 145–159.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics 50*(3), 665–690.

Tetlock, P. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance 62*, 1139–1168.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*(1), 267–288.

Welch, I. and A. Goyal (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies 4*, 226–265.

Table 1: Pearson correlation coefficients
In this table, we present contemporaneous Pearson correlation coefficients (ρ) of the quantity of information (Q), the tone dispersion (INC), and the information load index (*InfLoad*), with proxies of investor attention, asymmetric information, investor uncertainty, financial uncertainty, and finally, news-based measures at a monthly frequency. Q, INC, and *InfLoad* are all introduced in Section 2.1. In row 1, we use the excess search frequency in Google (Search Volume Index) as a proxy of investor attention (ASVI) (Da et al., 2011; Akarsu and Süer, 2021). ASVI is the difference between cross-sectional and time series monthly averages of 579 U.S. firms' SVIs and the last quarter's SVIs. In row 2, we include the market excess correlation (CORR), obtained as the average monthly correlation of the Fama-French 25 equally-weighted portfolios formed on size and book-to-market in excess of its past 24 months average. In rows 3 and 4, we include the bid-ask spread (SPR) and the distance between the highest ask and lowest bid prices (DIST), respectively. We then include the market volatility (VOLA) in row 5, as the standard deviation of daily CRSP NYSE/AMEX/NASDAQ value-weighted portfolio returns for a given month. In rows 6 and 7, we consider measures of the analysts estimates dispersion using the Institutional Brokers' Estimate System (Refinitiv, IBES North American Summary & Detail Estimates, Level 2, Current & History Data, Adjusted and Unadjusted) summary database. We calculate the standard deviation of the analysts' earnings-per-share (EPS) forecasts for a given stock and time period (DISP1), and the average value of the absolute deviation of the highest EPS forecast from the actual value and the lowest EPS forecast from the actual value (DISP2). We include in rows 8 to 11, the CBOE Volatility Index (VIX) index, the uncertainty measure (BEX) of Bekaert et al. (2019), the Federal Reserve Banks of St. Louis index (FSTR1), and the Chicago financial stress index (FSTR2), respectively. Finally, we compare the *InfLoad* index with other news-based measures: the Economic Policy Uncertainty index (EPU) of Baker et al. (2016), the Geopolitical Risk Index (GPR) of Caldara and Iacoviello (2018), and News-implied volatility index (NVIX) of Manela and Moreira (2017), in rows 12 to 14, respectively. ***, ** and * denote 1%, 5%, and 10% of significance level.

		I	II	III
		ρ_Q	ρ_{INC}	$\rho_{InfLoad}$
1	ASVI	-0.053	-0.345***	-0.265**
2	CORR	0.061**	0.082***	0.082***
3	SPR	0.589***	0.306***	0.612***
4	DIST	0.240***	0.404***	0.420***
5	VOLA	0.196***	0.211***	0.240***
6	DISP1	0.041	0.310***	0.240***
7	DISP2	0.045	0.168***	0.150***
8	VIX	0.203***	0.368***	0.411***
9	BEX	-0.050	0.546***	0.320***
10	FSTR1	0.352***	0.397***	0.479***
11	FSTR2	0.371***	0.176***	0.439***
12	EPU	0.017	0.399***	0.296***
13	GPR	0.047*	0.021	0.049*
14	NVIX	0.046	0.346***	0.197***

Table 2: Summary statistics

This table reports the mean, minimum, maximum, standard deviation, the first and second autocorrelation, and the $p-$values corresponding to the Philips-Perron stationarity test results of the variables included in our analysis. The last three rows report the correlation of the series indicated at the column header with the information overload measures. $InfOver_t$(mean), $InfOver_t$(1sd), $InfOver_t$(2sd) are the information overload measures introduced in Section 3.2. Monthly excess market returns, rx_t^m are the CRSP NYSE/AMEX/NASDAQ value-weighted portfolio returns in excess of the one-month treasury bill rate. SENT_t is the market sentiment measure, calculated as the difference between the proportion of positive and negative words, following Garcia (2013). DY_t is the S&P 500 monthly dividend yield, ΔSTIR_t is the change in three-month T-bill rates, TS_t is the term spread, calculated as the difference between the ten-year Treasury bond and the three-month Treasury bill yields. Default spread, DS_t, is measured as the difference between BAA and AAA corporate bond spreads, CAY_t is the consumption–wealth ratio of Lettau and Ludvigson (2001) and obtained by using the most recently available quarterly observations, RVOLA_t is realized volatility and calculated as the standard deviation of market stock returns. Finally, ILLIQ_t is the Amihud's illiquidity measure. All of the variables are monthly estimates from March 1952 to December 2018. Data sources: ProQuest, TDM Studio. New York Times Historical Newspapers, Global Financial Data, Inc., GFDatabase, FRED, the Federal Reserve Bank of St Louis, Kenneth French's online data library.

	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
	rx_t^m	$InfOver_t$ (mean)	$InfOver_t$ (1sd)	$InfOver_t$ (2sd)	SENT_t	DY_t	ΔSTIR_t	TS_t	DS_t	CAY_t	RVOLA_t	ILLIQ_t
mean	0.612	0.488	0.153	0.027	-0.402	3.117	0.001	1.642	0.964	0.05	0.798	5.241
min	-23.24	0.25	0.00	0.00	-1.063	1.08	-3.85	-2.4	0.32	-4.945	0.174	0.321
max	16.1	0.742	0.355	0.133	-0.126	6.4	2.4	4.42	3.38	3.801	4.991	59.129
st. Dev.	4.247	0.08	0.057	0.027	0.125	1.2	0.442	1.361	0.434	1.876	0.488	5.741
AC1	0.077	-0.085	-0.257	-0.052	0.744	0.992	0.128	0.959	0.971	0.973	0.675	0.822
AC2	-0.023	0.053	-0.138	0.065	0.696	0.984	-0.024	0.914	0.927	0.946	0.546	0.749
PP–stationarity	0.01	0.01	0.01	0.01	0.01	0.208	0.01	0.01	0.01	0.03	0.01	0.01
$\rho(InfOver(\text{mean}),i)$	-0.103	1.00	0.326	-0.06	-0.057	0.056	-0.017	-0.037	0.099	-0.09	0.064	0.027
$\rho(InfOver(\text{1sd}),i)$	-0.107	0.326	1.00	0.219	-0.019	-0.012	-0.036	0.064	0.065	0.022	0.09	0.047
$\rho(InfOver(\text{2sd}),i)$	-0.028	-0.06	0.219	1.00	0.034	-0.064	-0.026	0.029	-0.142	0.065	0.013	-0.052

Table 3: Information overload and market excess returns
In this table, we report the estimated coefficients of the time-series regressions introduced in (29) for $h = 1$. Volume is the S&P 500 trading volume, demeaned and log-transformed. The rest of the variables are described in Table 2. Columns I and II presents the results when the main independent variable is information load and the rest of the columns when it is information overload. In columns III and IV, information overload considers the days when the information load index is above the historical mean. In columns V and VI, we use a one-standard-deviation band, and in columns VII and VIII, we use two-standard-deviation bands. All of the variables are in monthly frequency from March 1952 to December 2018, where available and standardized to ease the interpretation of the coefficients. Newey-West standard errors are reported. ***, **, and * denote significance at the 1%, 5%, and 10% level (two-sided), respectively. Data sources: ProQuest, TDM Studio. New York Times Historical Newspapers, Global Financial Data, Inc., GFDatabase, the Federal Reserve Bank of St Louis, Kenneth French's online data library.

	I	II	III	IV	V	VI	VII	VIII
Dep. var.	rx^m_{t+1}	volume	rx^m_{t+1}	volume	rx^m_{t+1}	volume	rx^m_{t+1}	volume
	InfLoad$_t$		*InfOver*$_t$(mean)		*InfOver*$_t$(1sd)		*InfOver*$_t$(2sd)	
InfLoad/Over$_t$	0.03	-0.83***	-0.19	0.10**	0.30**	0.05	0.35***	-0.11***
	(0.199)	(0.099)	(0.147)	(0.041)	(0.150)	(0.035)	(0.132)	(0.043)
SENT$_t$	-0.24	-0.14*	-0.23	-0.01	-0.25	-0.00	-0.23	-0.00
	(0.201)	(0.079)	(0.201)	(0.090)	(0.200)	(0.091)	(0.201)	(0.090)
DY$_t$	0.23	-1.63***	0.26	-2.15***	0.25	-2.14***	0.24	-2.14***
	(0.241)	(0.124)	(0.208)	(0.103)	(0.210)	(0.103)	(0.208)	(0.103)
ΔSTIR$_t$	-0.40*	0.11**	-0.40*	0.11*	-0.39	0.10*	-0.39	0.10*
	(0.237)	(0.053)	(0.235)	(0.055)	(0.237)	(0.054)	(0.237)	(0.053)
TS$_t$	0.31*	0.23***	0.30*	0.38***	0.30*	0.38***	0.28*	0.38***
	(0.174)	(0.074)	(0.164)	(0.073)	(0.165)	(0.073)	(0.165)	(0.073)
DS$_t$	0.10	0.79***	0.10	0.93***	0.08	0.93***	0.16	0.91***
	(0.234)	(0.079)	(0.232)	(0.088)	(0.229)	(0.088)	(0.232)	(0.089)
CAY$_t$	0.37*	0.38***	0.35*	0.36***	0.37**	0.35***	0.36*	0.35***
	(0.189)	(0.068)	(0.189)	(0.083)	(0.189)	(0.083)	(0.189)	(0.083)
RVOLA$_t$	-0.53**	0.20***	-0.51**	0.09	-0.55**	0.09	-0.55**	0.10
	(0.248)	(0.067)	(0.249)	(0.075)	(0.250)	(0.075)	(0.254)	(0.075)
ILLIQ$_t$	0.17		0.18		0.17		0.19	
	(0.169		(0.170)		(0.171)		(0.168)	
Adj. R^2 (%)	3.22	73.15	2.74	73.10	3.07	72.97	3.22	73.15
N Obs.	792	790	792	790	792	790	792	790

Table 4: Long-run effects of information overload on market excess returns

In this table, we report the estimated coefficients of the time-series regressions introduced in (29) for $h = 1, 2, ..., 24$. $InfOver_t$(2sd) is the information overload measure introduced in Section 3.2, where it considers the days when the information load index is above its historical two-standard-deviation bands. The rest of the variables are described in Table 2. Information overload is standardized to ease the interpretation of the coefficients. All of the variables are monthly frequency from March 1952 to December 2018. We report Hodrick (1992) standard errors to address overlapping observations, which induces serial correlation in the residuals. ***, **, and * denote significance at the 1%, 5%, and 10% level (two-sided), respectively. Data sources: ProQuest, TDM Studio. New York Times Historical Newspapers, Global Financial Data, Inc., GFDatabase, the Federal Reserve Bank of St Louis, Kenneth French's online data library.

horizon (h)	1	2	3	4	5	6	7	8	9	10	11	12
$InfOver_t$(2sd)	0.347**	0.199**	0.241***	0.276***	0.232***	0.202***	0.194***	0.183***	0.148***	0.130**	0.115**	0.128***
	(0.132)	(0.099)	(0.084)	(0.076)	(0.071)	(0.067)	(0.064)	(0.060)	(0.058)	(0.058)	(0.054)	(0.051)
Adj. R^2 (%)	3.22	3.53	5.36	7.53	8.99	10.35	12.31	14.30	15.85	17.58	19.30	20.67
N Obs.	792	791	790	789	788	787	786	785	784	783	782	781
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
horizon (h)	13	14	15	16	17	18	19	20	21	22	23	24
$InfOver_t$	0.117**	0.10**	0.101**	0.095**	0.101**	0.094**	0.083*	0.083*	0.066	0.062	0.061	0.059
	(0.051)	(0.051)	(0.050)	(0.049)	(0.047)	(0.045)	(0.044)	(0.043)	(0.041)	(0.042)	(0.041)	(0.041)
Adj. R^2 (%)	21.93	23.09	23.91	25.53	26.92	27.96	29.02	30.41	31.37	32.16	33.20	34.12
N Obs.	780	779	778	777	776	775	774	773	772	771	770	769
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table 5: Predictive ability of *InfOver* and other text-based measures
In this table, we examine the explanatory power of information overload on future market returns in comparison to other text-based measures. We run the time-series regressions introduced in (29) for $h = 1$, of market excess returns on $InfOver_t$(2sd). SENT is the market sentiment via Garcia (2013), EPU is the Economic Policy Uncertainty index of Baker et al. (2016), GPR is the Geopolitical Risk Index of Caldara and Iacoviello (2018), and NVIX is the News-implied Volatility Index of Manela and Moreira (2017). All of the other variables are introduced in Table 2. All of the variables are in monthly frequency from March 1952 to December 2018, where available and standardized to ease the interpretation of the coefficients. Newey-West standard errors are reported. ***, **, and * denote significance at the 1%, 5%, and 10% level (two-sided), respectively. Data sources: ProQuest, TDM Studio. New York Times Historical Newspapers, Global Financial Data, Inc., GFDatabase, the Federal Reserve Bank of St Louis, Kenneth French's online data library.

	I	II	III	IV	V	VI	VII	VIII	IX
$InfOver_t$(2sd)	0.32**		0.36***						0.39**
	(0.131)		(0.131)						(0.193)
$SENT_t$				-0.25				-0.10	-0.08
				(0.207)				(0.409)	(0.403)
EPU_t					0.73**			0.57*	0.59*
					(0.306)			(0.346)	(0.344)
GPR_t						0.17		0.24	0.21
						(0.178)		(0.233)	(0.240)
$NVIX_t$							0.27	0.23	0.25
							(0.205)	(0.289)	(0.290)
DY_t		0.23	0.22	0.22	-0.10	0.19	0.27	0.03	0.03
		(0.177)	(0.175)	(0.178)	(0.529)	(0.178)	(0.180)	(0.602)	(0.599)
$\Delta STIR_t$		-0.42*	-0.42*	-0.41*	0.29	-0.42*	-0.42*	0.41	0.38
		(0.246)	(0.245)	(0.247)	(0.487)	(0.247)	(0.247)	(0.512)	(0.512)
TS_t		0.31*	0.28	0.32*	-0.25	0.29*	0.26	-0.33	-0.36
		(0.174)	(0.174)	(0.171)	(0.273)	(0.173)	(0.177)	(0.306)	(0.308)
DS_t		0.11 0.18	0.09	-0.14	0.11	0.13	-0.24	-0.14	
		(0.233)	(0.235)	(0.234)	(0.466)	(0.235)	(0.238)	(0.541)	(0.530)
CAY_t		0.26	0.25	0.37*	0.23	0.27	0.30*	0.26	0.27
		(0.165)	(0.164)	(0.190)	(0.257)	(0.166)	(0.171)	(0.352)	(0.349)
$RVOLA_t$		-0.43	-0.46*	-0.53**	-0.57**	-0.42	-0.55**	-0.63**	-0.71**
		(0.264)	(0.265)	(0.255)	(0.285)	(0.263)	(0.272)	(0.291)	(0.294)
$ILLIQ_t$		0.21	0.23	0.18	0.34	0.22	0.19	0.28	0.27
		(0.176)	(0.173)	(0.177)	(0.229)	(0.175)	(0.173)	(0.237)	(0.232)
Adj. R^2 (%)	0.44	2.61	3.17	2.67	1.97	2.63	2.82	1.42	1.97
N Obs.	800	792	792	792	407	792	759	374	374

Table 6: Out-of-Sample Predictive Ability
The table reports the results of three exercises to assess the out-of-sample performance of information overload over one-month-ahead forecasts of market excess returns. In the first exercise, *Historical*, we regress the excess returns on *InfOver*(2sd) and compare the predicted value with the historical moving-average excess returns. In the second exercise, *Base*, we compare the predicted value of excess market returns obtained from estimating (29) with a benchmark model that includes all controls but *InfOver*(2sd) in the covariates. Finally, in the third exercise *Lasso*, the benchmark model is the "best model" obtained by dynamically selecting control variables via rolling LASSO regressions. The training period is 45 years. We report the out-of-sample R^2 defined in (31) and the difference in mean absolute errors (ΔMAE). We test the statistical differences by employing the Diebold and Mariano (1995) test and report the corresponding p-value. ***, **, and * denote significance at the 1%, 5%, and 10% level, respectively.

Exercise type	N obs.	out-of-sample R^2 (%)	ΔMAE	p-value
Historical	795	1.05	0.043**	0.0143
Base	795	1.29	0.055***	0.0053
Lasso	795	1.02	0.043**	0.0293

Table 7: Effects of information overload on cross-section of returns
In this table, we test the effects of information overload over the cross-section of returns. In Columns I through IV, the dependent variables are returns of portfolios formed by the big minus small stocks, high variance minus low variance stocks, high beta minus low beta stocks, and high operating profit minus low operating profit stocks, respectively. $InfOver_t$(2sd) is the information overload measure introduced in Section 3.2, where it considers the days when the information load index is above the historical two-standard-deviation bands. SENT+*Controls* are the covariates of model (29). *Carhart-4* is factor model of Carhart (1997). *FF-5* is the Fama-French 5 factors of Fama and French (2015). The Fama-French factors are constructed using the 6 value-weight portfolios formed on size and book-to-market, the 6 value-weight portfolios formed on size and operating profitability, and the 6 value-weight portfolios formed on size and investment. The momentum factor uses six value-weight portfolios formed on size and prior (2-12) returns. All of the variables are in monthly frequency from March 1952 to December 2018. Newey-West standard errors are reported. ***, **, and * denote significance at the 1%, 5%, and 10% level (two-sided), respectively. Data sources: ProQuest, TDM Studio. New York Times Historical Newspapers, Global Financial Data, Inc., GFDatabase, FRED, the Federal Reserve Bank of St Louis, Kenneth French's online data library.

Portfolio type	I size	II variance	III beta	IV op. profit
Panel A				
$InfOver_t$(2sd)	-0.45**	0.89***	0.76***	-0.48***
	(0.183)	(0.277)	(0.234)	(0.134)
Adj. R^2 (%)	0.59	1.95	2.44	2.15
N Obs.	792	657	648	657
SENT+*Controls*	Yes	Yes	Yes	Yes
Carhart-4	No	No	No	No
FF-5	No	No	No	No
Panel B				
$InfOver_t$(2sd)	-0.39**	0.90***	0.70***	-0.43***
	(0.163)	(0.264)	(0.220)	(0.133)
Adj. R^2 (%)	7.27	3.54	3.14	1.76
N Obs.	872	663	654	663
SENT+*Controls*	No	No	No	No
Carhart-4	Yes	Yes	Yes	Yes
FF-5	No	No	No	No
Panel C				
$InfOver_t$(2sd)	-0.54**	0.90***	0.70***	-0.42***
	(0.211)	(0.265)	(0.222)	(0.129)
Adj. R^2 (%)	7.88	3.90	3.29	3.79
N Obs.	662	662	654	662
SENT+*Controls*	No	No	No	No
Carhart-4	No	No	No	No
FF-5	Yes	Yes	Yes	Yes
Panel D				
$InfOver_t$(2sd)	-0.54**	0.90***	0.70***	-0.42***
	(0.212)	(0.264)	(0.221)	(0.129)
Adj. R^2 (%)	7.83	3.75	3.15	3.67
N Obs.	662	662	654	662
SENT+*Controls*	Yes	Yes	Yes	Yes
Carhart-4	Yes	Yes	Yes	Yes
FF-5	Yes	Yes	Yes	Yes

Table 8: Robustness

In this table, we present the results of the robustness analysis. Column I reports the baseline specification. In Column II, we calculate information load via principal component analysis instead of taking the average of quantity and inconsistency components. In Columns III and IV, we report the results when we include the quantity and inconsistency components separately. In Column V, we calculate the threshold using two months of historical data, instead of two weeks. In Column VI we include the NBER recession dates in the control set. In Column VII, instead of calculating volatility as the standard deviation of daily market returns, we employ the GARCH(1,1) model. Finally, in Columns VIII–XII, we repeat the baseline specification using data for the early period (1926–1945), from 1946, from 1960, from 1980, and from 1990, respectively. $InfOver_t$(2sd) is the information overload measure introduced in Section 3.2, where it considers the days when the information load index is above the historical two-standard-deviation bands. The rest of the variables are described in Table 2. All of the explanatory variables are standardized to ease the interpretation of the coefficients. All of the variables are in monthly frequency from March 1952 to December 2018, except for the specifications in columns VIII–XII. Newey-West standard errors are reported. ***, **, and * denote significance at the 1%, 5%, and 10% level (two-sided), respectively. Data sources: ProQuest, TDM Studio. New York Times Historical Newspapers, Global Financial Data, Inc., GFDatabase, the Federal Reserve Bank of St Louis, Kenneth French's online data library.

Specification:	I baseline	II pca	III Q	IV INC	V 2months	VI NBER	VII GARCH	VIII 1926–1945	IX Post–1946	X Post–1960	XI Post–1980	XII Post–1990
$InfOver_t$(2sd)	0.35***	0.35***	0.24	0.14	0.34**	0.35***	0.32**	-0.07	0.32**	0.39***	0.38**	0.34*
	(0.132)	(0.131)	(0.166)	(0.152)	(0.133)	(0.132)	(0.135)	(0.525)	(0.127)	(0.144)	(0.177)	(0.191)
$SENT_t$	-0.24	-0.24	-0.26	-0.24	-0.28	-0.31	0.06	1.54*	-0.04	-0.07	-0.21	-0.23
	(0.205)	(0.205)	(0.207)	(0.205)	(0.206)	(0.201)	(0.255)	(0.876)	(0.179)	(0.201)	(0.319)	(0.394)
DY_t	0.23	0.23	0.28	0.23	0.20	0.35*	0.42**	-0.35	0.34**	0.10	0.13	-0.05
	(0.197)	(0.197)	(0.204)	(0.197)	(0.199)	(0.206)	(0.190)	(1.177)	(0.136)	(0.235)	(0.395)	(0.475)
$\Delta STIR_t$	-0.40	-0.40	-0.39	-0.41*	-0.38	-0.46*	-0.38	1.64	-0.39*	-0.45*	-0.14	0.03
	(0.243)	(0.243)	(0.243)	(0.242)	(0.239)	(0.242)	(0.243)	(1.300)	(0.233)	(0.269)	(0.275)	(0.282)
TS_t	0.29*	0.29*	0.31*	0.31*	0.30*	0.25	0.32*	-0.08	0.33**	0.28	0.14	-0.05
	(0.169)	(0.169)	(0.169)	(0.169)	(0.169)	(0.169)	(0.170)	(0.892)	(0.163)	(0.186)	(0.215)	(0.249)
DS_t	0.16	0.16	0.10	0.10	0.10	0.22	-0.17	-0.91	0.08	0.21	0.06	-0.09
	(0.235)	(0.235)	(0.232)	(0.233)	(0.232)	(0.236)	(0.246)	(1.343)	(0.207)	(0.276)	(0.452)	(0.503)
CAY_t	0.36*	0.36*	0.34*	0.37**	0.36*	0.36*	0.16					
	(0.189)	(0.189)	(0.188)	(0.189)	(0.190)	(0.187)	(0.213)					
$RVOLA_t$	-0.56**	-0.56**	-0.56**	-0.53**	-0.58**	-0.49*	0.25	0.33	-0.36	-0.54*	-0.66**	-0.61
	(0.256)	(0.256)	(0.253)	(0.253)	(0.256)	(0.265)	(0.224)	(0.835)	(0.244)	(0.285)	(0.335)	(0.430)
$ILLIQ_t$	0.20	0.20	0.17	0.18	0.17	0.22	0.11	1.53	0.25	0.34*	0.39*	0.60
	(0.172)	(0.172)	(0.177)	(0.175)	(0.172)	(0.172)	(0.177)	(1.115)	(0.162)	(0.184)	(0.226)	(0.398)
$NBER_t$						-1.16*						
						(0.607)						
Adj. R^2 (%)	3.22	3.21	2.86	2.66	3.19	3.79	2.35	4.44	2.67	2.87	1.35	0.9
N Obs.	792	792	792	792	792	792	785	228	865	697	462	346