ARTICLE IN PRESS
JID: COMCOM [m5G;July 8, 2015;20:35]

Computer Communications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Twitter volume spikes and stock options pricing

Wei Wei [a], Yuexin Mao [b,*], Bing Wang [b]

[a] FinStats.com, United States
[b] Computer Science & Engineering Department, University of Connecticut, United States

ARTICLE INFO

Article history:
Available online xxx

Keywords:
Twitter
Stock
Option
Twitter volume spikes
Stock options trading

ABSTRACT

The stock market is a popular topic in Twitter. The number of tweets concerning a stock varies over days, and sometimes exhibits a significant spike. In this paper, we investigate the relationship between Twitter volume spikes and stock options pricing. We start with the underlying assumption of the Black–Scholes model, the most widely used model for stock options pricing, and investigate when this assumption holds for stocks that have Twitter volume spikes. We find that the assumption is less likely to hold in the time period before a Twitter volume spike, and is more likely to hold afterwards. In addition, the volatility of a stock is significantly lower after a Twitter volume spike than that before the spike. We also find that implied volatility increases sharply before a Twitter volume spike and decreases quickly afterwards. In addition, put options tend to be priced higher than call options. Last, we find that right after a Twitter volume spike, options may still be overpriced. Based on the above findings, we propose a put spread selling strategy for stock options trading. Realistic simulation of a portfolio using one year stock market data demonstrates that, even in a conservative setting, this strategy achieves a 34.3% gain when taking account of commissions and ask-bid spread, while S&P 500 only increases 12.8% in the same period.

1. Introduction

Twitter has rapidly gained popularity since its creation in March 2006. As of July 2014, it has more than 500 million users, with more than 271 million being active users [1]. The stock market is a popular topic in Twitter. Many traders, investors, financial analysts and news agencies post tweets about the stock market in Twitter, which may be further retweeted. As a result, there can be thousands of tweets each day related to certain stocks. In general, the number of tweets concerning a stock varies over days, and sometimes exhibits a significant spike, indicating a sudden increase of interests in the stock. Since a collection of tweets reflect the collective wisdom of the users who post the tweets, a Twitter volume spike about a stock may contain important information regarding the stock. In this paper, we investigate the relationship of Twitter volume spikes and stock options pricing. The reason for focusing on stock options is because they are valuable investment vehicles but are very difficult to understand [23]. Our goal is to investigate whether Twitter volume spikes can shed light on the behavior of stock options pricing, and whether the insights thus obtained can help to assist stock options trading.

A stock option is a financial contract that gives the owner the right, but not the obligation, to buy or sell an underlying asset (stock) at a specified strike price on or before a specified date. Specifically, *call option* gives the owner the right to buy a stock; *put options* give the owner the right to sell a stock. The Black–Scholes model is the most widely used model for stock options pricing. It has led to a boom in options trading ever since it was introduced in 1970s. We start from the underlying assumption of the Black–Scholes model, i.e., stock price follows a geometric Brownian motion and hence stock return follows a lognormal distribution, and investigate when this assumption holds for stocks that have Twitter volume spikes. We then proceed to investigate implied volatility (derived from the Black–Scholes model) as well as the actual volatility around a Twitter volume spike. Our results demonstrate that Twitter volume spikes can be very helpful in understanding stock options pricing. In addition, using Twitter volume spikes, one can design highly profitable options trading strategies. Our main contributions are:

- We find that in a time period with a Twitter volume spike, stock return is less likely to follow a lognormal distribution, indicating that Twitter volume spikes are correlated with extreme changes in stock prices. On the other hand, for a short time period after a Twitter volume spike, the lognormal assumption is likely to hold. In addition, the volatility of a stock is significantly lower after a Twitter volume spike than that before the spike. We further investigate stock price model selection, and find that a three-parameter model that uses the same drift and different volatilities before and

* Corresponding author. Tel.: +1 860 486 3411; fax: +1 860 486 4817.
E-mail address: yuexin.mao@engr.uconn.edu (Y. Mao).

http://dx.doi.org/10.1016/j.comcom.2015.06.018
0140-3664/© 2015 Published by Elsevier B.V.

Please cite this article as: W. Wei et al., Twitter volume spikes and stock options pricing, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.06.018

after a Twitter volume spike provides the highest gain in the likelihood value.

- We find a clear pattern in implied volatility (IV) around a Twitter volume spike. Specifically, IV increases sharply before a Twitter volume spike and decreases quickly afterwards. Furthermore, IV of put options tends to be larger than IV of call options. We also find that the volatility around a Twitter volume spike is particularly high. In addition, options may still be overpriced right after a Twitter volume spike. This is particularly true for put options, which confirms that people tend to strongly prefer avoiding losses to acquiring gains [14].
- Based on our findings, we propose a put spread selling strategy for stock options trading. Realistic simulation of a portfolio using one year stock market data demonstrates that, even in a conservative setting, this strategy achieves a 34.3% gain when taking account of commissions and ask-bid spread, while S&P 500 only increases 12.8% in the same period.

While several studies relate social media and the financial market (e.g., [8,18,19,21,29], see more details in Section 2), to the best of our knowledge, our study is the first that analyzes the relationship between Twitter volume spikes and stock options pricing. Our results indicate that social media can be a powerful tool to help understand the behavior of stock options, and further assist the trading of stock options.

The rest of the paper is organized as follows. Section 2 briefly reviews related work. Section 3 describes how we collect data and identify Twitter volume spikes. Section 4 briefly describes the lognormal stock price model and the Black–Scholes model. Section 5 analyzes the relationship between Twitter volume spikes and stock price model. Section 6 analyzes the relationship between Twitter volume spikes and stock options pricing. Section 7 presents a stock options trading strategy and evaluates its performance. Section 8 briefly discusses the choice of threshold for identifying Twitter volume spikes. Last, Section 9 concludes the paper and presents future work.

2. Related work

Existing studies on Twitter have investigated the general characteristics of the Twitter social network (e.g., [13,17]) and the social interactions within Twitter [11]. Several studies use tweets to predict real-world events such as earthquakes [26], box-office revenues of movies [4,9], seasonal influenza [2], the popularity of a news article [5], and popular messages in Twitter [10].

The studies that are closest to ours are those that relate Twitter to the financial market. Kanungsukkasem et al. [16] propose a method to recognize NASDAQ stock symbols in a stream of tweets. Bar-Haim et al. [6] predict stock price movement by analyzing tweets to find expert investors and collect experts' opinions. Several studies use Twitter sentiment data to predict the stock market. Bollen et al. [8] find that specific public mood states in Twitter are significantly correlated with the Dow Jones Industrial Average (DJIA), and thus can be used to forecast the direction of DJIA changes. Zhang et al. [29] find that emotional tweet percentage is correlated with DJIA, NASDAQ and S&P 500. Later on, Mao et al. [19] find that Twitter sentiment indicator and the number of tweets that mention financial terms in the previous 1–2 days can be used to predict the daily market return. Makrehchi et al. [18] propose an approach that uses event based sentiment tweets to predict the stock market movement, and develop a stock trading strategy that outperforms the baseline. In our prior study [20], we find that the daily number of tweets that mention S&P 500 stocks is correlated with certain stock market indicators at three different levels, from the stock market, to industry sector, and then to individual company stocks. The study [25] also reports the correlation between trading volume and the daily number of tweets for individual company stocks. In another prior study [21], we provide insight on when Twitter volume spikes occur and possible causes of these spikes. Furthermore, we develop a Bayesian classifier that uses Twitter volume spikes to assist stock trading.

Our current study differs from all the above in that we focus on Twitter volume spikes and stock options pricing. To the best of our knowledge, this is the first study that investigates how Twitter volume spikes can be used to understand stock options pricing and assist stock options trading.

3. Methodology

In this section, we describe our methodology of collecting data and identifying Twitter volume spikes.

3.1. Stock market data

We obtain daily stock market data and stock option data for the 500 stocks in the S&P 500 index. For stock market data, we consider stock daily closing price for each stock. For stock option data, we consider the call and put options of a stock. We only consider short term options that will expire in around 30 days.

3.2. Twitter data

In Twitter community, people usually mention a company's stock using the stock symbol prefixed by a dollar sign. For example, $AAPL represents the stock of Apple Inc. When collecting public tweets on S&P 500 stocks, we only search for tweets that follow the above convention (i.e., having a dollar sign before a stock symbol). This is because many stock symbols (e.g., A, CAT, GAS) are common words, and hence using search keywords without the dollar sign will result in a large number of spurious tweets. Unless otherwise stated, the results reported in the paper are based on Twitter data collected over one year, from August 1, 2013 to August 6, 2014.

3.3. Twitter volume spikes

We next describe how we identify Twitter volume spikes. Consider a stock. Roughly, a Twitter volume spike happens when the number of tweets related to the stock is significantly larger than usual. Therefore, one way to identify Twitter volume spikes is as follows. We first obtain the number of tweets for the stock on a day and the average number of tweets for the stock in the past N days. Then if the former is significantly larger than the latter, we say there is a Twitter volume spike. The above approach uses the absolute number of tweets to identify Twitter volume spikes, which may not provide robust identification. For instance, it can lead to false Twitter volume spikes when the numbers of tweets for a large number of stocks are inflated (for instance, due to abuse of some users, as we have observed in the collected data). Therefore, for a stock, instead of using the absolute value of the number of tweets, we use the relative value, i.e., the number of tweets for the stock on a day over the total number of tweets for all S&P 500 stocks on that day, to identify Twitter volume spikes. Specifically, if this relative value is at least K times of the average relative value in the past N days, then we say the stock has a Twitter volume spike. Unless otherwise stated, we use $N = 70$ and $K = 3$ in this paper. In Section 8, we further investigate the choice of K.

The above definition only considers the number of tweets, while does not consider the users who post the tweets. In our context, a large number of tweets about a stock is interesting only if it indicates that many users show significantly increased interests in the stock. Therefore, we add two additional conditions when identifying Twitter volume spikes. First, the number of unique users has to be sufficiently large. Specifically, we say a stock has a Twitter volume spike on a day only if the number of unique users that post the tweets is larger than a threshold. We choose the threshold to be 10 in this paper. Even when the number of unique users is sufficiently large, majority of the tweets can be from a small number of users. To avoid such a scenario, we further require that the tweets have to be from a diverse set of users. Specifically, we define a *user diversity index*, and require that the index to be larger than a threshold. Suppose M unique users tweet about a stock on a day. Let p_i denote the fraction of tweets from user i. Then we define user diversity index as

$$I = \frac{-\sum_{i=1}^{M} p_i \log p_i}{\log M} \tag{1}$$

where the numerator is the entropy, while the denominator is the maximum value of the entropy (i.e., when each of the M users posts the same number of tweets, i.e., $p_i = p_j, \forall i \neq j$). Therefore, $I \in (0, 1]$. Furthermore, it is easy to see that the value of I is independent of the base of the logarithm by applying change of base in the logarithm. In this paper, we say a stock has a Twitter volume spike on a day only if the user diversity index is above a threshold, chosen as 0.4.

In summary, we use three conditions, one on the number of tweets, one on the number of unique users that post the tweets, and the third on the diversity of the users that post the tweets, when identifying Twitter volume spikes. For the Twitter data that we collected (i.e., tweets that contain S&P 500 stock symbols from August 1, 2013 to August 6, 2014), we find that all the 500 stocks have at least one Twitter volume spike, and there are a total of 3288 Twitter volume spikes, which are used in the analysis in the rest of the paper.

4. Background

The Black–Scholes model assumes that stock price follows a geometric Brownian motion. In the following, we first briefly describe the geometric Brownian motion model, and then describe the Black–Scholes model.

4.1. Stock price model

Let S_t denote stock price on day t. Let μ denote the drift rate of the stock, and let σ denote the stock volatility. The most widely used model for stock price [7,12,22,24] is the Geometric Brownian motion model, that is,

$$dS_t = \mu S_t\, dt + \sigma S_t\, dW_t, \tag{2}$$

where W_t is a Brownian motion. On the right hand side of (2), the first term is used to model deterministic trends, while the second one is used to model unpredictable events. For an arbitrary initial value S_0, the stochastic differential equation (2) has the analytic solution

$$S_t = S_0 \exp\left(\left(\mu - \frac{\sigma^2}{2}\right)t + \sigma W_t\right). \tag{3}$$

Let R_t denote the *log return* (i.e., logarithm of stock return) on day t. Then

$$R_t = \ln \frac{S_t}{S_{t-1}} = \left(\mu - \frac{1}{2}\sigma^2\right) + \sigma(W_t - W_{t-1}), \tag{4}$$

where $W_t - W_{t-1}$ is the usual Brownian increment that follows a normal distribution. The above shows that when assuming stock price follows a Geometric Brownian motion, log return follows a normal distribution, or stock return follows a lognormal distribution. Given m samples of log returns, denoted as $\{R_1, \ldots, R_m\}$, the two parameters, μ and σ, can be estimated empirically as

$$\mu = \frac{\sum_{t=1}^{m} R_t}{m}, \quad \sigma = \sqrt{\frac{\sum_{t=1}^{m} (R_t - \overline{R})}{m-1}} \tag{5}$$

where $\overline{R}$ is the mean of the m samples.

4.2. Black–Scholes model for stock option pricing

As we briefly mentioned earlier, a stock option is a financial contract that gives the buyer (owner) the right to buy or sell an underlying asset at a specified price (*strike price*) on or before a specified date (*expiration date*) [28]. Stock options are in two categories: *call options* and *put options*. A call option of a stock gives the buyer the right to buy the stock at the strike price; a put option gives the buyer the right to sell the stock at the strike price.

The Black–Scholes model [7] is a widely used mathematical model for estimating the price of a stock option. In its basic form, it assumes that the market consists of a risky asset (i.e., a stock) and a riskless asset. The rate of return on the riskless asset is constant, and thus called the risk-free interest rate, denoted as r. The stock does not pay a dividend, and its price follows a Geometric Brownian motion with drift μ and volatility σ. There is no arbitrage opportunity (i.e., there is no way to make a riskless profit). The market is frictionless (i.e., transactions do not incur any fees or costs).

Let t denote time. Let S_t denote the stock price at time t, which is as modeled in (2). Let $V(S, t)$ be the price of the stock option, which is a function of time t and stock price S. The Black–Scholes equation is a partial differential equation that describes the price of the option over time. Specifically, it is

$$\frac{\partial V}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2 V}{\partial S^2} + rS\frac{\partial V}{\partial S} - rV = 0, \tag{6}$$

where we write $V(S, t)$ simply as V and S_t as S for ease of notation.

The Black–Scholes equation can be used to estimate the price of call and put options. Let T denote its expiration date. Let E denote the strike price of the option. If the option is a call option, it has a payoff of $S_T - E$ if S_T is larger than E. Otherwise, the payoff is zero. That is, the payoff is

$$\max(S_T - E, 0)$$

Using the above condition and the Black–Scholes equation, the price of the call option at time t is

$$S_t N(d_1) - e^{-r(T-t)} E N(d_2), \tag{7}$$

where $N(d)$ is the cumulative distribution function of the standard normal distribution, and

$$d_1 = \frac{\ln \frac{S_t}{E} + \left(r + \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}}, \tag{8}$$

$$d_2 = \frac{\ln \frac{S_t}{E} + \left(r - \frac{\sigma^2}{2}\right)(T-t)}{\sigma\sqrt{T-t}} = d_1 - \sigma\sqrt{T-t}. \tag{9}$$

If the option is a put option, it has a payoff of $E - S_T$ if S_T is smaller than E. Otherwise, the payoff is zero. That is, the payoff is

$$\max(E - S_T, 0)$$

Using the above condition and the Black–Scholes equation, the price of the put option at time t is

$$-S_t N(-d_1) + e^{-r(T-t)} E N(-d_2), \tag{10}$$

where $N(d)$, d_1 and d_2 are as defined earlier.

While the above model assumes no dividend, the case with dividend can also be handled [28]. We address dividend in all the results presented in the paper.

5. Twitter volume spikes and stock price model

While log return is widely assumed to follow normal distribution (see Section 4.2), this assumption does not always hold in practice [23]. Specifically, the distribution of log returns can possess much heavier tails than those of normal distribution. In other words, log

Please cite this article as: W. Wei et al., Twitter volume spikes and stock options pricing, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.06.018

Table 1
Percentage of samples that follow a normal distribution.

τ	15	30	50	100	150
Percentage	77.2%	66.2%	53.8%	31.2%	19.6%

returns can grow or drop much more sharply than that in normal distribution. Intuitively, sharp increases or decreases in stock returns can trigger more discussions about them, and hence Twitter volume spikes. Therefore, extreme stock prices might be correlated with Twitter volume spikes. In the following, we investigate whether this is indeed the case. After that, we investigate the characteristics of the stock price before and after a Twitter volume spike, and how to choose model parameters in the presence of Twitter volume spikes.

5.1. Twitter volume spikes and lognormal assumption

For a stock, consider a time series of log returns over 2τ days around day t, $\mathcal{R}_{t,\tau} = \{R_{t-\tau+1}, \ldots, R_t, \ldots, R_{t+\tau}\}$. In the following, we vary τ from 15 to 150, and identify when $\mathcal{R}_{t,\tau}$ is likely to follow a normal distribution. For this purpose, we consider the log returns of all the S&P 500 stocks from February 21, 2012 to August 1, 2014. For each stock, we random pick a time t and use Shapiro–Wilk test [27] to test whether $\mathcal{R}_{t,\tau}$ follows a normal distribution. Table 1 shows the percentage of the samples that follow a normal distribution for different values of τ. We can see that as τ increases, the percentage of samples that follow a normal distribution decreases. This indicates that the assumption of normal distribution is more likely to hold for short-term data and is less likely to hold for long-term data. In the rest of the paper, we choose $\tau \le 30$.

We next investigate whether extreme stock returns are related to Twitter volume spikes. For this purpose, we consider all the Twitter volume spikes (there are 3288 such samples). Suppose for a stock, a Twitter volume spike happens on day t, we then use Shapiro–Wilk test [27] to test whether $\mathcal{R}_{t,\tau}$ follows a normal distribution. Table 2 shows the percentage of samples that follow a normal distribution, where $\tau = 15$ or 30. For comparison, the results for a day that is chosen randomly are also shown in the table. For fair comparison, the samples of random chosen days are constructed as a one-to-one mapping with those of Twitter volume spikes. Specifically, if for a stock, there is a Twitter volume spike on day t, then we randomly choose a day, t', as a sample that corresponds to the sample for Twitter volume spike. From Table 2, we see that the log returns around a Twitter volume spike are much less likely to follow a normal distribution than those around a random day.

We next consider the time periods before and after a Twitter volume spike separately. Let $\mathcal{R}^-_{t,\tau}$ denote the series of log returns of τ days, from $t-\tau+1$ to t. We again use Shapiro–Wilk test to test whether $\mathcal{R}^-_{t,\tau}$ follows a normal distribution. Similarly, let $\mathcal{R}^+_{t,\tau}$ denote the set of log returns of τ days, from $t+1$ to $t+\tau$, we test whether it follows a normal distribution. Table 2 also shows, for each of the two sub-periods, the percentage of samples that follow a normal distribution. We observe that the log returns in the two sub-periods are more likely to follow normal distributions than those in the entire period. Furthermore, the log returns in the latter sub-period (i.e., after the Twitter volume spike, excluding the day with Twitter volume spike) are more likely to follow a normal distribution than those in the former sub-period. This implies that the log returns on the days around a Twitter volume spike, and especially on the day with Twitter volume spike, are more likely to be extreme values. To further confirm this, we remove 6 days, from $t-2$ to $t+3$, in each sample. Specifically, let $\mathcal{R}'^-_{t,\tau}$ denote the set of log returns from $t-\tau+1$ to $t-3$, and let $\mathcal{R}'^+_{t,\tau}$ denote the set of log returns from $t+4$ to $t+\tau$. The results are shown in Table 3. We observe that log returns are indeed more likely to follow a normal distribution after removing these 6 days. In fact, the results are comparable to those when choosing a random day, which further confirms that extreme log returns are correlated with Twitter volume spikes.

5.2. Twitter volume spikes and stock price model selection

We have observed that stock price exhibits different behaviors before and after a Twitter volume spike. Specifically, log returns are more likely to follow a normal distribution after a Twitter volume spike. In the following, we first compare the stock volatility in the time periods before and after a Twitter volume spike. The results will provide insights on whether different model parameters are needed for the two time periods. Based on our results in Section 5.1, all the results below are restricted to a short time period surrounding a Twitter volume spike. Specifically, suppose that a Twitter volume spike happens on day t. Then we only consider the days in $[t-\tau+1, t+\tau]$, where $\tau \le 30$. Let σ^-_τ denote the stock volatility derived from the log returns from day $t-\tau+1$ to t. Let σ^+_τ denote the stock volatility derived from the log returns from day $t+1$ to $t+\tau$. Both σ^-_τ and σ^+_τ are *empirical* volatility that are obtained using (5). We use paired t-test to compare σ^-_τ and σ^+_τ for all 3288 Twitter volume spikes. The null hypothesis is $\sigma^-_\tau \le \sigma^+_\tau$. Table 4 shows the p-values of the t-tests when varying τ from 15 to 30. The very small p-values indicate that we can reject the null hypothesis, indicating that there is strong evidence that $\sigma^-_\tau > \sigma^+_\tau$. For comparison, we also show the t-test results

Table 2
Percentage of samples that follow a normal distribution for the days around a Twitter volume spike. The results for randomly chosen days are also presented for comparison.

Testing set	$\tau = 15$		$\tau = 30$	
	Twitter volume spike	Random day	Twitter volume spike	Random day
$\mathcal{R}_{t,\tau}$	57.4%	76.6%	45.8%	59.4%
$\mathcal{R}^-_{t,\tau}$	69.7%	86.0%	60.4%	73.7%
$\mathcal{R}^+_{t,\tau}$	83.7%	86.3%	74.6%	76.6%

Table 3
Percentage of samples that follow a normal distribution after excluding days from $t-2$ to $t+3$. The results for randomly chosen days are also presented for comparison.

Testing set	$\tau = 15$		$\tau = 30$	
	Twitter volume spike	Random day	Twitter volume spike	Random day
$\mathcal{R}'^-_{t,\tau}$	87.9%	88.5%	77.5%	75.9%
$\mathcal{R}'^+_{t,\tau}$	88.2%	88.5%	78.1%	77.8%

Please cite this article as: W. Wei et al., Twitter volume spikes and stock options pricing, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.06.018

Table 4
p-values of the t-tests for $\sigma_\tau^- > \sigma_\tau^+$.

τ	Twitter volume spike	Random day
15	2.5×10^{-71}	0.6
20	7.1×10^{-66}	0.5
25	5.5×10^{-60}	0.5
30	2.1×10^{-46}	0.4

Table 5
p-values of the t-tests for likelihood improvement.

τ	$I_3 > I_3'$	$I_4 > I_4'$
15	8.6×10^{-13}	5.7×10^{-8}
20	2.5×10^{-12}	1.1×10^{-9}
25	5.1×10^{-10}	1.8×10^{-8}
30	3.9×10^{-8}	2.1×10^{-7}

when choosing a random day, which exhibit large p-values, indicating no strong evidence that $\sigma_\tau^- > \sigma_\tau^+$.

The above observation (i.e., $\sigma_\tau^- > \sigma_\tau^+$) indicates that we may need to use different parameters for the two time periods before and after the Twitter volume spike. In the following, we consider three models. The first model uses two parameters for drift and volatility respectively, denoted as $\mu_{2\tau}$ and $\sigma_{2\tau}$, that are estimated from the entire time period (i.e., 2τ days, indicated by the subscripts) using (5), respectively. The second model estimates three parameters, $\mu_{2\tau}$, σ_τ^- and σ_τ^+, where $\mu_{2\tau}$ is the drift estimated using the entire time period, and σ_τ^-, σ_τ^+ are the volatilities that are estimated using the first τ and last τ days, respectively. The third model uses four parameters, μ_τ^-, μ_τ^+, σ_τ^- and σ_τ^+, where μ_τ^- and σ_τ^- are estimated using the first τ days, and μ_τ^+ and σ_τ^+ are estimated using the last τ days.

To decide which model is the best, we use AICc [3], i.e., Akaike information criterion (AIC) with a correction for finite sample sizes, as a measure of the relative quality of each model. Specifically, for a given statistical model for m samples, AICc is defined as

$$AICc = AIC + \frac{2k(k+1)}{m-k-1},$$
$$AIC = 2k - 2\ln L,$$

where k is the number of parameters in the model and L is the maximized value of the likelihood function for the model. A smaller value of AICc indicates the model is preferred. As shown above, AIC is a measure that deals with the trade-off between the goodness of fit of the model and the complexity of the model; AICc enhances AIC by adding greater penalty for extra parameters.

For each Twitter volume spike, we calculate the AICc values for the three models described above, denoted as $AICc_2$, $AICc_3$ and $AICc_4$, respectively, where the subscript corresponds to the number of parameters in a model. The value of τ is chosen to be 15, 20, 25 and 30. After that, we use paired t-test to pairwise compare the AICc values for the three models. We find that, for all the settings that we consider, there is strong evidence that $AICc_2 > AICc_3$, $AICc_2 > AICc_4$ and $AICc_4 > AICc_3$. That is, $AICc_2 > AICc_4 > AICc_3$. This is consistent with the earlier results that the volatilities before and after a Twitter volume spike differ significantly, which justifies that they should be estimated separately. On the other hand, the result that the model with three parameters outperforms that with four parameters indicates that it is undesirable to use too many parameters.

Last, we investigate the gain obtained when using a proper model. Let L_2, L_3 and L_4 denote the maximized value of the likelihood function for the three model (the subscript represents the number of parameters in a model). Define the likelihood improvement when using three parameters over using two parameters as $I_3 = L_3/L_2 - 1$. Similarly, define $I_4 = L_4/L_2 - 1$ for the improvement using four parameters over using two parameters. For comparison, we also investigate a randomly chosen time period $[t-\tau+1, t+\tau]$, when t is chosen randomly, and denote the likelihood improvements as I_3' and I_4', respectively (in this case, our t-tests also indicate that $AICc_2 > AICc_4 > AICc_3$). We find that the gain when t is a random day is less significant than that when there is a Twitter volume spike on day t. Specifically, we perform t-tests to compare I_3 and I_3', and compare I_4 and I_4'. The null hypotheses are $I_3 \le I_3'$ and $I_4 \le I_4'$. Table 5 shows the p-values of the t-tests when varying τ from 15 to 30. The very small p-values indicate that we can reject the null hypothesis. That is, there is strong evidence that the likelihood improvement corresponding to the case of Twitter volume spike is larger than that of a random day.

In summary, the above results demonstrate that it is important to take Twitter volume spikes into account while studying and modeling stock prices. Specifically, the behavior of stock prices differs significantly before and after a Twitter volume spike: the empirical volatility is lower after a Twitter volume spike, and a three-parameter model that provides separate estimation of the volatilities before and after a Twitter volume spike provides the highest gain in the likelihood value.

6. Twitter volume spikes and stock options pricing

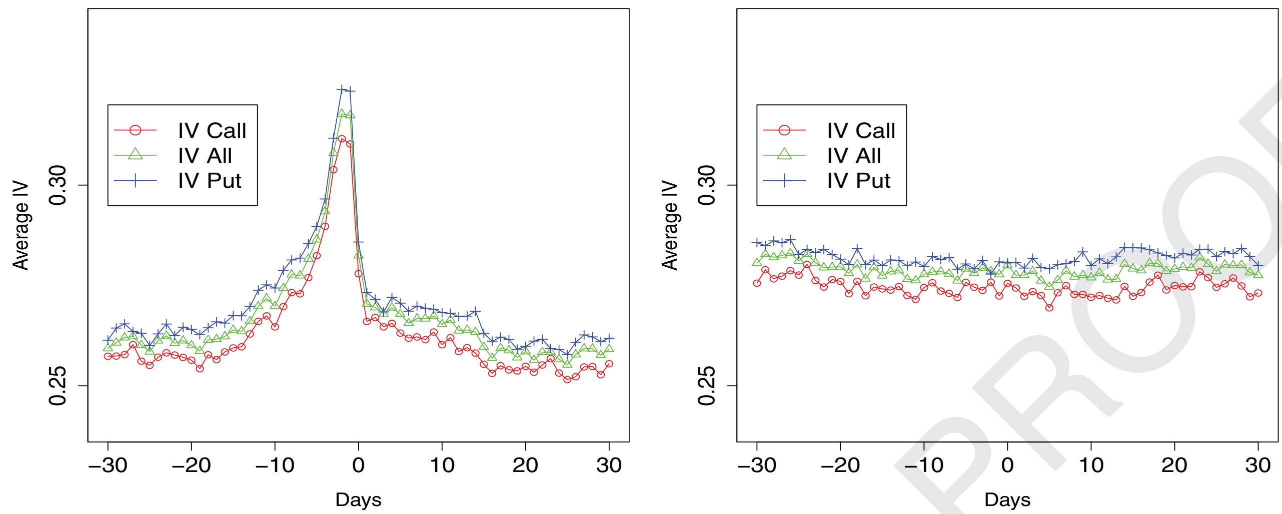
Having investigated the relationship of Twitter volume spikes and the lognormal stock price model, we now investigate the relationship of Twitter volume spikes and the Black–Scholes model for stock options pricing. Using the Black–Scholes model, one can derive implied volatility (IV) of an option contract, which is an estimate of the volatility. In the following, we first investigate IV around a Twitter volume spike, and then investigate volatility around a Twitter volume spike.

6.1. IV around a Twitter volume spike

We only consider short term options that will expire in around 30 days after a Twitter volume spike since such short-term options are more likely affected by a Twitter volume spike, while long term options are less affected by a Twitter volume spike. Consider a stock. On day t, for a given option price, a given strike price with an expiration date, and the current stock price, we can use (7) to solve for σ to obtain the IV corresponding to the call option; similarly, we can use (10) to obtain the IV corresponding to the put option. We obtain IV at the end of a trading day. The stock price is the daily closing price. The price of an option is taken as the average of the ask and bid prices to take account of ask-bid spread (ask price is the lowest price for which a seller is willing to sell, bid price is the highest price that a buyer is willing to pay for, and these two prices can be very different).

We next investigate the IV around a Twitter volume spike. For convenience, we represent time as relative to when a Twitter volume spike happens; negative values correspond to days before a Twitter volume spike, while positive values correspond to days after a Twitter volume spike. Suppose that one Twitter volume spike is for a particular stock, and happens on day t_0. We consider all the strikes (that will expire in around 30 days after t_0) for this stock on day t (relative to t_0), and obtain the IVs for the put and call options for each strike on day t. After doing the above for all the Twitter volume spikes, we can obtain the average IV for day t over all the Twitter volume spikes, denoted as $\overline{\sigma_t}$. Specifically, $\overline{\sigma_t}$ is a weighted sum of all the IVs (for each Twitter volume spike, we obtain a set of IVs, one IV for one option), where the weight for an IV is the trading volume of its corresponding option at the end of the trading day. We further obtain two more quantities that are similar to $\overline{\sigma_t}$, denoted as $\overline{\sigma_t^c}$ and $\overline{\sigma_t^p}$, which differ from $\overline{\sigma_t}$ in that $\overline{\sigma_t^c}$ is obtained by only considering call options, while $\overline{\sigma_t^p}$ is obtained by only considering put options.

Please cite this article as: W. Wei et al., Twitter volume spikes and stock options pricing, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.06.018



(a) Twitter Volume Spike (b) Random Day

Fig. 1. (a) The average IV for each of the 30 days before and after a Twitter volume spike. Three cases, when only consider call options, only consider put options, and consider all options, are plotted in the figure. (b) The corresponding results for randomly chosen days.

We next investigate how $\overline{\sigma_t}$, $\overline{\sigma_t^c}$, and $\overline{\sigma_t^p}$ change with t. Fig. 1(a) plots these three quantities for $t \in [-30, 30]$. In the figure, for each of these three quantities, the value for day t is an average value that is obtained considering all the instances of Twitter volume spikes, excluding those for which we cannot obtain one of the three quantities (e.g., there may not exist a call or put option with short-term expiration date). We observe that all the three quantities, $\overline{\sigma_t}$, $\overline{\sigma_t^c}$, and $\overline{\sigma_t^p}$, increase sharply before a Twitter volume spike and decrease quickly afterwards. The IV is still high right after a Twitter volume spike, and then decreases to a low value. For comparison, we also investigate how IV changes before and after a day that is chosen randomly. The results are shown in Fig. 1(b). Different from the results in Fig. 1(a), the results in Fig. 1(b) do not indicate a significant increase in IV before a random day or a significant decrease in IV after a random day.

We also observe from Fig. 1 (a) that $\overline{\sigma_t^p}$ is larger than $\overline{\sigma_t^c}$ for all of the 61 days, which indicates that put options may be priced higher compared to call options. We next use t-test to further confirm the above results. The null hypothesis is $\overline{\sigma_t^p} \le \overline{\sigma_t^c}$ for $t \in [-30, 30]$. For all of the 61 days, 55 days have p-value less than 0.05. The very small p-values on most days indicate that we can reject the null hypothesis, that is, there is strong evidence that $\overline{\sigma_t^p} > \overline{\sigma_t^c}$, further confirming the results we observe from Fig. 1(a). For a randomly chosen day, Fig. 1 (b) also shows that $\overline{\sigma_t^p} > \overline{\sigma_t^c}$, which is also confirmed by t-test.

We next further explore the relationship between the IV obtained from put options and the IV obtained from call options. For each Twitter volume spike, for day t, we compare the average IV obtained from put options (again, the average is a weighted sum, where the weight for an IV is the trading volume of its corresponding option at the end of the trading day) and that obtained from call options. We then obtain the percentage that the former is larger than the latter considering all the instances of Twitter volume spikes for day t. The results are presented in Fig. 2(a), $t \in [-30, 30]$. We see that the percentage is above 60% for all 61 days. For the 2 days immediately before a Twitter volume spike, the percentages are particularly high, and then the percentage drops quickly afterwards. For comparison, Fig. 2(a) also plots the corresponding results for randomly chosen days, which shows that the percentages are also above 60%. On the other hand, we observe a more significant increase and a more significant decrease in percentage right before and after a Twitter volume spike, compared to the case of random days. The above results again confirm that IV of put options is larger than that of call options. To further illustrate the above points, we obtain the ratio of the average IV obtained from put options over the average IV obtained from call options for each instance of Twitter volume spike on day t, and then obtain the average ratio over all the instances for day t. Fig. 2(b) plots the average ratio for each of the 30 days before and after a Twitter volume spike. The 95% confidence intervals are also plotted in the figure. We can see that ratios are above 1.03 for most days, providing further evidence that IV of put options is larger than that of call options.

In summary, we observe that the IV is still high right after a Twitter volume spike. A natural question is whether it accurately predicts the actual volatility. In addition, we observe that put options are priced higher than call options. A natural question is whether it is rational, or it is due to people's tendency of loss aversion (i.e., people tend to strongly prefer avoiding losses to acquiring gains) [14,15]. We next answer these two questions by investigating volatility around a Twitter volume spike.

6.2. Volatility around a Twitter volume spike

When investigating volatility around a Twitter volume spike, to gain insights, we make a simplifying assumption that the price of a stock follows a Brownian motion (instead of Geometric Brownian motion). That is, we ignore the deterministic term in the right hand side of (2). This is reasonable since we are only interested in short-term (i.e., within 60 days) behavior. Under this assumption, we have log return on day t as

$$R_t = \ln \frac{S_t}{S_{t-1}} = \sigma (W_t - W_{t-1}),$$

where W_t is a Brownian motion. From the above, we see that, under the simplifying assumption, R_t/σ follows a standard normal distribution.

We next explore R_t/σ for t around a Twitter volume spike, where t is relative to the day when a Twitter volume happens, $t \in [-29, 30]$. Since we do not know the real σ, we use the IV on day -30 to approximate σ. Specifically, for a stock, the IV on a day is a weighted average considering all the strikes (both call and put options) for the stock (again we only consider options that will expire in around 30 days), where the weight is the trading volume of an option at the end of the trading day. We only consider Twitter volume spikes for which we can obtain the IV on day -30. For each such Twitter volume spike, we can obtain one instance of R_t/σ for day $t \in [-29, 30]$.

Please cite this article as: W. Wei et al., Twitter volume spikes and stock options pricing, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.06.018

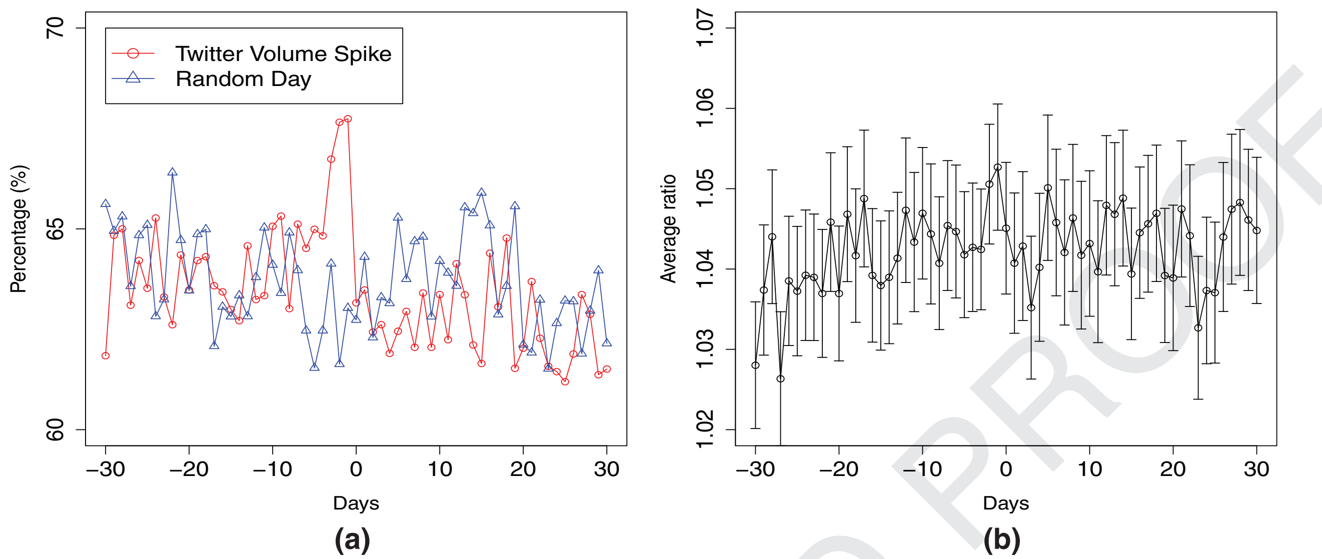


Fig. 2. (a) Percentage that IV obtained from put options is larger than that from call options. (b) Average ratio of IV obtained from put options over IV obtained from call options (with 95% confidence interval).

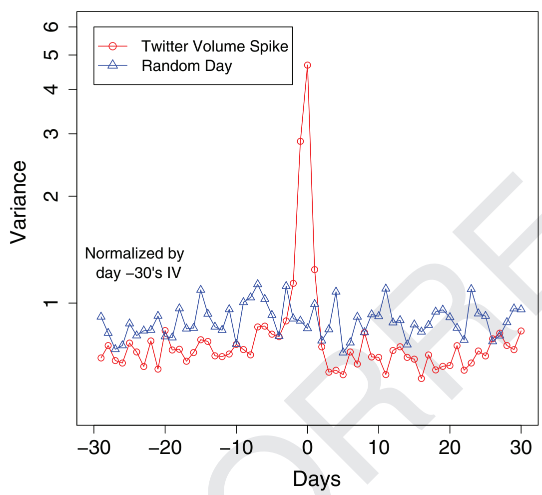


Fig. 3. Variance of normalized log returns around a Twitter volume spike.

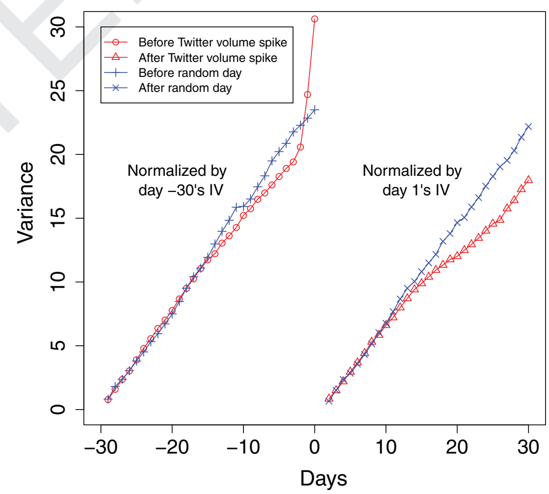


Fig. 4. Variance of normalized cumulative log returns around a Twitter volume spike. For comparison, the corresponding results for randomly chosen days are also plotted in the figure.

We then use the sample variance to approximate the variance of R_t/σ for $t \in [-29, 30]$. Fig. 3 plots the results. For comparison, Fig. 3 also plots the corresponding results for randomly chosen days. We see that for the case of Twitter volume spikes, the variances from day -1 to day 1 are much larger than the corresponding values for the case of randomly chosen days. The difference is most significant on day 0 (the former is 6 times of the latter). On the other hand, for most of the days after 0, i.e., 28 out of 30 days, the variances of the former are lower than those in the latter. The results indicate that the price of a stock is very volatile around a Twitter volume spike (related to this stock), particularly for the days immediately before and after the Twitter volume spike (i.e., for days -1 to $+1$). After that, the volatility is even lower than usual.

The above considers the variance of log returns. We next consider the variance of *cumulative log return*. Consider a stock. Define $R_{t+n,t}$ as the cumulative log return on day $t+n$ relative to day t, $n \geq 1$. Then under the simplifying assumption that stock price follows a Brownian motion, we have

$$R_{t+n,t} = \ln \frac{S_{t+n}}{S_t} = \sigma (W_{t+n} - W_t).$$

Therefore, $R_{t+n,t}/\sigma$ follows a normal distribution with variance n.

We now investigate $R_{t+n,t}/\sigma$ around a Twitter volume spike. Again, t is relative to the day when a Twitter volume happens. We consider $t \in [-30, 30]$. Based on the earlier observation that the variance of log return on a day with a Twitter volume spike is significantly larger than the variances of other days, we divide the time period into two parts, one from day -30 to 0 and the other from day 1 to 30. For the first part, the cumulative log return on day i is $\ln \frac{S_i}{S_{-30}}$, $i \in [-29, 0]$, where S_i is the stock price on day i, and we normalize it by the average IV on day -30. For the second part, the cumulative log return on day i is $\ln \frac{S_i}{S_1}$, $i \in [2, 30]$ and we normalize it by the average IV on day 1.

Fig. 4 plots the results for $t \in [-30, 30]$, where the value for t is the average over all the instance of Twitter volume spikes. Specifically, for each t, we have 2955 samples (excluding 333 Twitter volume spikes for which we cannot obtain the IV on day -30 or day 1). For both parts, we use the sample variance of the normalized cumulative log returns to approximate the variance. If the Brownian motion assumption holds, the variance of the normalized cumulative log return will increase linearly with time. For comparison, the corresponding

Please cite this article as: W. Wei et al., Twitter volume spikes and stock options pricing, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.06.018

results for randomly chosen days are also plotted in the figure. For the case of randomly chosen days, for both parts (i.e., days [−30, 0] and [1, 30]), the variance of cumulative log returns indeed increases approximately linearly with time. For the case of Twitter volume spikes, for both parts, the variance increases linearly with time except for days −2, −1 and 0, which have particularly large variance. We use least squares estimation to estimate the slopes of all the linear curves (for the case of Twitter volume spike, days −2, −1 and 0 are omitted in the estimation). For the case of Twitter volume spikes, the slopes of the two parts are 0.73 and 0.58, respectively, while for the case of random chosen days, the slopes of the two parts are 0.82 and 0.76, respectively. The significantly lower slope of the second part when there are Twitter volume spikes compared to that for randomly chosen days (i.e., 0.58 versus 0.76) indicates that the IV of day 1 (i.e., the day immediately a Twitter volume spike, which is used to normalize $\ln \frac{S_i}{S_1}$, $i \in [2, 30]$) may still be higher than usual, and hence option prices on that day may still be overpriced. This indicates that we can use Twitter volume spike as a trading signal: right after a Twitter volume spike, we can utilize the overpriced options to gain profit, which will be descried in detail in Section 7.

7. Application in stock option trading

Our earlier analysis indicates that put options tends to be priced higher than call options, and option prices may still be overpriced right after a Twitter volume spike. Based on the above results, we conjecture that selling put options right after a Twitter volume spike can be a profitable trading strategy. In the following, we first describe one such strategy and then evaluate its performance.

7.1. Put spread selling strategy

Before describing the strategy, we first describe put option selling in more detail. As described earlier, put option is a financial contract between a buyer and seller of the option. It gives the buyer the right to sell a stock at the strike price on the option expiration day. As an example, suppose that a seller sells a put, which gives a buyer the right to sell 100 shares of the stock of a company, say XYZ, at the strike price of $80 at expiration (i.e., on the expiration day). To purchase the option, the buyer pays the premium of $2 per share (premium is paid to the seller of the option and is quoted on a per-share basis). If the stock price is $82 at expiration, which is higher than the strike price, then the seller can keep the premium, gaining a profit of $2 \times 100 = \$200$. On the other hand, if the stock price drops to $70 at expiration, then the profit of the buyer is $(80 - 70 - 2) \times 100 = \$800$, while the seller loses $800. In other words, for a buyer, one of the purposes of buying put option is similar to buying an insurance: it limits the loss of the buyer during unfavorable events with the payment of the premium. For a seller, selling put option can lead to profits through the premium. On the other hand, when the stock price drops significantly, then a seller can lose a substantial amount of money. For instance, in the previous example, if the stock price falls to zero (XYZ bankrupts), then the loss of the seller will be $(80 - 2) \times 100 = \$7800$.

Options spread is widely considered as an option trading strategy to limit the risk. In this paper, we consider one type of option spread strategy, called put spread selling. Specifically, the put spread strategy is *bull spread* [23]. It is established with put options by buying a put with a lower strike price and simultaneously selling a put with a higher strike price; the two puts have the same expiration date. This strategy limits the amount of loss. For instance, in the previous example, suppose that a trader buys a put option with the strike price of $75 at the premium of $1 per share and sells a put with the strike price of $80 at the premium of $2 per share. Then even if the stock price falls to zero, the loss of the trader is limited to $(80 - 2 - 75 + 1) \times 100 = \$400$. Fig. 5 illustrates the maximum profit and loss (in a negative value) using the above strategy. When the stock price at expiration is no less than $80, the trader earns a profit of $(2 - 1) \times 100 = \$100$; when the stock at expiration is no more than $75, the trader has a loss of $400; and when the stock price at expiration is between $75 and $80, the profit of the trader is between −$400 and $100, and is a linear function of the stock price at expiration.

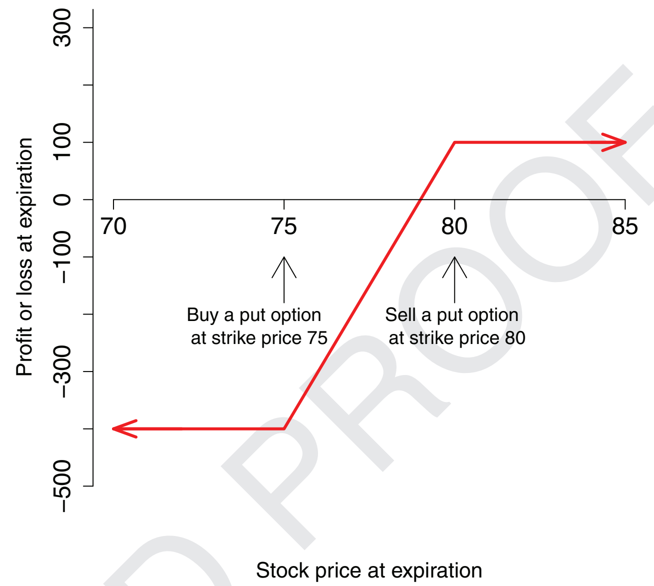


Fig. 5. An example illustrating put spread strategy. In the example, the strategy is established by buying a put with the strike price of $75 at the premium of $1 per share and selling a put with the strike price of $80 at the premium of $2 per share.

Based on our observations in earlier sections, we propose the following put spread selling strategy. Suppose that for a stock, a Twitter volume spike happens on day t. Then a trader uses a put spread strategy on a day right after t. Specifically, he will choose a put spread that will expire in a few weeks after t, and buy and sell puts with δ value in different ranges (δ represents the rate of change of option value with respect to changes in the stock price [28]. For put options, δ values are negative, from −1 to 0. As a rough example, a put option with a δ of −0.5 will decrease by $0.50 for every $1 increase in the underlying stock price).

7.2. Performance evaluation

We next evaluate the performance of the above strategy. We first consider a simplified simulation scenario, and then consider realistic simulation settings.

7.2.1. Simplified trading simulation

In the simplified scenario, we do not consider commission. In addition, the price of an option is set to be the average of the ask and bid prices. The performance metric we use are *premium retention ratio* and *fraction of winning trades*. The premium retention ratio is the amount of profit divided by the amount of premium collected for all traded options. For instance, in the earlier example on bull spread, when the stock price is $82 at expiration, the premium retention ratio is 1; while when the stock price is $75 at expiration, the premium retention ratio is $-400/(200 - 100) = -4$. The fraction of winning trades is defined as the ratio of trades that have positive profit.

Table 6 shows the results, where the trade is on t, $t + 1$ or $t + 2$ (a Twitter volume spike happens on day t) and the expiration date is 4 weeks after t. For simplicity, we only sell put options with δ between −0.5 and 0. For the put options that we buy, it is only sensible to choose options with strike prices lower than those of the put options that we sell (see the example in Fig. 5). In addition, to control the risk

Please cite this article as: W. Wei et al., Twitter volume spikes and stock options pricing, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.06.018

Table 6
Performance of the put spread selling strategy in simplified trading simulation.

Q3

δ range		Premium retention ratio			Fraction of winning trades		
Sell	Buy	t	$t+1$	$t+2$	t	$t+1$	$t+2$
$[-0.5, -0.4]$	$[-0.3, -0.2]$	32.6%	31.5%	**33.6%**	74.2%	74.3%	**74.6%**
$[-0.4, -0.3]$	$[-0.2, -0.1]$	44.9%	**50.9%**	47.0%	82.5%	**85.6%**	84.4%
$[-0.3, -0.2]$	$[-0.1, 0]$	48.0%	**59.8%**	54.3%	88.9%	**91.4%**	91.1%

level, we let the δ level of the options that we buy to be roughly 0.2 lower than that of the options that we sell. Summarizing the above, we have the three settings of δ listed in Table 6. That is, δ for the options that we sell is chosen to be in $[-0.5, -0.4]$, $[-0.4, -0.3]$, or $[-0.3, -0.2]$. Correspondingly, the δ for the options that we buy is chosen to be in $[-0.3, -0.2]$, $[-0.2, -0.1]$, or $[-0.1, 0]$, respectively. For the options that we sell or buy, if the δ is chosen to be in range $[a, b]$ and there are multiple candidate options in the range, we choose the option whose δ value is closest to b. From Table 6, we see that the strategy gains profit in all the settings. Specifically, the average premium retention ratio varies from 31.5% to 59.8%, and the fraction of winning trades varies from 74.2% to 91.4%.

7.2.2. *Realistic trading simulation*

We next evaluate the performance of the put spread strategy through realistic trading simulation. In the simulation, we use a portfolio that can have up to 20 spread positions. Initially, the cash balance is \$100,000, the number of open positions is 0, and the number of available positions is 20. After we apply put spread strategy for a stock (i.e., sell a put at a high strike price and buy a put at a low strike price), the number of available positions is reduced by one until these options are settled on their expiration day. During the simulation, we try to keep the amount of cash that is allocated to a position to be balanced. Specifically, if c is the current cash balance and n is the number of available positions, then the maximum amount of cash to a position is c/n. For instance, at the beginning, the amount of cash that can be allocated to a position is 100,000/20. Suppose at a later time, there are already two open positions and the amount of cash is 90,000. Then the number of available positions becomes 18, and the maximum amount of cash to a position is 90,000/18. For one position, the number of put spread is $\lfloor c/(nb) \rfloor$, where b is the margin requirement of the put spread (i.e., 100 times the difference of the two strike prices, e.g., in the example in Section 7.1, the margin requirement is $(80-75) \times 100 = \$500$). For each put spread, we assume the commission is \$2 (\$1 for selling and \$1 for buying a put). In addition, to be realistic, we take ask-bid spread into account, that is, we buy an option at the ask price and sell an option at the bid price. At any point of time, the number of open positions is no more than 20.

The performance metric is *percentage gain*, that is, the relative difference of the cash balance from the beginning to the end of the

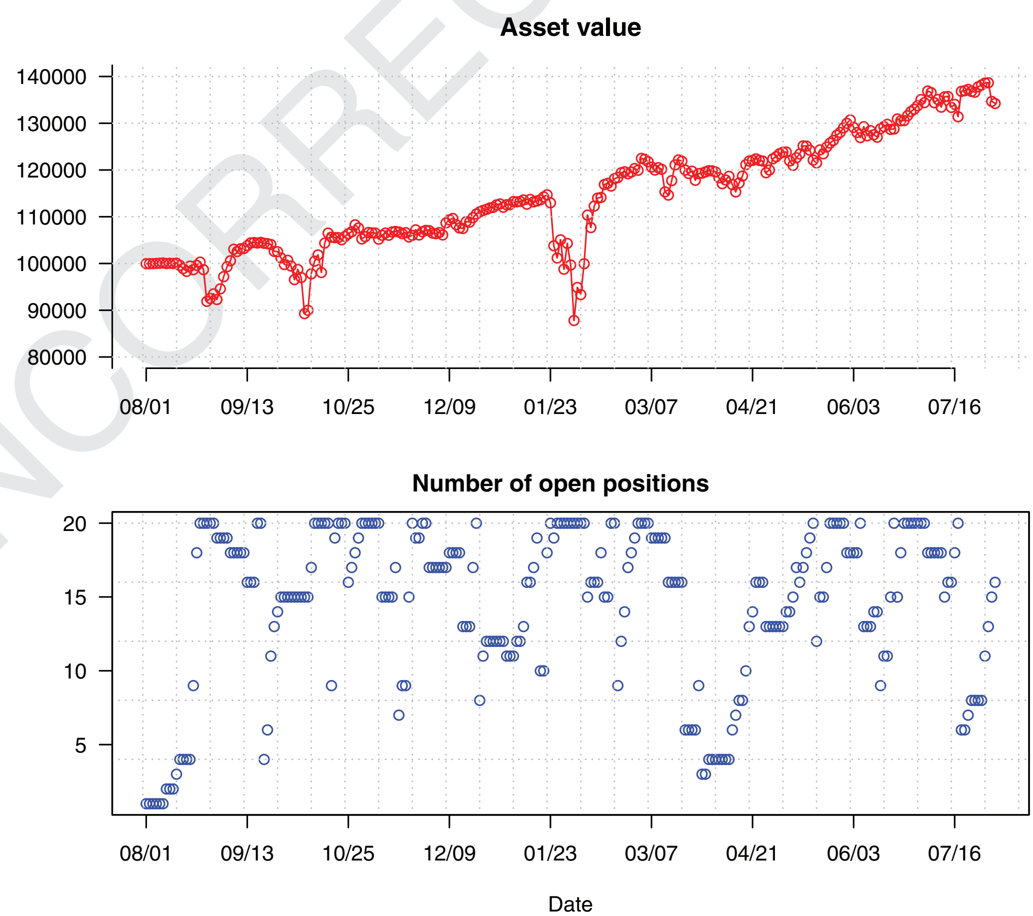


Fig. 6. Put spread simulation. The setting is: sell options with $\delta \in [-0.3, -0.2]$ and buy options with $\delta \in [-0.1, 0]$. The upper figure shows the value of the asset (available cash plus value of the options) on each day; the lower figure shows the number of open positions in the portfolio.

Please cite this article as: W. Wei et al., Twitter volume spikes and stock options pricing, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.06.018

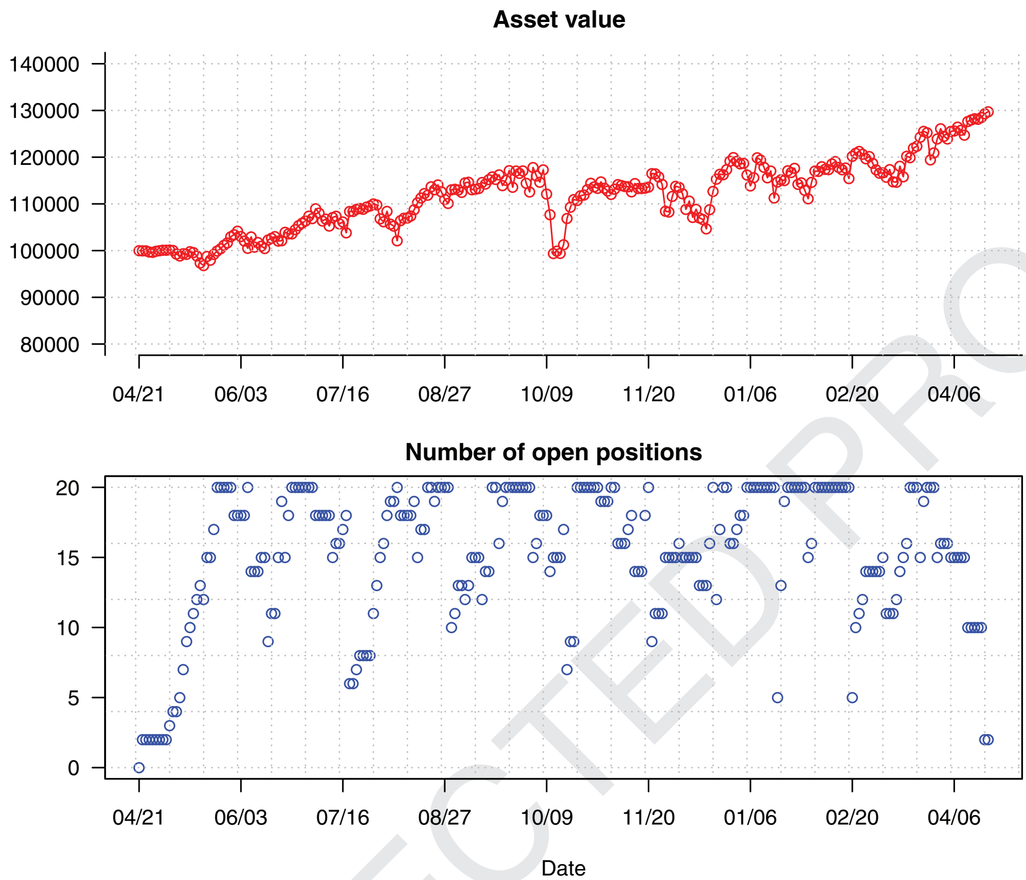


Fig. 7. Put spread simulation (using data collected from April 21, 2014 to April 20, 2015). The setting is: sell options with $\delta \in [-0.3, -0.2]$ and buy options with $\delta \in [-0.1, 0]$. The upper figure shows the value of the asset (available cash plus value of the options) on each day; the lower figure shows the number of open positions in the portfolio.

Table 7
Performance of the put spread selling strategy in realistic trading simulation.

δ range		Trading day		
Sell	Buy	t	$t+1$	$t+2$
$[-0.5, -0.4]$	$[-0.3, -0.2]$	−18.3%	−32.5%	**10.6%**
$[-0.4, -0.3]$	$[-0.2, -0.1]$	19.9%	**50.6%**	37.4%
$[-0.3, -0.2]$	$[-0.1, 0]$	−0.8%	**34.3%**	33.8%

Table 8
Performance of the put spread selling strategy in realistic trading simulation (using data collected from April 21, 2014 to April 20, 2015).

δ range		Trading day		
Sell	Buy	t	$t+1$	$t+2$
$[-0.5, -0.4]$	$[-0.3, -0.2]$	−17.6%	**15.9%**	11.5%
$[-0.4, -0.3]$	$[-0.2, -0.1]$	24.0%	**36.7%**	13.6%
$[-0.3, -0.2]$	$[-0.1, 0]$	10.8%	**29.7%**	27.3%

simulation. Table 7 shows the simulation results. This strategy achieves 50.6% gain when selling options with $\delta \in [-0.4, -0.3]$ and buying options with $\delta \in [-0.2, -0.1]$ on the day following a Twitter volume spike. Although this setting achieves high rate of return, the stock volatilities for this setting are also relatively large, indicating that the trading risk for this setting is large. When selling options with $\delta \in [-0.3, -0.2]$ and buying options with $\delta \in [-0.1, 0]$, which is a lower risk setting, the strategy still achieves 34.3% gain. Fig. 6 plots the simulation result for this setting. The upper figure shows the value of the asset (available cash plus value of the options) on each day, and the lower figure shows the number of open positions on each day. We observe that both quantities change in a stable fashion. Last, of all 180 trades, only 17 trades lose money. The fraction of winning trades is 90.6%.

The above results are for the data collected from August 1, 2013 to August 6, 2014. We also repeat the above evaluation for the data collected from April 21, 2014 to April 20, 2015, and observe similar results. Table 8 shows the simulation results, which are consistent with the results in Table 7. Fig. 7 plots the value of the asset (available cash plus value of the options) and the number of occupied trading positions on each day. We observe that both of them change in a stable fashion. Last, of all 190 trades in this setting (selling options with $\delta \in [-0.3, -0.2]$ and buying options with $\delta \in [-0.1, 0]$), only 23 trades lose money. The fraction of winning trade is 87.9%.

8. Choice of threshold

So far, we have used threshold $K = 3$ when identifying Twitter volume spikes (see Section 3.3). In this section, we investigate how to choose K. The approach we use is based on the insights on how average IV changes around a Twitter volume spike. Let $\mathcal{D}$ denote the set of Twitter volume spikes that are identified using $K = 3$. Let $\mathcal{D}'$ denote the set of Twitter volume spikes that are identified using $K' \neq K$. It is clear that $\mathcal{D} \subseteq \mathcal{D}'$ when $K' < K$. When $K = 3$, let $\overline{\sigma_t}$ denote the average IV over all the instances of Twitter volume spikes as calculated in Section 6.1, where t is relative to the day when a Twitter volume spike happens, $t \in [-30, 30]$. For $K' < K$, let $\overline{\sigma_t'}$ denote the average IV over all the instances of Twitter volume spikes in $\mathcal{D}'$, and let $\overline{\sigma_t''}$ denote the average IV over all the instances of Twitter volume spikes in $\mathcal{D}' \setminus \mathcal{D}$,

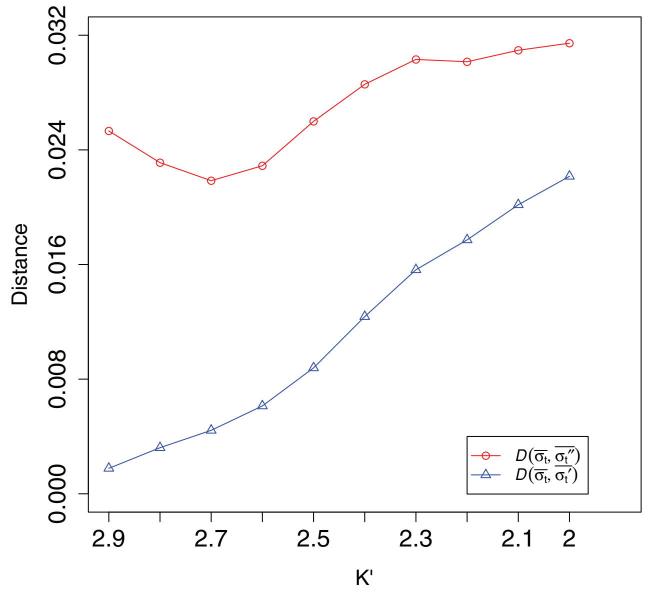


Fig. 8. The distance between $\overline{\sigma_t}$ and $\overline{\sigma_t'}$ (the lower curve with triangles) and the distance between $\overline{\sigma_t}$ and $\overline{\sigma_t''}$ (the upper curve with circles) when K' decreases from 2.9 to 2.

that is, $\overline{\sigma_t''}$ is the average IV from the additional Twitter volume spikes when choosing a smaller K'.

Define the distance between $\overline{\sigma_t}$ and $\overline{\sigma_t'}$ as the normalized Euclidean distance. Similarly, define the distance between $\overline{\sigma_t}$ and $\overline{\sigma_t''}$. That is,

$$D\left(\overline{\sigma_t}, \overline{\sigma_t'}\right) = \sqrt{\frac{\sum_{t=-30}^{30}\left(\overline{\sigma_t} - \overline{\sigma_t'}\right)^2}{61}},$$

$$D\left(\overline{\sigma_t}, \overline{\sigma_t''}\right) = \sqrt{\frac{\sum_{t=-30}^{30}\left(\overline{\sigma_t} - \overline{\sigma_t''}\right)^2}{61}}$$

Fig. 8 plots the distances defined above when K' decreases from 2.9 to 2. As expected, $D\left(\overline{\sigma_t}, \overline{\sigma_t'}\right)$ increases when K' decreases (i.e., deviates more from 3). The slope of the increase is lower at the beginning and becomes larger afterwards. The distance $D\left(\overline{\sigma_t}, \overline{\sigma_t''}\right)$ is the minimum when $K' = 2.7$. The larger distance when K' is larger than 2.7 is due to a small number of samples in $\mathcal{D}' \setminus \mathcal{D}$. When K' is smaller than 2.7, more Twitter volume spikes are identified; on the other hand, $\overline{\sigma_t''}$ deviate more from $\overline{\sigma_t}$, leading to larger distances. The above results indicate that the threshold can be chosen from 2.7 to 3, which may achieve similar performance as that when choosing the threshold to 3.

To further confirm this, we use $K = 2.7$ and the same thresholds for the number of unique users and user diversity index to identify Twitter volume spikes. In this case, we identify 4088 Twitter volume spikes (24.3% higher than that when using $K = 3$). We then repeat the analysis presented in Section 5–7 using the new set of Twitter volume spikes. Indeed, we find that the observations on stock price, IV and volatility are similar as those when $K = 3$, and the performance of the put spread trading strategy is similar as that when $K = 3$.

9. Conclusion and future work

In this paper, we have investigated the relationship between Twitter volume spikes and stock options pricing. We started with the underlying assumption of the Black–Scholes model, and investigated when this assumption holds for stocks that have Twitter volume spikes. We next investigated stock volatility around a Twitter volume spike and found that a three-parameter model that uses the same drift and different volatilities before and after a Twitter volume spike provides the highest gain in the likelihood value. We also found a clear pattern in IV around a Twitter volume spike: IV increases sharply before a Twitter volume spike and decreases quickly afterwards. In addition, put options tend to be priced higher than call options. Last, we found that right after a Twitter volume spike, options may still be overpriced. Based on the above findings, we propose a put spread selling strategy. Realistic simulation using one year stock market data demonstrates that, even in a conservative setting, this strategy achieves a 34.3% gain when taking account of commissions and ask-bid spread, while S&P 500 increases 12.8% in the same period.

As future work, we are looking into the content of tweets to understand their impact on stock options pricing. The results will be shown at http://www.finstats.com.

References

[1] Twitter reports second quarter 2014 results, Twitter, 2014-07-29.
[2] H. Achrekar, A. Gandhe, R. Lazarus, S.H. Yu, B. Liu, Twitter improves seasonal influenza prediction, in: Proceedings of Annual International Conference on Health Informatics (HEALTHINF), Vilamoura, Algarve, Portugal, 2012.
[3] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723.
[4] S. Asur, B. Huberman, Predicting the future with social media, in: Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, Canada, 2010.
[5] R. Bandari, S. Asur, B.A. Huberman, The pulse of news in social media: forecasting popularity, CoRR abs/1202.0332 (2012).
[6] R. Bar-Haim, E. Dinur, R. Feldman, M. Fresko, G. Goldstein, Identifying and following expert investors in stock microblogs, in: Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2011.
[7] F. Black, M. Scholes, The pricing of options and corporate liabilities., J. Pol. Econ. 81 (3) (1973) 637–654.
[8] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, J. Comput. Sci. 2 (1) (2011) 1–8.
[9] J. Du, H. Xu, X. Huang, Box office prediction based on microblog, Expert Syst. Appl. 41 (4, Part 2) (2014) 1680–1689.
[10] L. Hong, O. Dan, B.D. Davison, Predicting popular messages in twitter, in: Proceedings of International Conference Companion on World Wide Web (WWW), 2011.
[11] B.A. Huberman, D.M. Romero, F. Wu, Social networks that matter: Twitter under the microscope, First Monday 14 (1) (2009).
[12] J.C. Hull, Options, Futures and Other Derivatives, 8th ed., Prentice-Hall, 2011.
[13] A. Java, X. Song, T. Finin, B. Tseng, Why we Twitter: understanding microblogging usage and communities, in: Proceedings of WebKDD and SNA-KDD Workshop on Web Mining and Social Network Analysis, San Jose, CA, 2007.
[14] D. Kahneman, Thinking, Fast and Slow, reprint ed., Farrar, Straus and Giroux, 2013.
[15] D. Kahneman, A. Tversky, Prospect theory: an analysis of decision under risk, Econometrica 47 (1979) 263–291.
[16] N. Kanungsukkasem, P. Netisopakul, T. Leelanupab, Recognition of nasdaq stock symbols in tweets, in: Proceedings of International Conference on Knowledge and Smart Technology (KST), 2014.
[17] B. Krishnamurthy, P. Gill, M. Arlitt, A few chirps about Twitter, in: Proceedings of Workshop on Online Social Networks (WOSN), Seattle, WA, 2008.
[18] M. Makrehchi, S. Shah, W. Liao, Stock prediction using event-based sentiment analysis, in: Proceedings of IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013.
[19] H. Mao, S. Counts, J. Bollen, Predicting financial markets: comparing survey, news, Twitter and search engine data, arXiv:1112.1051 (2011).
[20] Y. Mao, B. Wang, W. Wei, B. Liu, Correlating S&P 500 stocks with Twitter data, in: Proceedings of ACM International Workshop on Hot Topics on Interdisciplinary Social Networks Research (HotSocial), Beijing, China, 2012.
[21] Y. Mao, W. Wei, B. Wang, Twitter volume spikes: analysis and application in stock trading, in: Proceedings of Workshop on Social Network Mining and Analysis (SNAKDD), Chicago, Illinois, 2013.
[22] R. McDonald, D. Siegel, The value of waiting to invest, Q. J. Econ. 101 (4) (1986) 707–727.
[23] L.G. McMillan, Options as a Strategic Investment, 5th, Prentice-Hall, 2012.
[24] R.C. Merton, Optimum consumption and portfolio rules in a continuous-time model, J. Econ. Theory 3 (4) (1971) 373–413.
[25] E.J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, A. Jaimes, Correlating financial time series with micro-blogging activity, in: Proceedings of ACM International Conference on Web Search and Data Mining (WSDM), Seattle, WA, 2012.
[26] T. Sakaki, M. Okazaki, Y. Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, in: Proceedings of International Conference on World Wide Web (WWW), Raleigh, NC, 2010.
[27] S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrika 52 (3–4) (1965) 591–611.
[28] P. Wilmott, Paul Wilmott on Quantitative Finance, John Wiley & Sons, 2013.
[29] X. Zhang, H. Fuehres, P.A. Gloor, Predicting stock market indicators through Twitter "I hope it is not as bad as I fear", Procedia – Soc. Behav. Sci. 26 (2011) 55–62.

Please cite this article as: W. Wei et al., Twitter volume spikes and stock options pricing, Computer Communications (2015), http://dx.doi.org/10.1016/j.comcom.2015.06.018