Predicting Cryptocurrency Price Bubbles Using Social Media Data and Epidemic Modelling

Ross C. Phillips and Denise Gorse

Department of Computer Science
University College London
Gower Street, London, WC1E 7JE, UK

{r.phillips, d.gorse}@cs.ucl.ac.uk

***Abstract* – Financial price bubbles have previously been linked with the epidemic-like spread of an investment idea; such bubbles are commonly seen in cryptocurrency prices. This paper aims to predict such bubbles for a number of cryptocurrencies using a hidden Markov model previously utilised to detect influenza epidemic outbreaks, based in this case on the behaviour of novel online social media indicators. To validate the methodology further, a trading strategy is built and tested on historical data. The resulting trading strategy outperforms a buy and hold strategy. The work demonstrates both the broader utility of epidemic-detecting hidden Markov models in the identification of bubble-like behaviour in time series, and that social media can provide valuable predictive information pertaining to cryptocurrency price movements.**

Index Terms – cryptocurrency price bubbles, social media data mining, hidden Markov model, trading strategy, epidemic detection.

I. Introduction

Cryptocurrencies, of which the best known is Bitcoin [1], have emerged as a new asset class: the total market capital of the cryptocurrency market hit $10 billion in 2013, $20 billion at the start of 2017, and over $50 billion by May 2017. Though a fast growing market, in March 2017 the SEC disapproved a proposal for a Bitcoin ETF (investment vehicle), citing price volatility and the price being driven by speculation as two concerns. These undesirable features could be the result of bubbles that have been observed in cryptocurrency prices [2] [3], which makes the early detection of such bubbles an important topic of research.

Though the term 'bubble' is used in many areas (such as real estate, stocks, and commodities) it does not have a widely accepted definition. A price bubble may be separated into five phases, as described by [4] and [5]. First, a *displacement* occurs—possibly a new technology or innovation. This leads to an initial *boom phase* characterized by increasing investment in the area. The asset price rises; slowly at first, but then increasingly quickly, increasing the excitement about the asset, which leads then to a *euphoria* phase in which there is frenzied trading in the now overvalued asset, and where prices increase more quickly. Over time more sophisticated investors start to reduce their positions (*profit taking phase*), but less sophisticated investors continue to buy. At some point price rises stall and a *panic phase* may follow, where a downward movement in prices causes worried investors to reduce their positions rapidly.

Shiller [6] comments that the burst (panic) phase, though fitting with the metaphor of a bubble, is not essential to the formation of a bubble, and notes history shows this burst phase does not always occur, or if it does occur can be followed by a continued boom. Shiller in fact favors a more epidemic-like definition, describing a bubble as occurring by psychological contagion, where the news of price increases spurs investors' enthusiasm which spreads contagiously and brings in a larger group of investors, drawn in by envy and excitement about the previous price rises [7].

It is our hypothesis that it is possible to examine patterns in social media usage to detect the earlier stages of a cryptocurrency price bubble, the boom phase referenced above and the 'increasing interest' described by Shiller. This allows early detection of the formation of a bubble. The link between cryptocurrency prices and social media usage has already been demonstrated in the literature (discussed further in Section II) and suggests that social media should provide an intuitive data source.

The remainder of this paper is structured as follows. Section II reviews the background and relevant literature on bubbles, epidemic detection, and financial prediction via social media mining. Section III details a hidden Markov model (HMM) previously used to track influenza epidemics, which will be used here to track cryptocurrency bubbles. Section IV explores the input factors to be used. Section V details the considerations made while designing the experiments. Section VI presents the results, and Section VII concludes with a brief discussion.

II. Background

A. Social media data mining

Word of mouth has been shown to be an important factor in investment decisions; an investor's peers have been seen to share similar investment characteristics [8]. It has also been seen that neighbors' investment decisions are likely to be linked [9]. In an increasingly digital age, peers may now be online and geographically located around the world.

Much research has focused on predicting stock market prices using social media, with Twitter being a common platform choice [10]. However thriving online communities in which trading is discussed are even more prevalent in the case of cryptocurrencies; frequently discussing and forming opinion on relevant market events [11]. Online discussion can

978-1-5386-2726-6/17/$31.00 ©2017 IEEE

be harnessed to produce price predictions, for example, using sentiment to produce trading signals [12]. Twitter's relationship with cryptocurrency markets has been explored in several papers (for example, [2][12]). However, Hernandez et al. [13] discovered Twitter users communicating about Bitcoin behaved differently to the majority of Twitter users, in that they were not engaging in general social interaction but were focusing on a specific area of interest. Reddit is a social media platform that caters explicitly to subsets of users with particular interests, including cryptocurrencies, and for this reason is the data source used in this work.

B. Bubbles and epidemic detection

One approach to modelling financial asset bubbles is to repurpose models originally created by epidemiologists designed to track the spread of disease. A common model used in epidemiology is the SIR model, where the population is split into three categories: susceptible (S), infected (I), and recovered/removed (R). Members of the population transition from one category to another based on pre-defined rate formulae. Numbers in the infected category often exhibit a hump shaped pattern, rising rapidly at first and then declining, similar to the shape of a financial asset bubble [14].

An alternative approach to epidemic modelling used an HMM, applied to differenced incidence rates, to classify influenza time series data into epidemic and non-epidemic states [15]. Differenced incidence rates were used to de-trend the data and thus allow autoregressive modelling. The model worked well in an online environment (when it received new data points one by one) and, at each timestamp, produced a probability of the system being in the epidemic state. Abdullah [16] later experimented with the application of this model to Twitter message volumes to attempt to classify tweets into 'trending' and 'non-trending'.

III. Hidden Markov Models

In this work an HMM will be used to detect epidemic and non-epidemic states of social media usage and trading volume, due to its successful use in influenza epidemic outbreak prediction [15], where bubble-like behaviour is seen in relation to the number of individuals infected. An HMM has a number of underlying hidden states, which are transitioned between. Each state has associated possible observations. Given an observed series of data an HMM can be used to identify the most likely hidden state the model is in at each data point. The model has also previously been applied to Twitter data to categorise 'trending' vs 'non-trending' topics [16]. Components of this particular model are:

A. Number of hidden states

The model uses two hidden states, epidemic and non-epidemic, which are unobserved. E_t is an unobserved random variable to denote whether the system is in the epidemic state (1) or not (0) at time t.

B. Observation probability distribution

The hidden states have associated emission probabilities. Emission probabilities give the likelihood of seeing particular output values, and can be sampled from different distributions depending on which state the system is in. Differenced time series data (for example, for one of the social media indicators) is observed, where I_t represents the difference between the time series values at time t and t-1.

The model definition specifies that the conditional distribution of I_t is sampled from either an autoregressive process of order 1 (AR(1)) for the epidemic state ((1a) and (1b)) or sampled from a Gaussian white noise distribution for the non-epidemic state ((2a) and (2b)). Essentially, the epidemic state has interrelated changes, whereas the non-epidemic state has small random changes.

Hence the conditional distribution is defined as

$$I_1 \mid (E_1 = 1) \sim N(0, \sigma_1^2), \quad (1a)$$
$$I_t \mid (E_t = 1) \sim N(\rho * I_{t-1}, \sigma_1^2), \quad (1b)$$
$$I_1 \mid (E_1 = 0) \sim N(0, \sigma_0^2), \quad (2a)$$
$$I_t \mid (E_t = 0) \sim N(0, \sigma_0^2). \quad (2b)$$

As well as AR(1) being used in the aforementioned influenza model, it has been shown in previous work that an autoregressive process is an appropriate process to model time series dynamics during financial asset bubbles [17].

C. Transition probabilities

The HMM transitions between hidden states according to a transition probability matrix. This gives the probabilities of transitioning from one hidden state to another. Transitions exhibit the Markov property whereby the transition probability depends only on the current state, and the state history is forgotten. $P_{k,l}$ denotes the probability of transitioning to state l at time $t+1$ given that the current state is k at time t, i.e.

$$P_{k,l} = P(E_{t+1} = l \mid E_t = k).$$

D. Initial state distribution (parameter priors)

Prior parameter distributions are specified based on an understanding of the context. The prior parameter definitions below ensure that σ_1 (the sigma (standard deviation) associated with an epidemic state) has a higher prior value than σ_0 (the sigma associated with a non-epidemic state). Uniform distributions are chosen for the standard deviations, as suggested by Gelman [18]:

$$\theta_{low} \sim Unif(a, b),$$
$$\theta_{mid1} \sim Unif(\theta_{low}, b),$$
$$\theta_{mid2} \sim Unif(\theta_{mid1}, b),$$
$$\theta_{high} \sim Unif(\theta_{mid2}, b),$$
$$\sigma_0 \sim Unif(\theta_{low}, \theta_{mid1}),$$
$$\sigma_1 \sim Unif(\theta_{mid2}, \theta_{high}).$$

In the above a and b are hyper-parameters. The prior value for b is set as the maximum difference between two successive time series points. The prior value for a (set to be 1/10 of b) represents a lower bound on the non-epidemic state standard deviation. The value of a is used to attempt to ensure that the standard deviation does not converge to 0. The prior values for the remaining parameters are defined as in [15]:

$$\rho \sim Unif(-1,1),$$
$$P_{0,0} \sim Beta(0.5,0.5),$$
$$P_{1,1} \sim Beta(0.5,0.5).$$

Once the model and priors have been established, estimates of the optimal parameter values can be found using expectation maximisation (EM), a commonly used process proven by its use in multiple applications. This is an iterative process to find maximum likelihood parameters given an observed set of data. The parameters converged upon are those that provide the best fit with the observed data. Section V (B) discusses how the data is partitioned into multiple moving windows, and how state probabilities are retrieved from the model.

As discussed in Section II (B) the HMM in [15] added to previous epidemic-detecting HMM literature by using differenced data rather than unmodified influenza data; this enabled use of an AR(1) process on the de-trended data. Differenced data will also be used here, to the same effect, with the positive by-product that immediate changes can be recognised quicker than if absolute values were used. Section IV outlines where the data for the model is sourced.

IV. Input Factors

A. Social media indicators

As discussed in Section II (A), Twitter may not be the best platform to monitor cryptocurrency related discussion. Another social media website, Reddit, focusses more on the discussion and sharing of ideas and knowledge, and will be used in this work. Reddit is divided into *subreddits*; a subreddit is an area of Reddit dedicated to a particular topic. Each major cryptocurrency has its own subreddit. These Reddit communities have become the primary location for information dissemination relating to cryptocurrencies. Table I outlines the social media indicators that will be used in the model. The time series of each indicator will capture usage relating to a particular cryptocurrency's subreddit.

TABLE I
Social Media Indicators (Reddit)

Indicator	Description	Justification
Posts	The number of posts that occur on a particular subreddit, per day.	Volume of posts relating to a particular term is often used in social media analysis [2].
Subscriber growth	The number of new subscribers that a particular subreddit has, per day.	This shows new interest in the area from members outside of the community.
New authors	The number of new authors that post on a particular subreddit, per day.	This shows new members contributing within the community.

B. Ancillary use of trading volume

Trading volume is added as a fourth indicator, as a confirmatory signal; while most discussion on cryptocurrency subreddits pertains to, and may result in, further price movements, there are occasionally cases where a large amount of discussion is associated with some other unrelated topic. Volume is therefore important to confirm that social media activity relates to market activity. A model without volume input was constructed, and in fact the resulting trading strategy proved somewhat more profitable than that to be described below. The profit-depressing effect of volume is because volume tends to lag social media usage in moving to an epidemic state, and therefore positions are entered later; however the risk benefit of the additional volume input was considered to outweigh this lessened profitability.

V. Experiment Design

A. Experimental data

Bitcoin is the most common cryptocurrency examined in academic work, though recent work has started to expand the universe considered [19]. Here, four cryptocurrencies will be used: Bitcoin, Litecoin, Ethereum, and Monero. The methodology of this work could be applied to further cryptocurrencies, assuming they have an active subreddit.

All the required cryptocurrency trading data (daily closing price and volume) is sourced from CryptoCompare, a cryptocurrency data provider and community platform. Lesser known cryptocurrencies are often priced against Bitcoin, and for the purpose of this experiment, all cryptocurrency prices are transformed, when required, to a price in USD (for example, by using the Monero/Bitcoin and Bitcoin/USD exchange rate, Monero/USD can be calculated).

The experimental period chosen here is April 2015 to September 2016, with the exception of Ethereum. Ethereum was first listed on trading exchanges on August 8th 2015, and so the Ethereum time series will start from August 2015.

B. Moving window and state probability

Data is grouped into windows of length 100 data points; based on preliminary examination of the time series this is sufficiently long to encompass typical bubble and non-bubble regimes but short enough to be computationally viable. These windows are moved forward in time according to the mechanism used in [20] (a new piece of data being added to the window, and the oldest data removed). This moving window approach means that the model is always considering the most recent (social media and volume) data as it becomes available, as visually outlined in Fig. 1.

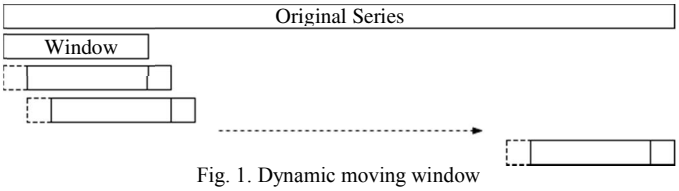

Fig. 1. Dynamic moving window

Once parameters have been fit, the HMM can output the probability of being in the epidemic state at each time series point. The most recent point (final point) in the moving window can be regarded as the current point (as no future data can be seen), and the probability of being in the epidemic state at this point is retrieved. Sherchan, Nepal, and Bouguettaya [21] used a similar technique of training an HMM on a moving window of data before retrieving the most probable state at the most recent data point in the window, as did Zhang [22] in the context of financial time series prediction.

In our case as the window moves forward a sequence of epidemic state probabilities are generated. The probability of being in the epidemic state can then be used within a trading strategy, as will be described in Section V (E).

C. *Preventing local maxima in the parameter fitting process*

The expectation maximization fitting process is susceptible to converging to local maxima. To overcome this, for each window of data 20 repeated parameter fittings are completed, and the fit that achieves the highest likelihood is then chosen as the final model. This multiple trial approach was used by Chan [23] while attempting to use HMMs to detect different regimes within trading markets.

D. *Use of distributed computation*

The parameter fitting process, combined with the moving window approach described above (where multiple near-identical data items are presented), takes considerable time to converge on a solution. It is common for researchers faced by such a situation to use cloud-based services to parallelize the computation [24]. One such facility is provided by Techila Technologies, which was used as a component of the work presented here. Techila is a service providing convenient integration with distributed cloud platforms such as Google Cloud and Amazon Web Services. The functionality was used in this work to run expectation maximization on multiple moving windows (examining different data) in parallel on a distributed grid of processors, allowing for a considerable reduction in computation time.

E. *Trading strategy*

Assessing the directional accuracy of predictions may be insufficient to assess their value. For example, the correctly predicted movements may frequently be small ones, possibly so small that trading costs would erode their profitability, while incorrect predictions could at the same time lead to large losses. A more persuasive way to validate the predictive power of the system is to convert the predictions of the HMM (that the system is in a state classed as either epidemic or non-epidemic) into a realistic trading strategy and assess its performance. Entry into an epidemic state is considered a buy signal, and exit from the epidemic state is considered a sell signal (to close the position and no longer hold the asset).

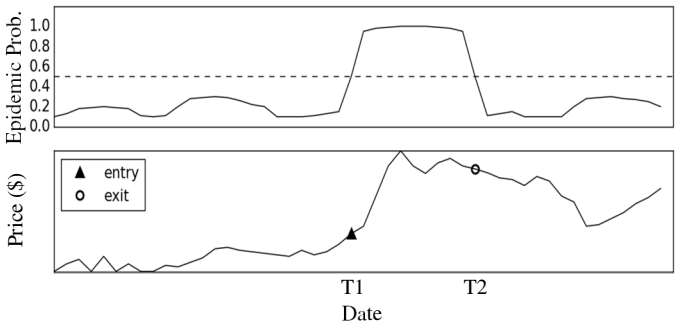

Fig. 2. Illustrative trading strategy entry and exit points (lower) based on probability of epidemic state (upper).

Fig. 2 shows a simulated example of this strategy. Fig. 2 (upper) shows the probability of the HMM being in the epidemic state at each time point. As the probability goes above 0.5 at time T1, the HMM is more likely to be in epidemic state than not, which is considered an entry point for the trading strategy; conversely, as the probability drops below 0.5 at time T2 the HMM is now more likely to be in the non-epidemic state, which is considered to be an exit signal. In Fig. 2 (lower), a corresponding simulated asset's price time series is shown with the entry and exit locations.

A separate HMM exists for each input factor, so several epidemic probabilities are produced at each timestamp. These can be combined into an overall prediction using either of the following mechanisms (both considered in Section VI):

1) Unanimous voting: A system where each HMM votes 'epidemic' or 'not epidemic'. The overall system needs to achieve consensus before an epidemic state is signaled.
2) Averaging: The individual probabilities are averaged. If the aggregated probability is above 0.5 the overall system is considered in an epidemic state.

Trading strategies can hold multiple assets at the same time. In this work, funds are allocated as follows: if one cryptocurrency is being signalled as epidemic, all the funds are allocated to this cryptocurrency; however if multiple cryptocurrencies are signalled as epidemic at the same time the funds are split between the cryptocurrencies equally.

The weightings are updated in the context of other positions; for example when one cryptocurrency is no longer signalled as being in the epidemic state, the position in that cryptocurrency is closed and the funds are reallocated to other open positions, purchasing additional units of these currencies.

A back-testing framework was built for evaluation of the trading strategy. Back-tests simulate a given trading strategy on historical data to determine its performance. The objective of a back-test is to produce results as close as possible to those that would have been achieved if the strategy had been trading real money over the tested period. As such, standard cryptocurrency exchange transaction fees have been included (chosen as 0.2% per transaction) in the simulations.

F. Benchmark strategy

To help assess the performance of the above trading strategy it is useful to define a benchmark strategy to which it can be compared. An equally weighted buy and hold strategy is used for this purpose: a total notional amount (in this case, $1,000) is divided by the number of assets being considered, with the assets being bought on the first day of the back-testing period and held until the last. This gives a reflection of how the overall market performed during the tested period. Buy and hold is generally regarded as a difficult benchmark to beat; this is especially true in the case of cryptocurrencies, for which prices have notoriously soared over short time horizons.

VI. Results

This section first examines the HMM state probabilities, and then validates the utility of these outputs by evaluating the multi-asset trading strategy defined above. Finally the parameters the HMM converges upon are considered, in order to better understand the reasons for the system's success.

As an example Fig. 3 (top) displays the probability of being in the epidemic state for the indicators relating to Ethereum. Fig. 3 (middle) shows how the price series for Ethereum evolves during the same time period. The best example of a bubble seen in the data gathered in this work occurs in the Ethereum price between January and April 2016 (where the price rises from around $1.50 to around $11); during this period all four indicator probabilities are consistently signaling the epidemic state, as would be expected. The entry and exit markers are placed in Fig. 3 (middle) to indicate when the overall system moves into (buy signal) and out of (exit signal) the epidemic state (using the unanimous voting methodology detailed in Section V (E)). The profitability of actions based on these signals will be considered later in the trading strategy commentary.

It should be noted, however, that some of the entry and exit points in Fig. 3 are located in regions exhibiting price behavior that is not bubble-like according to the definition in Section I. In the case of Ethereum this is due to sharp downward price movements causing brief epidemic-like social media and trading volume activity. One example of this can be seen in June 2016, where the price dropped from above $20 to just above $10. This resulted from the hack of an application built on Ethereum called the DAO, causing panic which was reflected in the price, and on social media for a number of weeks afterward.

Although the system profits from movements such as the above-described, as it detects the epidemic-like activity and generally enters at a low point in the price, this is still a fault in that it is a false positive for bubble detection. Ways to overcome this are left as extensions to the current work. For example, further indicators could be investigated, with an aim to find indicators that do not exhibit epidemic-like usage after a negative market event (which the current social media indicators appear susceptible to). Two candidates would be sentiment data and Google Trends. Additionally, while the focus of this work is exclusively on social media and volume the strategy could also use price related conditions (e.g. a requirement that price is rising) to avoid buying after crashes.

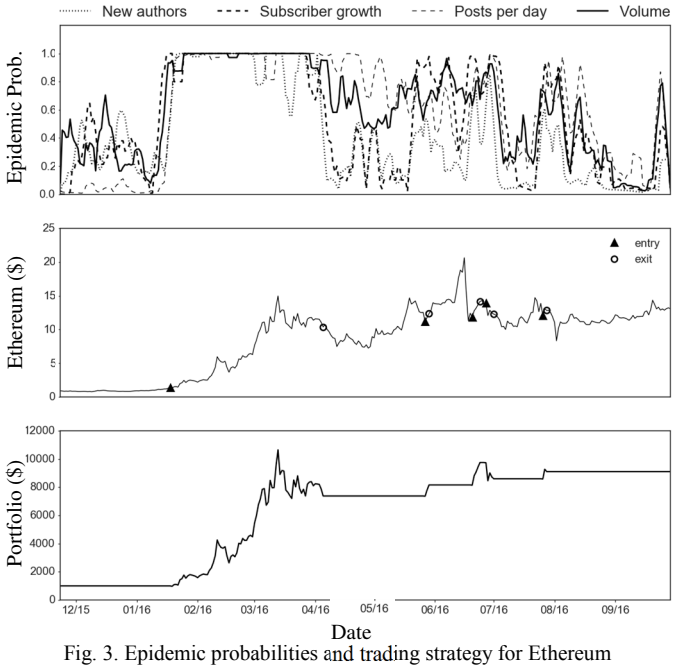

Fig. 3. Epidemic probabilities and trading strategy for Ethereum

The key advantage of the system presented in this paper is that trading positions are only entered for short periods when price rises are expected. For 222 days out of the 313 days in the tested period the strategy does not have a position (as shown by the entry and exit points in Fig. 3 (middle) and the portfolio value in Fig. 3 (bottom)); a multi-asset strategy as used below can at such a time allocate more funds to buying other cryptocurrencies signalled as being in their epidemic state. The next section examines the profitability of this.

A. Trading strategy: performance

This section shows the results of the multi-asset trading strategy described in Section V (E), which aims to allocate money to buy whichever cryptocurrencies are signaled as being in an epidemic state (using unanimous voting). Fig. 4 shows the price series for each cryptocurrency considered, with their entry and exit points.

As can be seen in Fig. 4 (top), Monero undergoes a sudden and substantial price rise near September 2016 as the price rises from around $2.20 to around $12; this is the second best example of a bubble in the data collected, after the Ethereum bubble already mentioned (January – April 2016). The strategy clearly profits from the Monero bubble. However Litecoin has no price rises comparable to Monero or Ethereum, while Bitcoin shows a general sustained price rise throughout the testing period but also some periods of increased growth from which the system can profit. Fig. 4 (bottom) shows the portfolio value over the tested period. The portfolio starts with $1000 and can be seen to make most profit in the time period around the Ethereum bubble (January – April 2016) and the Monero bubble (near September 2016).

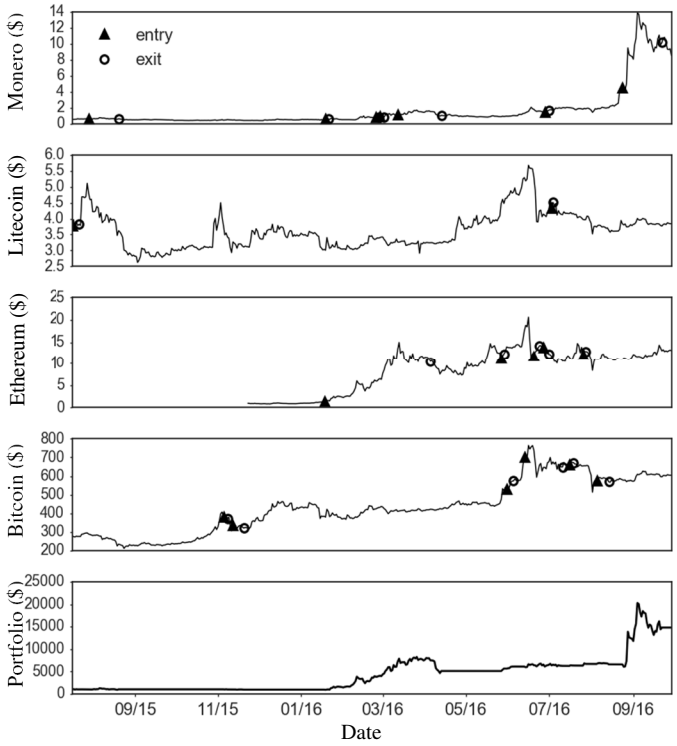

Fig. 4. Price series, and entry and exit points for each cryptocurrency, and overall portfolio value (last).

Table II shows the current unanimous voting HMM strategy evaluated against common trading strategy metrics. It can be seen that the strategy outperforms the buy and hold strategy on all metrics including the Sharpe and Sortino ratios. It also has a smaller percentage drawdown occurring over a shorter period. The final column in Table II shows the profitability of a modified version of the HMM strategy which will be discussed in the next section.

TABLE II
PERFORMANCE COMPARISON OF TRADING STRATEGIES

Metric	Buy and hold	HMM Strategy (unanimous voting)	HMM Strategy (averaging)
Ending portfolio (starting $1000)	$7,939	$14,804	$8,751
Returns	693.9%	1380.4%	775.1%
Sharpe ratio	1.77	1.93	1.48
Sortino ratio	2.63	2.64	2.29
Position number	4	20	33
Maximum drawdown (%)	50.03%	35.19%	32.68%
Maximum drawdown (duration)	47 days	12 days	12 days

B. *Trading strategy: voting vs. averaged probabilities*

An alternative to the unanimous voting used up to this point is to instead take the average of the factor-specific epidemic probabilities. Fig. 5 (top) visualizes the resulting epidemic probability of the system in the case of Ethereum, and Fig. 5 (bottom) displays the trades made by the averaging method (using the default epidemic threshold of 0.5).

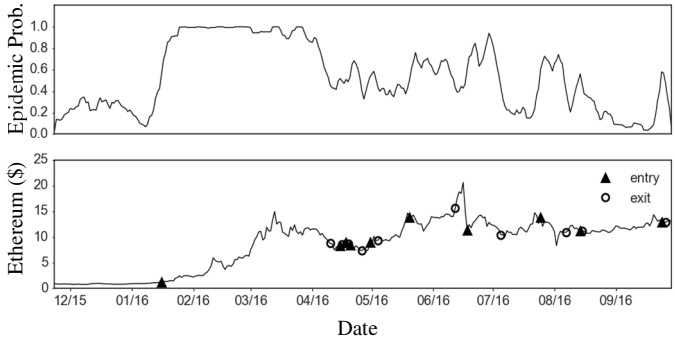

Fig. 5. Entry and exit points (lower) for Ethereum based on aggregated probability of epidemic state (upper).

During May 2016 the epidemic probability remains around 0.5, causing the averaging variant to repeatedly enter and exit positions, which is not ideal given the transaction fees associated with trading. Averaging (with the current epidemic/non-epidemic threshold of 0.5) takes 33 positions for the multi-asset strategy, instead of 20 for the voting method, as shown in Table II. Although the averaging method still outperforms buy and hold, the system's profitability is reduced greatly compared to unanimous voting.

Preliminary investigation into the impact of changing the epidemic/non-epidemic threshold suggests an increase in the threshold decreases the number of trades while increasing the overall profitability of those trades, thus improving overall strategy performance. Further work could explore setting a threshold via an optimization process examining the profitability of different threshold values on historical data.

C. *Parameters converged upon*

As mentioned in Section V (B) before state probabilities can be retrieved the HMM is trained on previous data observations in order to provide estimated values for a number of parameters (defined in Section III). Table III shows the values converged upon for one example cryptocurrency/factor combination: the 'new author' time series on the Ethereum subreddit (similar characteristics are found for the above parameters when examining other cryptocurrency/factor combinations). These values have for simplicity been averaged over those generated from all training window periods.

TABLE III
POSTERIOR MEAN OF PARAMETERS FOR ETHEREUM/NEW AUTHORS

Parameter	Posterior mean
ρ	0.80
σ_0	0.82
σ_1	3.20
$P_{0,0}$	0.86
$P_{1,1}$	0.72

The positive value of ρ shows the positive impact each data value has on the next, once in the epidemic state, and justifies the use of an autoregressive process. The values of

$P_{0,0}$ and $P_{1,1}$ suggest that once the HMM is in a particular state it is likely to remain in that state at the next data observation; this is expected as it is likely that epidemic states will continue for a number of data points; it is also advantageous for the trading strategy, as when receiving persistent signals the strategy does not change positions too frequently. The model is slightly more likely to exit the epidemic state ($P_{1,0}$ = 0.28) than it is to exit the non-epidemic state ($P_{0,1}$ = 0.14). This reflects the fact that epidemic states are expected to be shorter-lived than non-epidemic states. The sigma associated with the epidemic state, σ_1, is larger than the sigma associated with the non-epidemic state, as intended by the prior parameter choices.

The above values in Table III were time-averaged; however it should be noted that variation in parameter values can occur depending on the data period used for training. During the large bubble seen in the Ethereum price (between January and April 2016 – Fig. 3), the probability of remaining in the epidemic state ($P_{1,1}$) approaches 1. Such variability demonstrates the advantage of using a moving window trained on the most recent data.

VII. Conclusions

This work demonstrates how epidemic detection techniques can be applied to social media data to predict cryptocurrency price bubbles and provides some empirical evidence that bubbles mirror the social epidemic-like spread of an investment idea. To show this, an HMM methodology originally designed to detect influenza outbreaks was applied to community-based social media usage and trading volume relating to certain cryptocurrencies, categorizing usage into epidemic and non-epidemic states. The utility of state probabilities were validated by transforming them into a profitable trading strategy that outperformed a comparable benchmark. It is notable that no price-related trading signals were used in the trading strategy, only social media usage and trading volume were considered.

Aside from the HMM methodology used in this work, there is another well-known epidemic model, the SIR model, and it is intended to explore the use of this for cryptocurrency bubble detection in the future. This work has demonstrated a strong relationship between Reddit usage and cryptocurrency prices; as a result, the work has highlighted Reddit as a valuable source of information. Due to the way Reddit is structured into subreddits it also enables community dynamics to be analysed. There is a clear separation of users based on their interests. Other platforms such as Twitter do not have such distinct communities. It is hoped that this work will motivate further research into the role of Reddit within cryptocurrency markets and also encourage further exploration of Reddit within other areas in the field of social media data mining.

References

[1] S. Nakamoto, “Bitcoin: A Peer-to-Peer Electronic Cash system.” 2008.

[2] D. Garcia, C. Tessone, P. Mavrodiev and N. Perony, "The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy", Journal of The Royal Society Interface, vol. 11, no. 99, 2014.

[3] J. Fry and E. Cheah, "Negative bubbles and shocks in cryptocurrency markets", International Review of Financial Analysis, vol. 47, pp. 343-352, 2016.

[4] R. Z. Aliber and C. P. Kindleberger, Manias, panics and crashes: a history of financial crises. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, 2015.

[5] M. K. Brunnermeiera and M. Oehmkeb, “Bubbles, Financial Crises, and Systemic Risk”. Handbook of the Economics of Finance, vol. 2, pp. 1221–1288. 2013.

[6] R. J. Shiller, "Speculative asset prices", Amer. Econ. Rev., vol. 104, no. 6, pp. 1486-1517, 2014.

[7] R. J. Shiller, Irrational Exuberance Ed. 3. Princeton University Press, pp. 2, 2015.

[8] R. Heimer, "Peer Pressure: Social Interaction and the Disposition Effect", Review of Financial Studies, vol. 29, no. 11, pp. 3177-3209, 2016.

[9] V. Pool, N. Stoffman and S. Yonker, "The People in Your Neighborhood: Social Interactions and Mutual Fund Portfolios", The Journal of Finance, vol. 70, no. 6, pp. 2679-2732, 2015.

[10] J. Bollen, H. Mao and X. Zeng, "Twitter mood predicts the stock market", Journal of Computational Science, vol. 2, no. 1, pp. 1-8, 2011.

[11] M. Linton, E. Teo, E. Bommes, C. Chen and W. Härdle, "Dynamic Topic Modelling for Cryptocurrency Community Forums", Applied Quantitative Finance, pp. 355-372, 2017.

[12] D. Garcia and F. Schweitzer, "Social signals and algorithmic trading of Bitcoin", Royal Society Open Science, vol. 2, no. 9, 2015.

[13] I. Hernandez, M. Bashir, G. Jeon and J. Bohr, "Are Bitcoin Users Less Sociable? An Analysis of Users’ Language and Social Connections on Twitter", HCI International 2014 - Posters’ Extended Abstracts, pp. 26-31, 2014.

[14] R. J. Shiller and J. Pound, “Survey evidence on diffusion of interest and information among investors,” Journal of Economic Behavior & Organization, vol. 12, no. 1, pp. 47–66, 1989.

[15] M. A. Martínez-Beneito, D. Conesa, A. López-Quílez, and A. López-Maside, “Bayesian Markov switching models for the early detection of influenza epidemics,” Statistics in Medicine, vol. 27, no. 22, pp. 4455–4468, 2008.

[16] S. Abdullah and X. Wu, “An Epidemic Model for News Spreading on Twitter,” 2011 IEEE 23rd International Conference on Tools with Artificial Intelligence, 2011.

[17] E. Shtatland and T. Shtatland, "Another Look at Low-Order Autoregressive Models in Early Detection of Epidemic Outbreaks and Explosive Behaviors in Economic and Financial Time Series", NESUG’20 Proceedings, Paper 363, 2008.

[18] A. Gelman, “Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper),” Bayesian Analysis, vol. 1, no. 3, pp. 515–534, 2006.

[19] Y. Kim, J. Kim, W. Kim, J. Im, T. Kim, S. Kang and C. Kim, "Predicting Fluctuations in Cryptocurrency Transactions Based on User Comments and Replies", PLOS ONE, vol. 11, no. 8, 2016.

[20] S. H. Park, J. H. Lee, J. W. Song, and T. S. Park, “Forecasting Change Directions for Financial Time Series Using Hidden Markov Model,” Rough Sets and Knowledge Technology Lecture Notes in Computer Science, pp. 184–191, 2009.

[21] W. Sherchan, S. Nepal, and A. Bouguettaya, “A Trust Prediction Model for Service Web,” 2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications, 2011.

[22] Y. Zhang, "Prediction of financial time series with hidden Markov models," M.Sc. thesis, Simon Fraser University, May 2004.

[23] E. Chan, Machine trading, 1st ed. Hoboken, New Jersey: John Wiley & Sons, pp. 101-105, 2017.

[24] J. Yang, B. Lin, W. Luk, and T. Nahar, “Particle filtering-based Maximum Likelihood Estimation for financial parameter estimation,” 2014 24th International Conference on Field Programmable Logic and Applications (FPL), 2014.