Dipple *et al. Applied Network Science* (2020) 5:17
https://doi.org/10.1007/s41109-020-00259-1

Applied Network Science

RESEARCH Open Access

# Using correlated stochastic differential equations to forecast cryptocurrency rates and social media activities

Stephen Dipple[1,2*], Abhishek Choudhary[3], James Flamino[1,2], Boleslaw K. Szymanski[1,2,4] and G. Korniss[1,2]

*Correspondence:
swdipple@gmail.com
[1]Department of Physics, Applied Physics, and Astronomy, Rensselaer Polytechnic Institute, 110 8th Street, 12180-3590 Troy, NY, USA
[2]Network Science and Technology Center, Rensselaer Polytechnic Institute, 110 8th Street, 12180-3590 Troy, NY, USA
Full list of author information is available at the end of the article

**Abstract**

The growing interconnectivity of socio-economic systems requires one to treat multiple relevant social and economic variables simultaneously as parts of a strongly interacting complex system. Here, we analyze and exploit correlations between the price fluctuations of selected cryptocurrencies and social media activities, and develop a predictive framework using noise-correlated stochastic differential equations. We employ the standard Geometric Brownian Motion to model cryptocurrency rates, while for social media activities and trading volume of cryptocurrencies we use the Geometric Ornstein-Uhlenbeck process. In our model, correlations between the different stochastic variables are introduced through the noise in the respective stochastic differential equation. Using a Maximum Likelihood Estimation on historical data of the corresponding cryptocurrencies and social media activities we estimate parameters, and using the observed correlations, forecast selected time series. We successfully analyze and predict cryptocurrency related social media and the cryptocurrency market itself with a reasonable degree of accuracy. In particular, we show that our method has impressive accuracy in predicting whether a cryptocurrency market will increase or decrease a day in the future, a significant result with regards to investing and trading cryptocurrencies.

**Keywords:** Cryptocurrency, Social media activity, Stochastic differential equations, Maximum likelihood estimation, Predictive modeling

## Introduction

Increasingly interconnected financial systems and online social networks present both critical challenges and opportunities. Volatility in the former (e.g., cryptocurrency rates) can give rise to increased volume of activities in online social networks on relevant topics, while sentiments and rumors in online social networks can also have a significant impact on the corresponding financial time series. In this work, we aim to expand upon the study of (correlated) stochastic differential equations (SDEs) and their application to cryptocurrency markets. A large number of studies have demonstrated that complex socio-economic systems (Máté and Néda 2016), more specifically stocks (Merton 1973; 1971; Black and Scholes 1973; Wilmott et al. 1995; Bouchaud and Potters 2000; Mantegna and Stanly 2000; Reddy and Clinton 2016; Øksendal 2003), commodities (Schwartz 1997; Mejía Vega 2018), and cryptocurrencies (Cretarola and Figà-Talamanca 2019b; Cretarola

et al. 2019; Cretarola and Figà-Talamanca 2019) can be modeled using SDEs. Further, correlations do exist between some stocks (Mantegna and Stanly 2000; Teng et al. 2016; Plerou et al. 2002; Sándor and Néda 2015; Onnela et al. 2003), various cryptocurrencies (Saha 2018; Chaim and Laurini 2019), and possibly exist between online social media activities. We exploit these correlations and construct a general predictive method for sets of cryptocurrency markets. In addition, we incorporate social media information, not only to augment our predictions of the trading rates of cryptocurrencies, but also predict the activity in these social media in a similar fashion by modeling them with SDEs as well. Our results show that our prediction method yields fairly accurate results consistently outperforming our baseline measurements. Our method's most applicable result is the model's impressive accuracy in predicting whether the trading rate of cryptocurrencies would increase or decrease during the following day during volatile periods.

A prevalent and common choice is the Geometric Brownian Motion (GBM) which has been used for modeling stocks (Merton 1973; Black and Scholes 1973; Wilmott et al. 1995; Reddy and Clinton 2016; Øksendal 2003) and more recently also for cryptocurrencies (Cretarola et al. 2019; Cretarola and Figà-Talamanca 2019; Tarnopolski 2017; Kreuser and Sornette 2018; Wu et al. 2018). There is, however, debate on the extent these assets follow GBM dynamics. While GBM operates using Gaussian noise, there is evidence that the noise distribution of stocks such as the S&P 500 has curvature different than a normal distribution (Mantegna and Stanly 2000). Implication of this difference is most prominent in the frequency of rare events as a result of heavy-tailed noise, also possibly present for cryptocurrencies (Kreuser and Sornette 2018; Wu et al. 2018; Fry 2018). For small sample sizes, this difference is small enough that many continue to use GBM and other variants for modeling. In this paper, we will employ the GBM to model the dynamics of cryptocurrencies.

In general, stocks can influence each other and determining the mechanism for these influences is the focus of many areas of economics (Bouchaud and Potters 2000). There is also reasonable evidence to suggest that two (Teng et al. 2016) or multiple (Plerou et al. 2002) stocks can be correlated via their stochastic terms. The difficulty with these approaches is tackling parameter estimation.

Parameters of a GBM process can vary both with time and value of the function in non-linear ways (Bassler et al. 2007; Cretarola and Figà-Talamanca 2019a). This difficulty is compounded due to the introduction of noise, which makes disentangling complex parameters estimation, even from Gaussian noise, non-trivial. These considerations also apply to time variations in the correlation of multiple stocks (Teng et al. 2016). Another interesting consequence is that correlation can artificially arise due to random chance, which becomes important when a large number of time series are considered (Plerou et al. 2002).

Attempts have been made to include the role of social media (Rosati et al. 2018), such as studying the impact of "silent majority" vs. "vocal minority" (Mai et al. 2018)) and contextual or sentiment information from media sources (Lamon et al. 2017; Kim et al. 2016; Bollen et al. 2011; Phillips and Gorse 2017). The latter attempts have mostly utilized machine learning in some fashion, where coefficients are optimized to best fit the data. The trade off is the possibility of over-fitting causing predictions to become inaccurate. We instead model social media activity volumes by SDEs as well, where the correlations to cryptocurrencies are introduced via correlated noise. We can then invert this

method in order to predict social media activity using information about cryptocurrency markets. Various predictive methods have been developed for predicting activity in social media (Yao et al. 2018), and augmenting them with auxiliary information such as the cryptocurrency market can improve those methods. A related approach has also been developed in Cretarola and Figà-Talamanca (2019b); Cretarola et al. (2019); Cretarola and Figà-Talamanca (2019) using “market attention" as a relevant correlated stochastic variable.

Naturally, social media activity dynamics (referencing cryptocurrencies) and daily trading volumes exhibit significantly different features from those of the underlying cryptocurrencies. In particular, in these processes, after some possibly large spikes or drops (in response to some external “random" events) the corresponding stochastic variables have a tendency to return to some long-run mean. We hypothesize that these processes can be approximated by the Geometric Ornstein-Uhlenbeck (GOU) process (Schwartz 1997; Mejía Vega 2018).

In this paper, we expand applications of SDEs and provide a framework on how to use correlations between time series in a data set. This is demonstrated by generating synthetic data and testing our method's predictive power when the ground truth is known. We then model the cryptocurrency market, where we show our method's predictive power, most notably at correctly identifying increases and decreases in market value. Naturally, we also model social media using SDEs and examine our prediction's accuracy in this area.

## Theory and background

In this work, we begin with a Markovian SDE with a solution of the type,

$$S(t_i) = f(S(t_{i-1}), \psi(t_i), \overrightarrow{\beta}), \tag{1}$$

where $S(t_i)$ is the function value at time $t_i$, $\psi(t_i)$ is an independent, identically distributed, normal random variable realization at time $t_i$ with mean zero and variance unity, and $\overrightarrow{\beta}$ is the relevant parameter space. It is important to note that for the derivation of these equations, there is no correlation between time steps $\forall i, j \in \mathbb{Z}, \langle \psi(t_i), \psi(t_j) \rangle = \delta_{ij}$. While this may be true in theory, in practice there is typically some level of temporal correlation. In addition, we also restrict this family of solutions to one where $\psi$ can easily be isolated as seen in the following equation,

$$\psi(t_i) = g(S(t_i), S(t_{i-1}), \overrightarrow{\beta}). \tag{2}$$

While our framework can be applied to more general equations of the above forms, in this work we will exclusively use the GBM (Wilmott et al. 1995; Reddy and Clinton 2016) and the GOU process (Schwartz 1997; Mejía Vega 2018).

### The Geometric Brownian Motion

The GBM has been the simplest model to describe stock- or asset-price dynamics. In this work, we model cryptocurrency closing price as a GBM. The essence of this asset price dynamics is that the relative price change $dS/S$ can be split into a deterministic and a random component (Wilmott et al. 1995),

$$dS(t) = S(t)\mu dt + S(t)\sigma dW(t), \tag{3}$$

where $\mu$ is the expected rate of return over time, $\sigma$ is the volatility (the amplitude of the noise), and $dW(t) = W(t+dt) - W(t)$ is the infinitesimal Wiener Process ($\langle dW(t)\rangle = 0$ and $\langle dW^2(t)\rangle = dt$). Defining $U(t) \equiv \ln S(t)$ and using Itô's Lemma (Wilmott et al. 1995; Gardiner 1985; Itô 1944) one arrives at

$$dU(t) = \bar{\mu}dt + \sigma dW(t)\ , \tag{4}$$

with $\bar{\mu} = \mu - \frac{\sigma^2}{2}$. Ito's stochastic calculus yields the exact solution (Wilmott et al. 1995; Gardiner 1985; Itô 1944)

$$U(t) = U(0) + \bar{\mu}t + \sigma\psi(t)\sqrt{t}, \tag{5}$$

where $\psi(t)$ is a normal random variable with zero mean and unit variance. Equation (5) implies that the transition probability distribution for the variable $U$ can be written as

$$P(U(t)|U(0)) = \frac{1}{\sigma\sqrt{2\pi t}}\exp\left(-\frac{(U(t) - U(0) - \bar{\mu}t)^2}{2\sigma^2 t}\right). \tag{6}$$

Further, we can solve Eq. (5) for $\psi$,

$$\psi(t) = \frac{1}{\sigma\sqrt{t}}\left(U(t) - U(0) - \bar{\mu}t\right)\ . \tag{7}$$

The above expression can be used to extract and measure the "noise" from the empirical data where the underlying process is hypothesized to be a GBM.

#### The Geometric Ornstein-Uhlenbeck process

In this paper, we will model relevant social-media activity volumes and daily trading volumes of cryptocurrencies as GOU processes (Schwartz 1997; Mejía Vega 2018). The GOU process (also referred to as the exponential Ornstein-Uhlenbeck process) has been used to model commodity prices. While the fluctuations here too can exhibit large spikes or drops, they have a tendency to revert to the some long-run mean,

$$dS(t) = S(t)\kappa(\theta - \ln S(t))dt + S(t)\sigma dW(t), \tag{8}$$

where $\kappa$ is the relaxation rate to the long-run mean of the logarithmic variable. Again, by defining the logarithmic variable $U(t) \equiv \ln S(t)$ and using Ito's lemma, one obtains

$$dU(t) = \kappa(\bar{\theta} - U(t))dt + \sigma dW(t)\ , \tag{9}$$

with $\bar{\theta} = \theta - \frac{\sigma^2}{2\kappa}$, which is the standard Ornstein-Uhlenbeck (OU) process (Gardiner 1985). Employing Itô's isometry (Øksendal 2003; Gardiner 1985; Itô 1944; Franco 2003), this can be integrated exactly,

$$U(t) = U(0)\mathrm{e}^{-\kappa t} + \bar{\theta}(1 - \mathrm{e}^{-\kappa t}) + \sigma\psi(t)\sqrt{\frac{1 - \mathrm{e}^{-2\kappa t}}{2\kappa}}\ , \tag{10}$$

where $\psi(t)$ is a normal random variable with zero mean and unit variance. For the transition probability density, one then has,

$$P(U(t)|U(0)) = \left(\frac{\pi\sigma^2}{\kappa}\left(1 - \mathrm{e}^{-2\kappa t}\right)\right)^{-\frac{1}{2}} \times \exp\left(-\frac{\left(U(t) - \bar{\theta} - (U(0) - \bar{\theta})\mathrm{e}^{-\kappa t}\right)^2}{\frac{\sigma^2}{\kappa}(1 - \mathrm{e}^{-2\kappa t})}\right). \tag{11}$$

Further, from Eq. (10), we find,

$$\psi(t) = \frac{U(t) - U(0)\mathrm{e}^{-\kappa t} - \bar{\theta}(1 - \mathrm{e}^{-\kappa t})}{\sigma\sqrt{\frac{1-\mathrm{e}^{-2\kappa t}}{2\kappa}}}. \tag{12}$$

which can be utilized to extract and measure the “noise" from the empirical data where the underlying process is hypothesized to be a GOU process.

#### Time-Series generation, transition probability distributions, and noise reconstruction

Both the GBM and the GOU processes are Markovian. Further, the above solutions for $U(t) = \ln S(t)$ and the corresponding transition probability densities are valid for an arbitrary time interval $(0, t)$. Due to time homogeneity, the above equations are valid for any initial starting value and can be applied beginning at time step $t_{i-1}$ and ending at $t_i$. We assume that these time steps are evenly spaced such that $\forall i \in \mathbb{Z}_+, t_i - t_{i-1} = \Delta t > 0$. Next at each time step, we will redefine initial conditions such that the initial time value at each time step is $t_{i-1}$. Therefore, for the GBM, for an arbitrary *finite* time difference $\Delta t$,

$$S(t_i) = S(t_{i-1})\exp\left(\bar{\mu}\Delta t + \sigma\psi(t_i)\sqrt{\Delta t}\right), \tag{13}$$

$$P(S(t_i)|S(t_{i-1})) = \frac{1}{S(t_i)\sigma\sqrt{2\pi\Delta t}}\exp\left(-\frac{\left(\ln\left(\frac{S(t_i)}{S(t_{i-1})}\right) - \bar{\mu}\Delta t\right)^2}{2\sigma^2\Delta t}\right), \tag{14}$$

$$\psi(t_i) = \frac{1}{\sigma\sqrt{\Delta t}}\left(\ln\left(\frac{S(t_i)}{S(t_{i-1})}\right) - \bar{\mu}\Delta t\right). \tag{15}$$

Similarly, for GOU process we have,

$$S(t_i) = \exp\left(\ln\left(S(t_{i-1})\right)\mathrm{e}^{-\kappa\Delta t} + \bar{\theta}(1 - \mathrm{e}^{-\kappa\Delta t}) + \sigma\psi(t_i)\sqrt{\frac{1 - \mathrm{e}^{-2\kappa\Delta t}}{2\kappa}}\right), \tag{16}$$

$$P(S(t_i)|S(t_{i-1})) = \frac{1}{S(t_i)}\left(\pi\frac{\sigma^2}{\kappa}\left(1 - \mathrm{e}^{-2\kappa\Delta t}\right)\right)^{-\frac{1}{2}} \times$$
$$\exp\left(-\frac{\left(\ln S(t_i) - \bar{\theta} - (\ln S(t_{i-1}) - \bar{\theta})\mathrm{e}^{-\kappa\Delta t}\right)^2}{\frac{\sigma^2}{\kappa}\left(1 - \mathrm{e}^{-2\kappa\Delta t}\right)}\right), \tag{17}$$

$$\psi(t_i) = \frac{\ln S(t_i) - \bar{\theta} - (\ln S(t_{i-1}) - \bar{\theta})\mathrm{e}^{-\kappa\Delta t}}{\sigma\sqrt{\frac{1-\mathrm{e}^{-2\kappa\Delta t}}{2\kappa}}}. \tag{18}$$

Note that the above equations for $S(t_i)$ and $\psi(t_i)$ are precisely of the form of Eqs. (1) and (2).

#### Maxim likelihood parameter estimation for single-variable GBM and GOU processes

Here, we attempt to find parameters that maximize the likelihood that the given time series is a realization of our governing equation. Our maximization function (the log likelihood) is then, $L(\vec{\beta}) = \ln\left(\prod_{i=1}^{N} P(U_i|U_{i-1};\vec{\beta})\right)$, where we use our previous definition of $U_i = \ln S(t_i)$ and our parameter space of $\vec{\beta}$. Full details on MLE and how we use it to recover parameters can be seen in Appendix 1 (Franco 2003; Tang and Chen 2009;

Johnson and Wichern 2002; Rayner 1985). When maximizing $L$ for the GBM process, one easily finds (Hurn et al. 2003) (Appendix 1.1),

$$\hat{\hat{\mu}} = \frac{1}{N\Delta t}\sum_{i=1}^{N}(U_i - U_{i-1}), \tag{19}$$

$$\hat{\sigma}^2 = \frac{1}{N\Delta t}\sum_{i=1}^{N}\left(U_i - U_{i-1} - \hat{\hat{\mu}}\Delta t\right)^2. \tag{20}$$

For the GOU process, parameters that maximize $L$ are Mejía Vega (2018); Franco (2003); Tang and Chen (2009) (Appendix 1.2),

$$\begin{aligned} e^{-\hat{\kappa}\Delta t} &= \frac{\left(\frac{1}{N}\sum_{i=1}^{N}U_iU_{i-1}\right)-\left(\frac{1}{N}\sum_{i=1}^{N}U_{i-1}\right)\left(\frac{1}{N}\sum_{i=1}^{N}U_i\right)}{\left(\frac{1}{N}\sum_{i=1}^{N}U_{i-1}^2\right)-\left(\frac{1}{N}\sum_{i=1}^{N}U_{i-1}\right)^2} \\ &= \frac{\frac{1}{N}\sum_{i=1}^{N}\left(U_i-\frac{1}{N}\sum_{j=1}^{N}U_j\right)\left(U_{i-1}-\frac{1}{N}\sum_{j=1}^{N}U_{j-1}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(U_{i-1}-\frac{1}{N}\sum_{j=1}^{N}U_{j-1}\right)^2}, \end{aligned} \tag{21}$$

$$\hat{\hat{\theta}} = \frac{1}{N(1-e^{-\hat{\kappa}\Delta t})}\sum_{i=1}^{N}(U_i - U_{i-1}e^{-\hat{\kappa}\Delta t}), \tag{22}$$

$$\hat{\sigma}^2 = \frac{2\hat{\kappa}}{N}\sum_{i=1}^{N}\frac{\left(U_i-\hat{\hat{\theta}}-(U_{i-1}-\hat{\hat{\theta}})e^{-\hat{\kappa}\Delta t}\right)^2}{1-e^{-2\hat{\kappa}\Delta t}}. \tag{23}$$

It is interesting to note that the parameter estimation for $e^{-\hat{\kappa}\Delta t}$ is precisely the normalized autocorrelation of $U$ with time difference $\Delta t$ (Gardiner 1985; Tang and Chen 2009). This parameter estimation scheme allows one to extract the noise from the individual time series via Eqs. (15) and (18) , and conversely, generate and forecast the individual time series using Eqs. (13) and (16), for the GBM and the GOU process, respectively

### Noise-Correlated stochastic differential equations and application of the cholesky decomposition

For multiple time series, each stochastic variable is modeled by a proper corresponding SDE (with its specific parameters). We will use the indices of the $S_j(t)$ variables to distinguish between the GBM or GOU processes, such that of a total number of $d = n_{\text{GBM}} + n_{\text{GOU}}$ stochastic variables, the first $n_{\text{GBM}}$ are GBM variables, while the remaining $n_{\text{GOU}}$ are GOU variables. Then for $j = 1, \ldots, n_{\text{GBM}}$ ($j \in$ GBM for short),

$$dS_j(t) = S_j(t)\mu_j dt + S_j(t)\sigma_j dW_j(t), \tag{24}$$

While for $j = n_{GBM} + 1, \ldots, d$ ($j \in$ GOU for short),

$$dS_j(t) = S_j(t)\kappa_j(\theta_j - \ln S_j(t))dt + S_j(t)\sigma_j dW_j(t). \tag{25}$$

Most importantly, we consider the above system of SDEs where correlations are possibly present among the infinitesimal Wiener processes (and in turn, *across* the various time series),

$$\langle dW_j(t)dW_k(t)\rangle = dt\,\rho_{jk} \tag{26}$$

($\rho_{jj} = 1$ for all $j$). Analogously to the single-variable cases (with $U_j(t) = \ln S_j(t)$), we define the (finite) normalized noise variables ($\Delta t = t_i - t_{i-1} > 0$),

$$\psi_j(t_i) = \frac{1}{\sigma_j\sqrt{\Delta t}}\left(U_j(t_i) - U_j(t_{i-1}) - \bar{\mu}_j\Delta t\right) \tag{27}$$

for $j \in$ GBM, and

$$\psi_j(t_i) = \frac{U_j(t_i) - U_j(t_{i-1})\mathrm{e}^{-\kappa_j \Delta t} - \bar{\theta}_j(1 - \mathrm{e}^{-\kappa_j \Delta t})}{\sigma_j \sqrt{\frac{1-\mathrm{e}^{-2\kappa_j \Delta t}}{2\kappa_j}}} \tag{28}$$

for $j \in$ GOU. The correlations between these finite-time normalized noise variables,

$$\langle \psi_j(t)\psi_k(t)\rangle = \rho^{\psi}_{jk} \tag{29}$$

($\rho^{\psi}_{jj} = 1$ for all $j$), can be obtained using Ito's calculus and can be expressed in terms of the correlations between the corresponding underlying infinitesimal Wiener processes (see Appendix 2). In what follows, we will use the correlations between the finite-time noise variables as our numerical scheme directly utilizes those. By construction, these normal variables have zero mean and unit variance. The empirically measured covariances between these variables are

$$\rho^{\psi}_{jk} = Cor(\psi_j, \psi_k) = \langle \psi_j \psi_k \rangle = \frac{1}{N}\sum_{i=1}^{N} \psi_j(t_i)\psi_k(t_i) \tag{30}$$

Note that these empirical correlations precisely correspond to the MLE for these parameters. This follows from the fact that the MLE for multivariate normal variables again are equal to the sample means and sample covariances (Johnson and Wichern 2002; Rayner 1985) (see also Appendix 3 for some details). Further, for the two-point correlation functions of the scaled GOU processes in the stationary regime one has (Gardiner 1985; Singh et al. 2018)

$$e^{-\kappa_j \Delta t} = \frac{\langle (U_j(t+\Delta t) - \langle U\rangle)(U_k(t) - \langle U\rangle)\rangle}{\langle (U_j(t) - \langle U\rangle)(U_k(t) - \langle U\rangle)\rangle}\,, \tag{31}$$

$j, k \in$ GOU. Setting $k = j$, this becomes the normalized autocorrelation function, and we can utilize the above expression as an empirical estimate for $\hat{\kappa}_j$ (equivalent to the single-variable MLE Eq. (21)).

Next, we decompose the original, possibly correlated noise variables into a linear combination of *independent* ones (Sauer 2013),

$$\psi_j = \sum_k C_{jk}\psi'_k, \tag{32}$$

where $\psi'$ variables are independent normal random variables $\left(\left\langle \psi'_j \psi'_k \right\rangle = \delta_{jk}\right)$. The matrix $C$ can be determined from the covariance matrix of the original noise variables,

$$\rho^{\psi} = CC^T\,. \tag{33}$$

There is, of course, a family of $C$ matrices that satisfies Eq. (33). In this paper, we employ the Cholesky decomposition (Cox and Hammarling 1990), allowing us to generate correlated noise variables producing "prescribed" (empirically observed) correlations across time series. The Cholesky decomposition is advantageous as it provides a lower triangle matrix, which minimizes the number of non-zero elements and is computationally efficient to invert. Assuming $\rho^{\psi}$ is full rank, $C$ is invertible (which is the case if each time series considered is unique).

Given complete information on $\psi$, we are also able to obtain complete information on $\psi'$. Let us now assume that at some time $t_l$ a time series is masked. Without loss of

generality, we take that time series to be last in the index set. As we will see next, it is useful to solve Eq. (32) for $\psi'$,

$$\sum_j (C^{-1})_{kj}\psi_j = \psi'_k. \tag{34}$$

If we have $n$ time series, the last row equation is,

$$\sum_{j=1}^{n} (C^{-1})_{n,j}\psi_j = \psi'_n. \tag{35}$$

Solving for $\psi_n$, we obtain

$$\psi_n = \frac{\psi'_n - \sum_{j=1}^{n-1}(C^{-1})_{n,j}\psi_j}{(C^{-1})_{n,n}}. \tag{36}$$

Because $\psi_n$ is masked at time $t_l$, we can not determine $\psi'_n$. While sufficient information for an exact prediction of $\psi_n$ is unavailable, we can still produce an educated guess for $\psi_n$ using Eq. (36) by generating a realization of $\psi'_n$. Because in this equation the left hand side is a normal random variable with unit variance, the sum of the variances for the variables on the right hand side of the equation must sum to unity. Due to this requirement as $|(C^{-1})_{n,n}|$ becomes large, $\psi'_n$'s contributions to the prediction become smaller relative to the summation. This reduces the uncertainty in $\psi_n$. In the limiting case of $|(C^{-1})_{n,n}| \to \infty$, our prediction method becomes deterministic.

Given a sufficient initial period of complete information, we can estimate the SDE equation parameter space using our described MLE method (Franco 2003; Tang and Chen 2009; Johnson and Wichern 2002; Rayner 1985) assuming the parameters are time invariant. We then use Eq. (2) to convert our sets of time series into series of normal random variables, which we can then use to determine $\rho$ and $C$. In order to predict information we are lacking in Eq. (36), we generate a realization of $\psi'_n$. Once we have estimated $\psi_n(t_i)$, we can get the corresponding $S_n(t_i)$ from Eq. (1) and then average our estimation over many realizations.

## Results

To test the models posed, we have two testing environments. The first is a synthetic (or simulated) environment where the underlying process is known precisely (GBM or GOU). Hence, parameters and correlations are easily controlled in order to test the feasibility robustness of the prediction scheme. In this environment, we are able to use multiple realizations to average results and obtain important observations that can not be seen in one realization.

The second environment is application on a cryptocurrency market and social media containing several major cryptocurrencies including Bitcoin, Ethereum, Litecoin, Monero and XRP. (Data was provided as part of the DARPA SocialSim Project (DARPA).) The trends of these time series can be seen in Fig. 1. While cryptocurrencies have no restrictions on when a trade can be placed, we find it useful to define two daily metrics to use as a daily time series. We define the closing price of a cryptocurrency as the market price at 22:00 GMT each day and the daily volume as the amount of currency traded in terms of USD in the 24 hours prior to 22:00 GMT. Hence, we use $\Delta t = 1$ day in our numerical scheme presented in this paper.

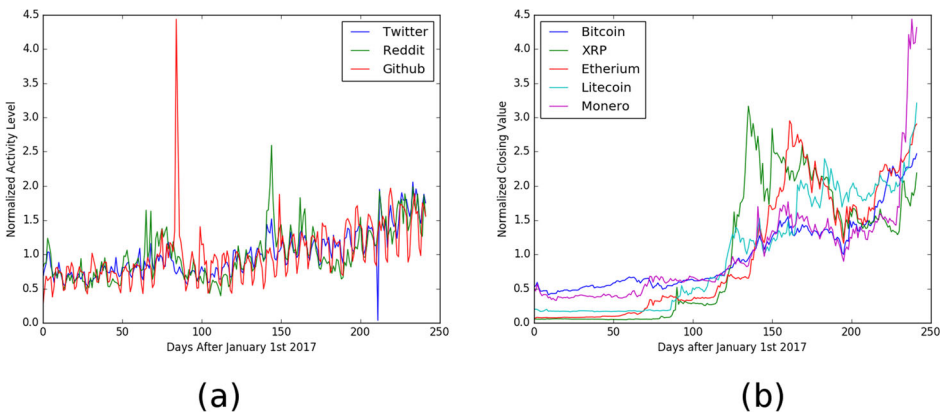


**Fig. 1** Time series details for (**a**) activity levels on social media and (**b**) market price at 14:00 GMT. Here all time series have been normalized such that their mean is unity

Our social media data sets contain tweets from Twitter, comments from Reddit, and events from GitHub that all pertain to *any cryptocurrency, including minor, less traded cryptocurrencies.* While the minor coins are included in the data set, they have relatively low number of tweets, comments, or events. We define daily statistics, even though this approach is generalizable to hour or even minute resolution. If a given time step returns zero due to the increased resolution, this becomes an absorbing state for both stochastic equations and parameters can not be estimated. For these reasons, we only use the total number of daily tweets/comments/events for each media to compose three time series to add to our data set. While increased resolution is possible, a sufficient density of activity at the increased resolution is required from a technical standpoint.

While we use GBM and GOU as our governing equations, we use them only as approximations. Indeed, the underlying equations which best describe these time series are likely more complex than what we have posed here. However, at the lowest order of improvement, any unaccounted effects for interactions could be simply cast as a more complex noise term. If these interactions persist between sets of time series, our correlated noise will allow these interactions to take place, even though we have not explicitly engineered these interactions in our underlying equation.

The duration of these data sets spans from January 1st 2017 to August 31st 2017. An important assumption was that the parameters can be approximated as time independent. Indeed, the parameters do have some time variance over long periods (Bassler et al. 2007). For this reason, we limit the training window for estimating correlations and parameters to the previous 120 days. Examples of the parameter estimation over time can be seen in Appendix 4.

In both synthetic and empirical data environments, we tested how well the method preforms at predicting market value and social media activity levels. We measured this in three ways. For the first, we used a sliding window approach where a single day at a time will be predicted using the previous 120 days. For the second method, we measured the accuracy for increased prediction duration. Here, we selected evenly spaced initial times and predicted the time series 14 days into the future. This will test how much uncertainty builds up as we extend our prediction further in time. Finally, we test our method's accuracy at predicting increases or decreases of the following day without considering the resulting values.

In order to test our predictions, we employ the Mean Absolute Percentage Error (MAPE) (Reddy and Clinton 2016; Maruddani and Trimono 2018),

$$\epsilon = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{S(t_i) - \hat{S}(t_i)}{\hat{S}(t_i)}\right|, \tag{37}$$

where $S(t_i)$ is the forecast value and $\hat{S}(t_i)$ is the actual (ground-truth) value. This provides an estimate for how much the prediction deviates from the ground truth.

#### Synthetic data from GBM and GOU processes

A large part of our results relies on predicting a given realization. In principle, all realizations are technically valid, even though some are increasingly unlikely. We show this in Fig. 2, where a particular realization can deviate outside the 1 $\sigma$ (STD) bounds produced from parameter estimation. Parameters from MLE and correlation estimation can be seen in Table 4 (Appendix 1). The ground truth can also deviate outside the 1 $\sigma$ (STD) bounds for the prediction using the correlated time series even for strong correlations. Because of the nature of these stochastic equations and that we only have one realization to work with in our data set, we expect some significant variations in the accuracy of our predictions.

To properly test our model, we use many realizations to show its robustness under ideal conditions. In this case, we generate data using known equations with set parameters. While our algorithm will estimate the parameters, the equation used will be known. This produces an idealization over our cryptocurrency data, where we can not say for certain what their underlying equation actually is. We also define the correlations and parameters to be constant in time, another idealization. We choose our parameters to produce significant variance in signal, but so that we do not lose accuracy due to our time step size. For our correlation matrix, we define all off-diagonal elements to be equal. We also explore the case where we add an exception to our correlation definition to observe its impact.

Figure 3 shows the accuracy of our prediction methods using various combinations of inputs and compares their result to a prediction assuming the time series is uncorrelated with the other time series. Figure 3a explores predicting the sign of the slope of the data. This shows a fairly linear trend towards 100% accuracy in predicting the sign of the slope with only small increases in correlation yielding significant improvements over the prediction without correlation.

Figure 3b shows the MAPE value when a single data point is masked. As expected, our prediction greatly exceeds the prediction without correlation. A more important observation is a small decrease as the number of time series increase. One would expect as the number of time series increase that the unaccounted noise in the prediction method would decrease, however we see a less significant decrease in MAPE value than expected. This may be a result of the homogeneity of our definition of correlations. To minimize the amount of unaccounted noise, each time series must be highly correlated with the time series intended to be predicted, but relatively uncorrelated with other time series used in the prediction method. In this way, each time series provides information that is as unique as possible. This of course produces some difficulties as, if one chooses to try to predict a different time series, all other time series are uncorrelated and do not contribute much to the prediction method.

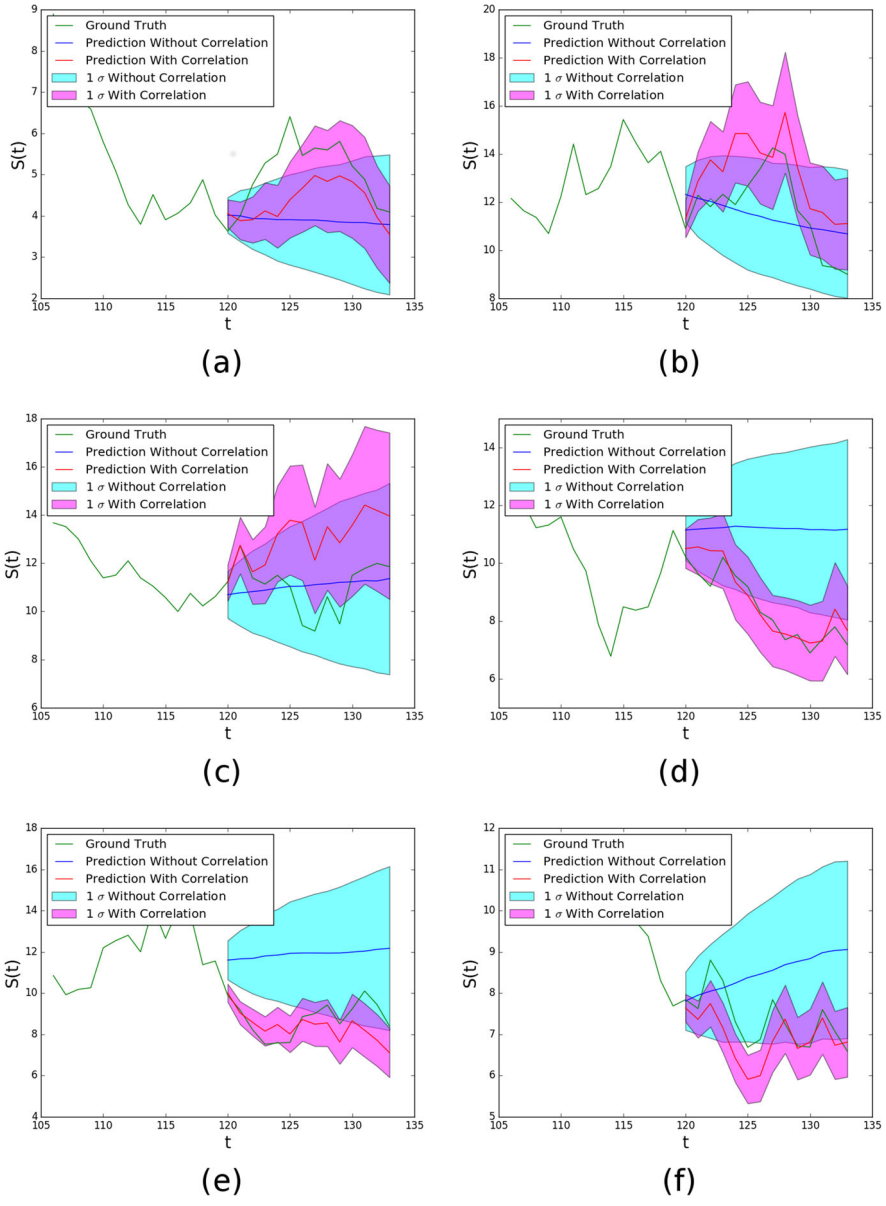


**Fig. 2** Predictions of simulated data using various models and correlations by masking the last 14 data points. Our prediction with correlation uses 11 other correlated time series, and we made a prediction assuming no correlation with those time series. A training window of 120 time steps was used in each case. (**a**, **c**, **e**) use GBM and (**b**, **d**, **f**) use GOU. (**a**, **b**) use a correlation matrix with all off-diagonal elements of the correlation matrix equal to .4 with (**c**, **d**) using .6 and (**e**, **f**) using .8. Estimated parameters and correlations for these examples can be seen in Table 4 (Appendix 4)

Figure 3c masks multiple time steps and the prediction method operates recursively to predict each subsequent time step using the previous time step's prediction starting with the last known value of the time series. As expected, the MAPE value increases as the number of time steps predicted increases as small inaccuracies in each prediction begin to accumulate. Figure 3d shows the decrease in MAPE value when an additional time series is added with similar correlations, but with a significantly increased correlation to the predicted time series. This increase in performance is significant, but not substantial as it

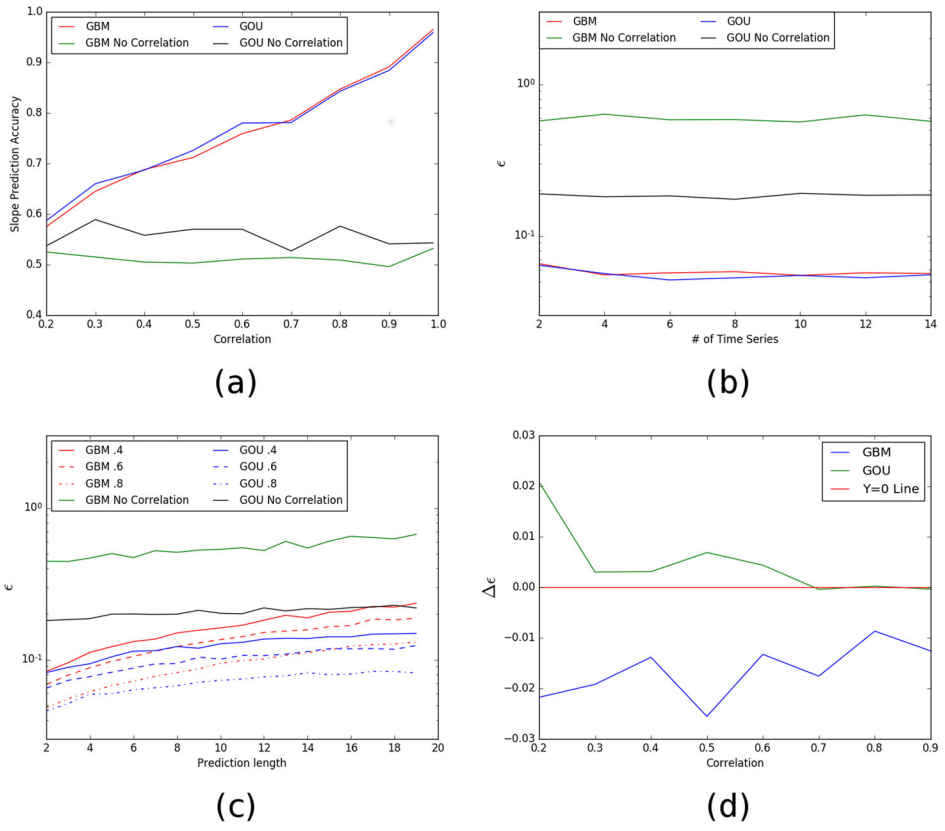


**Fig. 3** **a** Shows the accuracy in predicting whether the time series will increase or decrease for a training period of 120 time steps before the prediction. This is compared to a prediction without using correlated time series. **b** Here, we show the MAPE ($\epsilon$) for predicting the value of the time series one time step into the future for 90 and 120 time step training periods. After a small decrease in the MAPE value, there is not a significant change in MAPE as the number of time series increase. **c** shows the MAPE for predicting a variety of time steps into the future using correlations of 0.4, 0.6, 0.8. **d** shows the change in MAPE value ($\Delta\epsilon$) when a time series is added with improved correlation is added. This is done by setting the correlation of the added time series and the predicted time series to the square root of the original correlation

only provides a small increase. This indicates that adding additional time series in hopes of a better predictions is more complex and non-trivial than finding a well correlated time series.

As can be seen from the results of our simulation, our prediction algorithm performs well when the presented idealizations are in place. As we move to studying the cryptocurrency data, we expect our accuracy to decrease. In addition because our data is a single realization, we expect our prediction method to have various examples of poor performance.

#### Empirical data: forecasting cryptocurrency rates and social media activities

As we noted earlier, the dynamics of cryptocurrencies shares similarities to those of stocks. Hence, for cryptocurrency prices we employ the standard GBM. The tendency to “drift" in this process correspond to the expected rate of return, while the stochastic effects and fluctuations are captured by the appropriate noise term. Because of the drift, the price need not return to a baseline level.

On the other hand, to capture the dynamics of social media activities and trading volumes of cryptocurrencies, we need to make some observations and further hypotheses. For example, in the case of Twitter, a large volume of tweets can be produced in response to an event. After sufficient time, the activity around that event decays and the environment on the whole returns to a base-level activity until another event triggers an increase. The same observations can be made of the daily volumes traded. Hence, we use the GOU process for the social media activity and volume traded as when a fluctuation happens, they typically cannot be sustained for extended periods of time, and in time they return to some long-run mean. The geometric nature of the process ensures that the corresponding stochastic variable is non-negative, and fluctuations in the OU process have a tendency to revert to the mean in the long run.

For our prediction method we utilize only the previous 120 days before starting our prediction. In this way, we attempt to limit the effects of our approximation of time invariant parameters and correlations. This also includes a sufficient number of data points to have confidence in our parameter and correlation estimation. Other training periods were examined, but our choice performs best in most cases. Appendix 5 shows the analysis of the noise measured for each time series. Here, we see some deviations from a normal distribution and some temporal correlations which suggests the equations are not strictly Markovian.

Figure 4 shows two methods for predicting the masked time step. The current time prediction is fairly straight forward with concurrent information being used to generate predictions for a time series. However, it is not very useful to predict the value of Bitcoin after already knowing the value of other time series such as the closing price of other cryptocurrencies and the volume of Bitcoin traded. Instead, we give as inputs a similar time series and go forward with a similar prediction as the current time prediction, but with one important modification. We include in the data set a copy of the target time series, but time-shifted in such a way that the value of each time series at time $t_i$ is now coinciding with the copied time series' value at $t_{i+1}$. We now predict this copied time series which uses the known values of the various time series at time $t_i$ to predict the value of the copied and shifted time series at time $t_{i+1}$. This essentially allows us to make a future time prediction without any knowledge of other time series at $t_{i+1}$. This obviously holds only if the time-shifted copy is still correlated with other time series, which is true for some cases (See Appendix 5).

Table 1 shows the average MAPE over time for each subfigure in Fig. 4. Interestingly, the future time prediction produces slightly better MAPE values in comparison to the current time prediction in some cases. Social Media predictions perform significantly better than the prediction without correlation on average. This performance is less so for predicting cryptocurrency prices, but has overall low MAPE values. This result is slightly misleading as will discussed shortly when we examine predictions over multiple time steps.

We now discuss our method's performance for predicting multiple consecutive days. In this case, we will always use the current time prediction method. This methodology works best for examining a system where a media's history is available, but current and recent activity is unavailable. Being able to predict spikes in activity in this hidden media has many possible applications for information gathering. In the examples shown, our predictions use the same training period of 120 days and then predict the next 14 days using the concurrent time series. To reduce variations causing poor results, we examine three

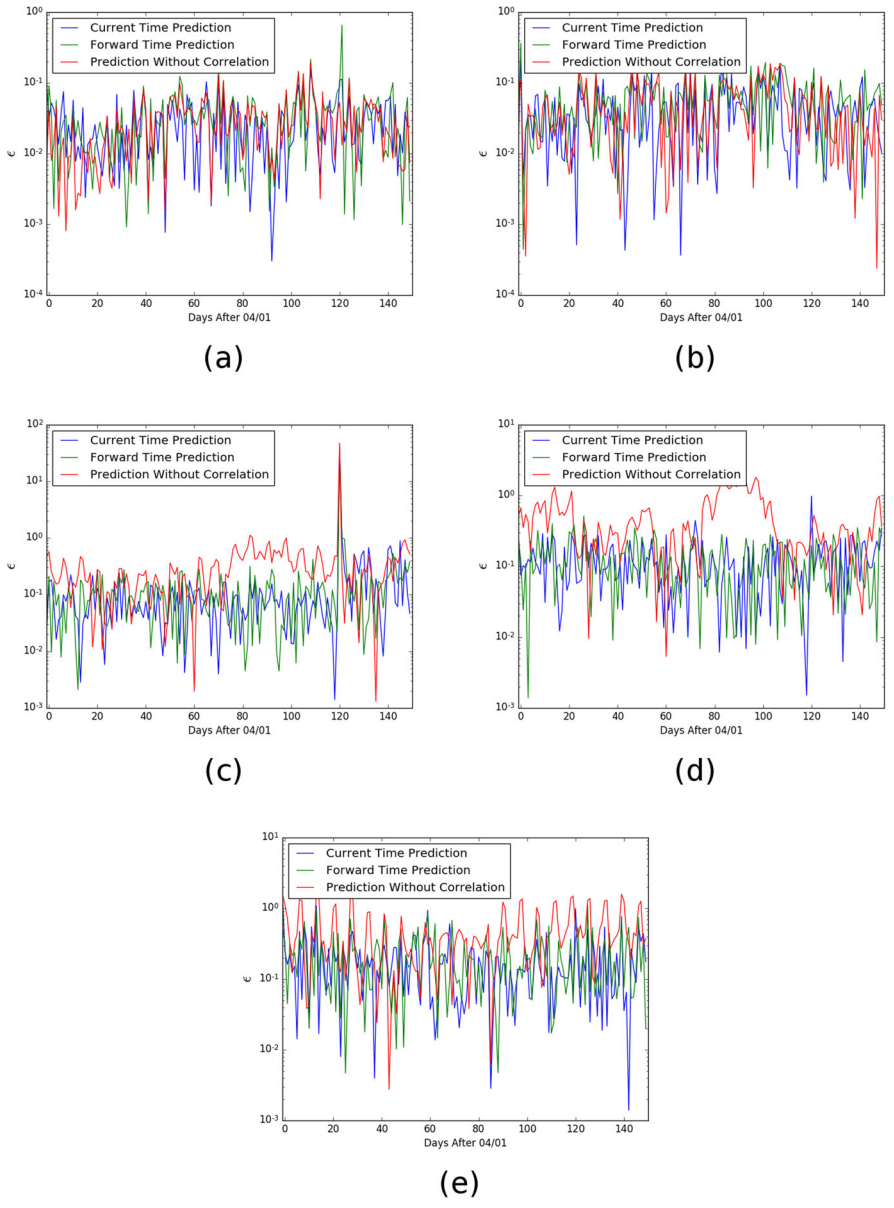


**Fig. 4** Each day in the data set is predicted after collecting a sufficient number of days for prediction. We measure the accuracy using the MAPE ($\epsilon$). **a** Shows the predictions for Twitter, **b** for Reddit, **c** for GitHub, **d** for Bitcoin, and **e** for Ethereum. Both methods for predictions are compared against the prediction without correlation and the average MAPE value can be seen in Table 1

**Table 1** Average MAPE of each time series from May 31st 2017 - Aug 30th 2017

| | Twitter | Reddit | GitHub | BTC | ETH |
|---|---|---|---|---|---|
| Current Time Prediction | 0.369 | 0.131 | 0.205 | 0.0308 | 0.0481 |
| Future Time Prediction | 0.336 | 0.129 | 0.231 | 0.0400 | 0.0654 |
| Linear Prediction | 0.616 | 0.495 | 0.488 | 0.0324 | 0.0544 |

different periods. We evenly space our predictions to ensure sufficient independence in our measurement accuracy.

Figures 5 and 6 show a suite of predictions for the various time series. The corresponding MAPE values for each of the predictions can be seen in Table 2. Parameters from MLE and correlation estimation can be seen in Table 5 (Appendix 1). As anticipated, our prediction method has difficulty matching the accuracy seen in our simulated data. We believe a likely cause is uncertainty in our measurements of the parameters, particularly the unaccounted effect of time variance. While this is present in our previous results, the recursive predictions accumulate uncertainty and amplify the uncertainty of our prediction.

Many of our prediction do follow the ground truth well. In cases that we do not have favorable MAPE values, our prediction still shows many of the characteristics present in the ground truth, such as coinciding peaks. Indeed, some of our predictions appear to be slightly translated compared to the ground truth. While social media predictions show fair results, the cryptocurrency predictions struggle significantly more. While the MAPE values are low for our cryptocurrency results compared to the social media prediction, this is mostly due to low volatility. With smaller volatility, the size of the noise

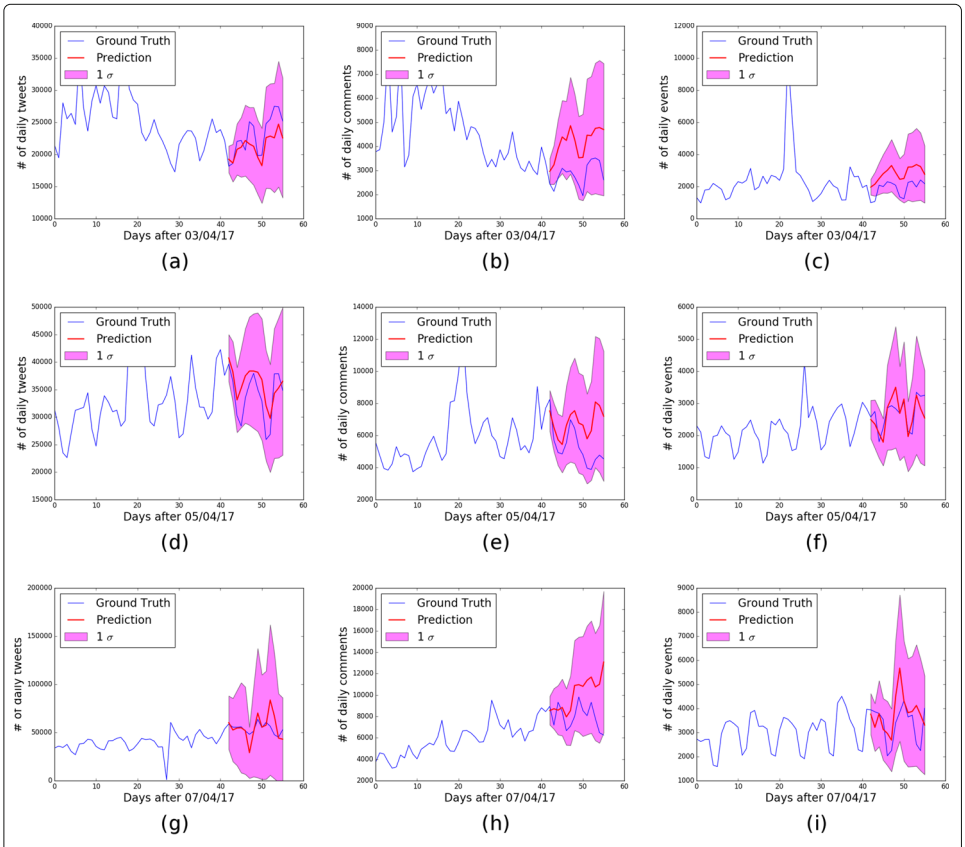


**Fig. 5** Each subfigure shows 14 days of prediction compared to the ground truth with the previous 42 days of ground truth as reference. (**a**, **b**, **c**) shows the prediction at the end of April 2017. (**d**, **e**, **f**) shows the prediction at the end of June 2017. (**g**, **h**, **i**) shows the predictions at the end of August 2017. (**a**, **d**, **g**) shows the predictions for Twitter at the corresponding time ranges, (**b**, **e**, **h**) for Reddit, and (**c**, **f**, **i**) for GitHub. Mape values for each subfigure can be seen in Table 2. Estimated parameters and correlations for these examples can be seen in Table 5 (Appendix 4)

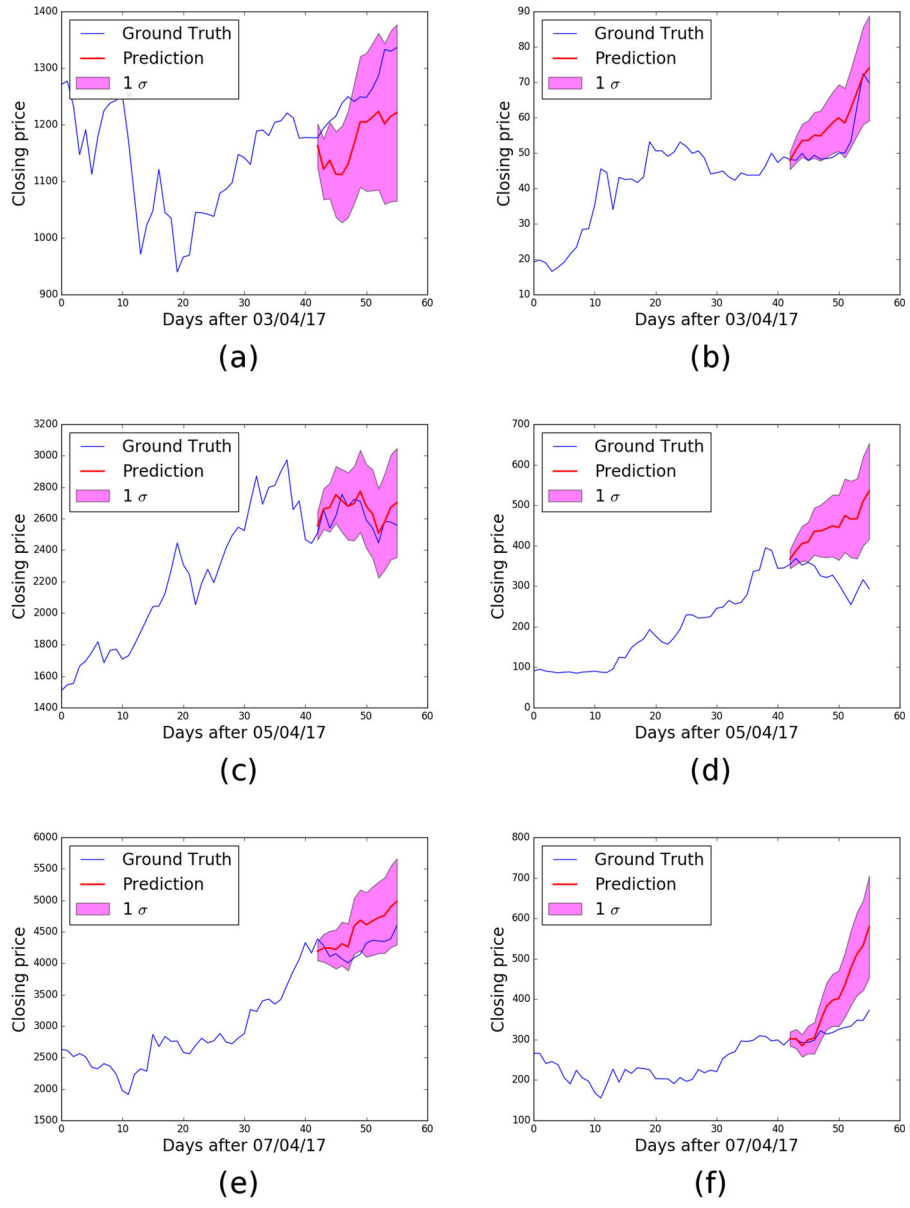


**Fig. 6** Each subfigure shows 14 days of prediction compared to the ground truth with the previous 14 days of ground truth as reference. (**a**, **b**) shows the prediction at the end of April 2017. (**c**, **d**) shows the prediction at the end of June 2017. (**e**, **f**) shows the predictions at the end of August 2017. (**a**, **c**, **e**) shows the predictions for Bitcoin at the corresponding time ranges and (**b**, **d**, **f**) for Ethereum. Mape values for each subfigure can be seen in Table 2. Estimated parameters and correlations for these examples can be seen in Table 5 (Appendix 4)

**Table 2** Average MAPE of time series for predictions during the months of April, June, and August in 2017

| | Twitter | Reddit | GitHub | Bitcoin | Ethereum |
|---|---|---|---|---|---|
| April | 0.0886 | 0.472 | 0.580 | 0.0647 | 0.107 |
| June | 0.0991 | 0.340 | 0.110 | 0.0247 | 0.406 |
| August | 0.1287 | 0.311 | 0.235 | 0.0749 | 0.226 |

also decreases. Because MAPE is a percentage metric, MAPE will typically decrease with smaller volatility. In addition, GBM is more sensitive to larger prediction periods compared to GOU, which makes the cryptocurrency predictions visually less accurate. In this way, the underlying equations for how cryptocurrency markets behave may require modification for more accurate long term predictions.

While cryptocurrencies may not be predicted well over large time ranges, there are more applications in predicting behavior in the future. Using our method for computing future time steps as mentioned above, we look to examine how well our prediction method does at predicting whether markets will increase or decrease into the next day.

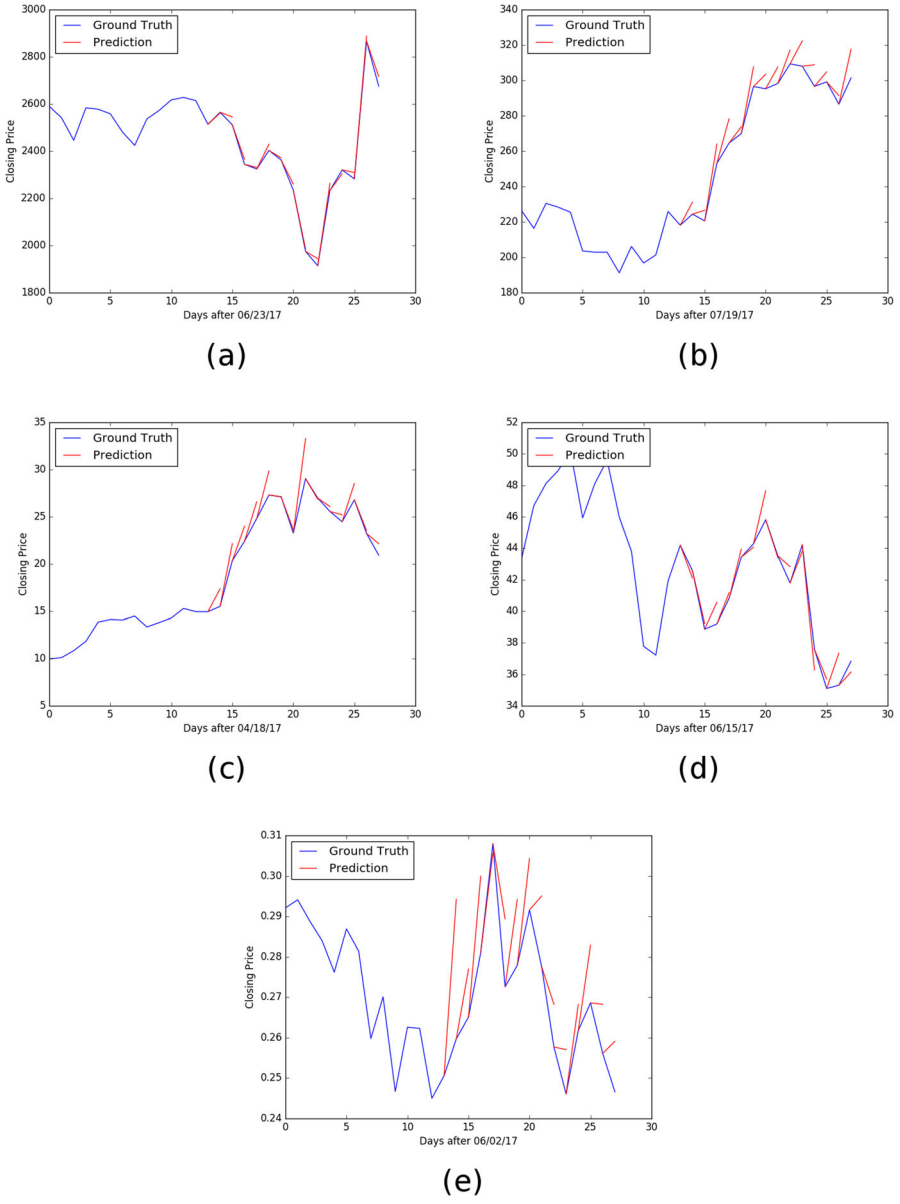


**Fig. 7** The ground truth for (**a**) Bitcoin, (**b**) Ethereum, (**c**) Litecoin, (**d**) Monero, and (**e**) XRP is compared against the prediction. Each prediction is based off of the previous known data point and each individual line is a separate prediction. This best shows for each prediction whether the closing price is predicted to increase or decrease. The number of positive identifications can be seen in Table 3

This has obvious applications in short term trading and investments. Figure 7 shows how well our prediction method performs at various time regions. Again, we choose time regions as evenly spaced as possible to provide as independent results as possible. We also choose the most obvious cases where our prediction method would be applicable; regions of high variations, peaks, and dips. Regions with many linear portions are biased due to the accuracy coming more from parameter estimation than from noise analysis.

Table 3 shows the number of days out of 14 that we correctly predict an increase or decrease in the following day. To compare, we also measure the number of instances where the market increased and then increased again on the following day as well as sequential decreases. If one were to randomly guess changes, we expect them to be correct 7 out of the 14 times. As can be seen, our prediction method performs exceptionally well with many perfect or close to perfect predictions and always outperforming our repeat and random measure.

## Future extensions

We have made the assumption through that the correlations and parameters are not time varying. A natural extension of this work is removing this assumption and creating models for behavior of the time dependence. There are many technical hurdles to overcome such as the modifications required to the underlying equations and preventing over fitting of the model. In addition, the underlying equations may be more complex than that presented here. As the approximation to the actual underlying equations becomes better, the accuracy of this method will also increase.

In other ways, our model is very generalizable. Our framework makes little assumption on the type of information given other than the model used for it. For example, it has been shown that Twitter “mood" or online sentiments are strongly correlated with stock market or cryptocurrency returns (Kim et al. 2016; Bollen et al. 2011). Hence, one could perform sentiment analysis on social media activities to separate events into two basic categories (bullish or bearish comments with respect to cryptocurrencies). It is likely that after exploiting this separation, the predictive power of our framework would improve. In addition, one can make a prediction for a cryptocurrency or social media using an independent method. This could then be added as a time series for our prediction method. If the independent prediction method is accurate, there should be a high-level of correlation between the prediction and the ground truth. In this way, the other time series can be used to fine tune this independent prediction method and increase the overall accuracy.

There is one major drawback to the independent prediction method. It would need to be applied to already known data so that a correlation can be measured between the prediction and ground truth. If the prediction is only one time step into the future, the prediction

**Table 3** The prediction shows the number days where whether the market will increase or decrease is correctly predicted over a 14 day window

| | Bitcoin | Ethereum | Litecoin | Monero | XRP |
|---|---|---|---|---|---|
| Prediction | 14 | 10 | 14 | 14 | 12 |
| Repeat | 6 | 5 | 8 | 8 | 9 |

The Repeat measure shows how well a prediction would work if it simply assumes the behavior of the previous step will hold for the next step (i.e. an increase in the market on the previous time step will indicate and increase in the next time step.)

method can be used recursively through the data set in order to achieve this. The difficulty arises when one intends to predict multiple days. As the duration of prediction increases, the accuracy of that prediction also decreases, which would in turn cause the correlation to decrease. This violates a key assumption of time invariant correlation, and if used, the method would require a derivation without the constraint of time invariant correlations.

## Conclusion

In this work, we proposed a method for predicting a time series exploiting stochastic correlations with other time series. In particular, we showed that the extracted "empirical" noise of each time series is correlated and provides some degree of predictability. This method is both robust in simulation and easily generalizable. Our framework can be applied to noise-coupled SDEs where parameter estimation is attainable (either through MLE or approximate empirical estimates for the means, covariances, and correlation functions where applicable). Its performance is dependent on using the correct underlying equation for each time series (i.e., a good hypothesis for the general forms of equations). For our synthetic data, where we knew precisely the forms of equations that generated the data, our method was rather accurate (as the only source of errors is coming from the parameter estimation). For the empirical data, our equations are only plausible hypothesis to begin with, which gives rise to more significant deviations between prediction and the ground truth. Specifically, more complex nonlinearities may be relevant or off-diagonal couplings may have to be included among the deterministic parts of the various stochastic processes (i.e., coupling may not be merely through the correlated noise). Also, our assumption of constant parameters, such as for the volatility, can break down, and one would need a more refined scheme for time-varying parameters.

Testing this prediction method on empirical data sets of social media relating to cryptocurrency and measures of the stochastic properties of those same cryptocurrencies shows reasonable results. Testing our method would have also benefited from larger data sets (longer time series), however, in this current project we only considered what we had access to as part of the DARPA SocialSim Project (DARPA). We find that the prediction algorithm is fairly consistent in predicting one day into the future. We also assert that our prediction method can be extended to predicting data on the next day assuming that time series are correlated through temporal shifts.

When we predict more than one day into the future, we begin to experience difficulties arising from growing uncertainty. We perform longer predictions by masking a 14 day period and using the other time series during the masked time period to estimate the daily changes. This produces a significant amount of uncertainty associated with recursive predictions, which in general causes uncertainty to accumulate. The cryptocurrency markets are particularly susceptible to this accumulation of uncertainty (Wu et al. 2018; Fry 2018). In spite of this, our prediction has many of the same characteristics as the ground truth characteristics such as coinciding peaks and troughs. This has particular applications in areas where a time series can not be directly observed in real time, but its history can be. This method can allow for an educated guess, especially for peaks, for the unobservable time series.

We showed it can be quite difficult for our method to predict the exact value a cryptocurrency market will assume on the following day. While this is an obvious shortcoming, the predictions often correctly predict whether the cryptocurrency markets will increase or decrease. Our method's performance in predicting increases and decreases was sufficiently better than a random prediction and our baseline prediction which assumes trends are likely to continue. These results have significant implications for trading and investing in cryptocurrencies. This method gives investors a reliable and accurate indicator when a currency is to increase or decrease.

## Appendix 1: Maximum likelihood estimation for single-variable GBM and GOU processes

Here, we provide some details for the maximum likelihood estimation (MLE) method for recovering the parameters of Eqs. (14) and (17). The probability of the realization of the time series $\{S(t_i)\}$ with initial condition $S(t_0)$ is Hurn et al. (2003),

$$\prod_{i=1}^{N} P[\, S(t_i)|S(t_{i-1}), ..., S(t_0); \overrightarrow{\beta}\,]\,. \tag{38}$$

where $\overrightarrow{\beta}$ is the parameter space. Because both GBM and GOU equations are Markovian, the transition probability distributions simplify to,

$$P[\, S(t_i)|S(t_{i-1}), ..., S(t_0); \overrightarrow{\beta}\,] = P[\, S(t_i)|S(t_{i-1}); \overrightarrow{\beta}\,]\,. \tag{39}$$

The log likelihood function is then,

$$L(\overrightarrow{\beta}) = \sum_{i=1}^{N} \ln\left(P[\, S(t_i)|S(t_{i-1}); \overrightarrow{\beta}\,]\right). \tag{40}$$

In order to recover the optimal model parameters $\overrightarrow{\beta}$, we determine the parameter set $\hat{\overrightarrow{\beta}}$ that maximizes $L(\overrightarrow{\beta})$.

### 1.1 MLE for GBM

For the GBM, we start with the transition probability distribution (Hurn et al. 2003) Eq. (14),

$$P[\, S(t_i)|S(t_{i-1}); \bar{\mu}, \sigma\,] = \frac{1}{S(t_i)\sigma\sqrt{2\pi\,\Delta t}} \exp\left(-\frac{\left(\ln\left(\frac{S(t_i)}{S(t_{i-1})}\right) - \bar{\mu}\,\Delta t\right)^2}{2\sigma^2\Delta t}\right). \tag{41}$$

Substituting our definition of $U_i = \ln S(t_i)$, we recover $\bar{\mu}$,

$$\left.\frac{\partial L}{\partial \bar{\mu}}\right|_{\hat{\bar{\mu}}} = 0 = \sum_{i=1}^{N} \frac{\partial}{\partial \hat{\bar{\mu}}} \frac{\left(U_i - U_{i-1} - \hat{\bar{\mu}}\Delta t\right)^2}{2\hat{\sigma}^2 \Delta t}, \tag{42}$$

$$0 = \sum_{i=1}^{N} \left(U_i - U_{i-1} - \hat{\bar{\mu}}\Delta t\right), \tag{43}$$

$$\hat{\bar{\mu}} = \frac{1}{N\Delta t} \sum_{i=1}^{N} \left(U_i - U_{i-1}\right). \tag{44}$$

Next, we recover $\sigma$,

$$\left.\frac{\partial L}{\partial \sigma}\right|_{\hat{\sigma}} = 0 = \sum_{i=1}^{N} \frac{\partial}{\partial \hat{\sigma}} \left( \ln\left(\hat{\sigma}\sqrt{2\pi\Delta t}\right) + \frac{\left(U_i - U_{i-1} - \hat{\bar{\mu}}\Delta t\right)^2}{2\hat{\sigma}^2 \Delta t} \right), \tag{45}$$

$$0 = \sum_{i=1}^{N} \frac{1}{\hat{\sigma}} - \frac{\left(U_i - U_{i-1} - \hat{\bar{\mu}}\Delta t\right)^2}{\hat{\sigma}^3 \Delta t}, \tag{46}$$

$$\hat{\sigma}^2 = \frac{1}{N\Delta t} \sum_{i=1}^{N} \left(U_i - U_{i-1} - \hat{\bar{\mu}}\Delta t\right)^2. \tag{47}$$

### 1.2 MLE for the GOU process

Obtaining the MLE parameter estimates from the transition probability distribution for the GOU process is more involved (Mejía Vega 2018; Franco 2003; Tang and Chen 2009),

$$P(S(t_i)|S(t_{i-1}); \bar{\theta}, \sigma, \kappa) = \frac{1}{S(t_i)} \left(2\pi \frac{\sigma^2}{2\kappa}\left(1 - \mathrm{e}^{-2\kappa\Delta t}\right)\right)^{-\frac{1}{2}} \times \exp\left(-\frac{\left(\ln S(t_i) - \bar{\theta} - (\ln S(t_{i-1}) - \bar{\theta})\mathrm{e}^{-\kappa\Delta t}\right)^2}{2\frac{\sigma^2}{2\kappa}\left(1 - \mathrm{e}^{-2\kappa\Delta t}\right)}\right). \tag{48}$$

For convenience, we make the following definitions,

$$x = \mathrm{e}^{-\kappa\Delta t}, \tag{49}$$

$$z = \hat{\bar{\theta}}(1 - x), \tag{50}$$

$$y^2 = \frac{\sigma^2}{2\kappa}(1 - x^2). \tag{51}$$

With these substitutions we can identify that the transition probability density is of similar form to the GBM process. We then recover the parameters $\hat{z}$ and $\hat{y}^2$ in the same manner as $\hat{\bar{\mu}}$ and $\hat{\sigma}^2$ for GBM respectively,

$$P(U_i|U_{i-1};x,y,z) = \frac{1}{y\sqrt{2\pi}}\exp\left(-\frac{(U_i - U_{i-1}x - z)^2}{2y^2}\right), \tag{52}$$

$$\left.\frac{\partial L}{\partial z}\right|_{\hat{z}} = 0 = -\sum_{i=1}^{N}\frac{\partial}{\partial \hat{z}}\frac{\left(U_i - U_{i-1}\hat{x} - \hat{z}\right)^2}{2\hat{y}^2}, \tag{53}$$

$$0 = \sum_{i=1}^{N}\left(U_i - U_{i-1}\hat{x} - \hat{z}\right), \tag{54}$$

$$\hat{z} = \frac{1}{N}\sum_{i=1}^{N}\left(U_i - U_{i-1}\hat{x}\right), \tag{55}$$

$$\hat{\hat{\theta}} = \frac{1}{N(1-\mathrm{e}^{-\hat{\kappa}\,\Delta t})}\sum_{i=1}^{N}\left(\ln S(t_i) - \ln(S(t_{i-1}))\mathrm{e}^{-\hat{\kappa}\,\Delta t}\right), \tag{56}$$

$$\left.\frac{\partial L}{\partial y}\right|_{\hat{y}} = 0 = \sum_{i=1}^{N}\frac{\partial}{\partial \hat{y}}\left(\ln\left(\hat{y}\sqrt{2\pi}\right) + \frac{\left(U_i - U_{i-1}\hat{x} - \hat{z}\right)^2}{2\hat{y}^2}\right), \tag{57}$$

$$0 = \sum_{i=1}^{N}\frac{1}{\hat{y}} - \frac{\left(U_i - U_{i-1}\hat{x} - \hat{z}\right)^2}{\hat{y}^3} \tag{58}$$

$$\hat{y}^2 = \frac{1}{N}\sum_{i=1}^{N}\left(U_i - U_{i-1}\hat{x} - \hat{z}\right)^2, \tag{59}$$

$$\hat{\sigma}^2 = \frac{2\hat{\kappa}}{N(1-\mathrm{e}^{-2\hat{\kappa}\,\Delta t})}\sum_{i=1}^{N}\left(\ln S(t_i) - \ln(S(t_{i-1}))\mathrm{e}^{-\hat{\kappa}\,\Delta t} - \hat{\hat{\theta}}(1-\mathrm{e}^{-\hat{\kappa}\,\Delta t})\right)^2. \tag{60}$$

Further,

$$\left.\frac{\partial L}{\partial x}\right|_{\hat{x}} = 0 = \frac{\partial}{\partial \hat{x}}\sum_{i=1}^{N}\left(-\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\hat{y}^2) - \frac{(U_i - U_{i-1}\hat{x} - \hat{z})^2}{2\hat{y}^2}\right) \tag{61}$$

$$0 = -\frac{1}{2\hat{y}^2}\frac{\partial}{\partial \hat{x}}\sum_{i=1}^{N}\left(U_i - U_{i-1}\hat{x} - \hat{z}\right)^2 = \frac{1}{\hat{y}^2}\sum_{i=1}^{N}\left(U_i - U_{i-1}\hat{x} - \hat{z}\right)U_{i-1} \tag{62}$$

Finally, using the expression for $\hat{z}$ earlier from Eq. (55), we obtain

$$\hat{x} = \frac{\left(\frac{1}{N}\sum_{i=1}^{N}U_iU_{i-1}\right) - \left(\frac{1}{N}\sum_{i=1}^{N}U_{i-1}\right)\left(\frac{1}{N}\sum_{j=1}^{N}U_j\right)}{\left(\frac{1}{N}\sum_{i=1}^{N}U_{i-1}^2\right) - \left(\frac{1}{N}\sum_{i=1}^{N}U_{i-1}\right)\left(\frac{1}{N}\sum_{j=1}^{N}U_{j-1}\right)} \tag{63}$$

$$= \frac{\frac{1}{N}\sum_{i=1}^{N}\left(U_i - \frac{1}{N}\sum_{j=1}^{N}U_j\right)\left(U_{i-1} - \frac{1}{N}\sum_{j=1}^{N}U_{j-1}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(U_{i-1} - \frac{1}{N}\sum_{j=1}^{N}U_{j-1}\right)^2}. \tag{64}$$

Thus the parameter estimation for $\hat{x} = \mathrm{e}^{-\hat{\kappa}\,\Delta t}$ is precisely the normalized empirical auto-correlation of $U(t_i)$ with time difference $\Delta t = t_i - t_{i-1}$ (Gardiner 1985; Tang and Chen 2009).

## Appendix 2: Covariances for the multivariate GBM-GOU system

We will use the indices of the (logarithmic) variables $U_j(t) = \ln S_j(t)$ to distinguish between the GBM or GOU processes, such that of a total of $d = n_{\mathrm{GBM}} + n_{\mathrm{GOU}}$ stochastic variables, the first $n_{\mathrm{GBM}}$ are GBM variables, while the remaining $n_{\mathrm{GOU}}$ are GOU variables. Then for $j = 1, \ldots, n_{\mathrm{GBM}}$ ($j \in$ GBM for short),

$$dU_j(t) = \bar{\mu}_j dt + \sigma_j dW_j(t) \ , \tag{65}$$

which can be formally integrated using stochastic calculus:

$$U_j(t) = U_j(0) + \bar{\mu}_j t + \sigma_j \int_0^t dW_j(t'), \tag{66}$$

Similarly, $j = n_{GBM} + 1, \ldots, d$ ($j \in$ GOU for short),

$$dU_j(t) = \kappa_j(\bar{\theta}_j - U_j(t))dt + \sigma_j dW_j(t) \ . \tag{67}$$

and by stochastic integration,

$$U_j(t) = \bar{\theta}_j + (U_j(0) - \bar{\theta}_j)\mathrm{e}^{-\kappa_j t} + \sigma_j e^{-\kappa_j t} \int_0^t e^{\kappa_j t'} dW_j(t') \ . \tag{68}$$

Noting that $\langle dW_j(t) dW_k(t) \rangle = \rho_{jk} dt$ (with $\rho_{jj} = 1$ for all $j$), the above expressions with Ito calculus can be used to obtain the covariances between any two variables (Gardiner 1985). The covariances between any two (scaled logarithmic) GBM processes ($j, k \in$ GBM),

$$\left\langle (U_j(t) - U_j(0) - \bar{\mu}_j t)(U_k(t) - U_k(0) - \bar{\mu}_k t) \right\rangle = \sigma_j \sigma_k \rho_{jk} t \ . \tag{69}$$

The covariances between any two (scaled logarithmic) GOU processes ($j, k \in$ GOU),

$$\left\langle (U_j(t) - \bar{\theta}_j - (U_j(0) - \bar{\theta}_j)\mathrm{e}^{-\kappa_j t})(U_k(t) - \bar{\theta}_k - (U_k(0) - \bar{\theta}_k)\mathrm{e}^{-\kappa_k t}) \right\rangle = \sigma_j \sigma_k \rho_{jk} \frac{1 - e^{-(\kappa_j + \kappa_k)t}}{\kappa_j + \kappa_k} \ . \tag{70}$$

The covariances between a (scaled logarithmic) GBM and GOU process ($j \in$ GOU and $k \in$ GBM),

$$\left\langle (U_j(t) - U_j(0) - \bar{\mu}_j t)(U_k(t) - \bar{\theta}_k - (U_k(0) - \bar{\theta}_k)\mathrm{e}^{-\kappa_k t}) \right\rangle = \sigma_j \sigma_k \rho_{jk} \frac{1 - e^{-\kappa_k t}}{\kappa_k} \ . \tag{71}$$

## Appendix 3: Maximum likelihood estimation for the multivariate GBM-GOU system

Noting that the above logarithmic variables $U_j(t) = \ln S_j(t)$ are coupled normal variables with the covariances provided above, the transition probability density $P(\boldsymbol{U}(t_i)|\boldsymbol{U}(t_{i-1}); \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\kappa})$ can be written exactly ($\Delta t = t_i - t_{i-1}$). Defining for short,

$$V_j^i \equiv U_j(t_i) - U_j(t_{i-1}) - \bar{\mu}_j \Delta t \tag{72}$$

for $j \in$ GBM and

$$V_j^i \equiv U_j(t_i) - \bar{\theta}_j - (U_j(t_{i-1}) - \bar{\theta}_j)\mathrm{e}^{-\kappa_j \Delta t} = U_j(t_i) - U_j(t_{i-1})\mathrm{e}^{-\kappa_j \Delta t} - \bar{\theta}_j(1 - \mathrm{e}^{-\kappa_j \Delta t}) \tag{73}$$

for $j \in$ GOU, one has

$$P(\boldsymbol{U}(t_i)|\boldsymbol{U}(t_{i-1}); \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\kappa}) = P(\boldsymbol{V}^i; \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\kappa}) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}\boldsymbol{V^i}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{V^i}} \tag{74}$$

where the covariances are given above with $t$ replaced by $\Delta t = t_i - t_{i-1}$,

$$\Sigma_{jk} = \sigma_j \sigma_k \rho_{jk} \Delta t \ . \tag{75}$$

for $j, k \in$ GBM,

$$\Sigma_{jk} = \sigma_j \sigma_k \rho_{jk} \frac{1 - e^{-(\kappa_j + \kappa_k)\Delta t}}{\kappa_j + \kappa_k} \ . \tag{76}$$

for $j, k \in$ GOU, and

$$\Sigma_{jk} = \sigma_j \sigma_k \rho_{jk} \frac{1 - e^{-\kappa_k \Delta t}}{\kappa_k} \,. \tag{77}$$

for $j \in$ GOU and $k \in$ GBM.

One then maximizes the log-likelihood function with respect to all parameters.

$$\begin{aligned} L(\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\kappa}) &= \ln \left( \prod_{i=1}^{N} P(\boldsymbol{U}(t_i) | \boldsymbol{U}(t_{i-1}); \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\kappa}) \right) = \sum_{i=1}^{N} \ln \left( P(\boldsymbol{U}(t_i) | \boldsymbol{U}(t_{i-1}); \boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\rho}, \boldsymbol{\kappa}) \right) \\ &= \frac{Nd}{2} \ln(2\pi) - \frac{N}{2} \ln \left( \det(\boldsymbol{\Sigma}) \right) - \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{V^i}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{V^i} \,. \end{aligned} \tag{78}$$

One can show (Johnson and Wichern 2002; Rayner 1985) that for multivariate normal distribution the MLE parameter estimates will be equal to the sample mean and sample covariances. Then, treating $\kappa_j$ and $x_j \equiv e^{-\kappa_j \Delta t}$ parametrically for $j \in$ GOU, for the corresponding MLE estimates for our scaled "means" we find

$$\hat{\mu}_j \Delta t = \frac{1}{N} \sum_{i=1}^{N} \left( U_j(t_i) - U_j(t_{i-1}) \right) \tag{79}$$

for $j \in$ GBM, and

$$\hat{\tilde{\theta}}_j (1 - \mathrm{e}^{-\hat{\kappa}_j \Delta t}) = \frac{1}{N} \sum_{i=1}^{N} \left( U_j(t_i) - U_j(t_{i-1}) \mathrm{e}^{-\hat{\kappa}_j \Delta t} \right) \tag{80}$$

for $j \in$ GOU. Further, the MLE estimates for the covariances one obtains

$$\hat{\Sigma}_{jk} \equiv \hat{\sigma}_j \hat{\sigma}_k \hat{\rho}_{jk} \Delta t = \frac{1}{N} \sum_{i=1}^{N} \left( U_j(t_i) - U_j(t_{i-1}) - \hat{\tilde{\mu}}_j \Delta t \right) \left( U_k(t_i) - U_k(t_{i-1}) - \hat{\tilde{\mu}}_k \Delta t \right) \tag{81}$$

for $j, k \in$ GBM,

$$\begin{aligned} \hat{\Sigma}_{jk} &\equiv \hat{\sigma}_i \hat{\sigma}_j \hat{\rho}_{jk} \frac{1 - e^{-(\hat{\kappa}_j + \hat{\kappa}_k) \Delta t}}{\hat{\kappa}_j + \hat{\kappa}_k} \\ &= \frac{1}{N} \sum_{i=1}^{N} \left( U_j(t_i) - U_j(t_{i-1}) \mathrm{e}^{-\hat{\kappa}_j t} - \hat{\tilde{\theta}}_j (1 - \mathrm{e}^{-\hat{\kappa}_j t}) \right) \left( U_k(t_i) - U_k(t_{i-1}) \mathrm{e}^{-\hat{\kappa}_k t} - \hat{\tilde{\theta}}_k (1 - \mathrm{e}^{-\hat{\kappa}_k t}) \right) \end{aligned} \tag{82}$$

for $j, k \in$ GOU, and

$$\begin{aligned} \hat{\Sigma}_{jk} &\equiv \hat{\sigma}_i \hat{\sigma}_j \hat{\rho}_{jk} \frac{1 - e^{-\hat{\kappa}_k \Delta t}}{\hat{\kappa}_k} \\ &= \frac{1}{N} \sum_{i=1}^{N} \left( U_j(t_i) - U_j(t_{i-1}) - \hat{\tilde{\mu}}_j \Delta t \right) \left( U_k(t_i) - U_k(t_{i-1}) \mathrm{e}^{-\hat{\kappa}_k t} - \hat{\tilde{\theta}}_k (1 - \mathrm{e}^{-\hat{\kappa}_k t}) \right) \end{aligned} \tag{83}$$

for $j \in$ GBM and $k \in$ GOU.

**Table 4** Parameter estimates for the simulated GBM and GOU processes shown in Fig. 2

| | $\hat{\bar{\mu}}$ | $\hat{\bar{\theta}}$ | $\hat{\sigma}$ | $\hat{\kappa}$ | $(C^{-1})_{n,n}$ |
|---|---|---|---|---|---|
| GBM 0.4 | -0.00397 | | 0.112 | | 1.296 |
| GOU 0.4 | | 8.89 | 0.1004 | 0.0786 | 1.406 |
| GBM 0.6 | 0.00381 | | 0.0904 | | 1.352 |
| GOU 0.6 | | 10.31 | 0.0960 | 0.0431 | 1.404 |
| GBM 0.8 | 0.00391 | | 0.0852 | | 1.950 |
| GOU 0.8 | | 9.88 | 0.0920 | 0.0544 | 2.006 |

Parameters were set as $\bar{\mu} = 0, \bar{\theta} = 10, \sigma = 0.1, \kappa = 0.1$. For a correlation of matrix with off diagonals 0.4, $(C^{-1})_{n,n} = 1.242$. For 0.6, $(C^{-1})_{n,n} = 1.517$ and for 0.8, $(C^{-1})_{n,n} = 2.143$

## Appendix 4: Parameter estimates by MLE

Parameter estimation for synthetic data shown in Fig. 2 can be seen in Table 4. Because of the nature of these stochastic equations and that we only have one realization to work with in our data set, we expect some significant variations in the accuracy of our predictions. In particular, the estimation of parameter $\kappa$ is known to exhibit very large relative errors for short sample sizes such as ours (Franco 2003; Tang and Chen 2009). Parameter estimation for the empirical data shown in Figs. 5 and 6 can be seen in Table 5.

## Appendix 5: Noise analysis and correlations in empirical data

In this section, we study the noise variables $\{\psi_i(t)\}$ extracted from the original time series (Fig. 1) by approximating the underlying processes by the SDEs as described in the main text. Figure 8 shows time series of the noise for the selected empirical data. Non-surprisingly, there is a periodicity of about seven days for the social media, indicating most users have a weekly periodicity to their social media usage. It is then natural to examine the temporal correlations present between the various noise time

**Table 5** Parameter estimates for the time series seen in Figs. 5 and 6

| | | $\hat{\bar{\theta}}$ | $\hat{\sigma}$ | $\hat{\kappa}$ | $(C^{-1})_{n,n}$ |
|---|---|---|---|---|---|
| Twitter | April | 19063 | 0.142 | 0.0103 | 1.26 |
| | June | 22878 | 0.150 | 0.0129 | 1.41 |
| | August | 24746 | 0.510 | 0.140 | 1.10 |
| Reddit | April | 2790 | 0.216 | 0.0338 | 1.18 |
| | June | 3944 | 0.214 | 0.0340 | 1.24 |
| | August | 4397 | 0.174 | 0.0238 | 1.13 |
| Github | April | 1337 | 0.309 | 0.0879 | 1.19 |
| | June | 1660 | 0.326 | 0.0947 | 1.32 |
| | August | 1651 | 0.278 | 0.0758 | 1.17 |
| | | $\hat{\bar{\mu}}$ | $\hat{\sigma}$ | | $(C^{-1})_{n,n}$ |
| Bitcoin | April | 0.00223 | 0.0404 | | 1.18 |
| | June | 0.00801 | 0.0393 | | 1.15 |
| | August | 0.01158 | 0.0484 | | 1.36 |
| Ethereum | April | 0.0196 | 0.0692 | | 1.32 |
| | June | 0.0308 | 0.0802 | | 1.34 |
| | August | 0.0179 | 0.0802 | | 1.40 |

Each three rows correspond to a different time series and each row in that subset correspond to the prediction in April, June, and August of 2017. As can be seen there is typically small changes over time for each parameter

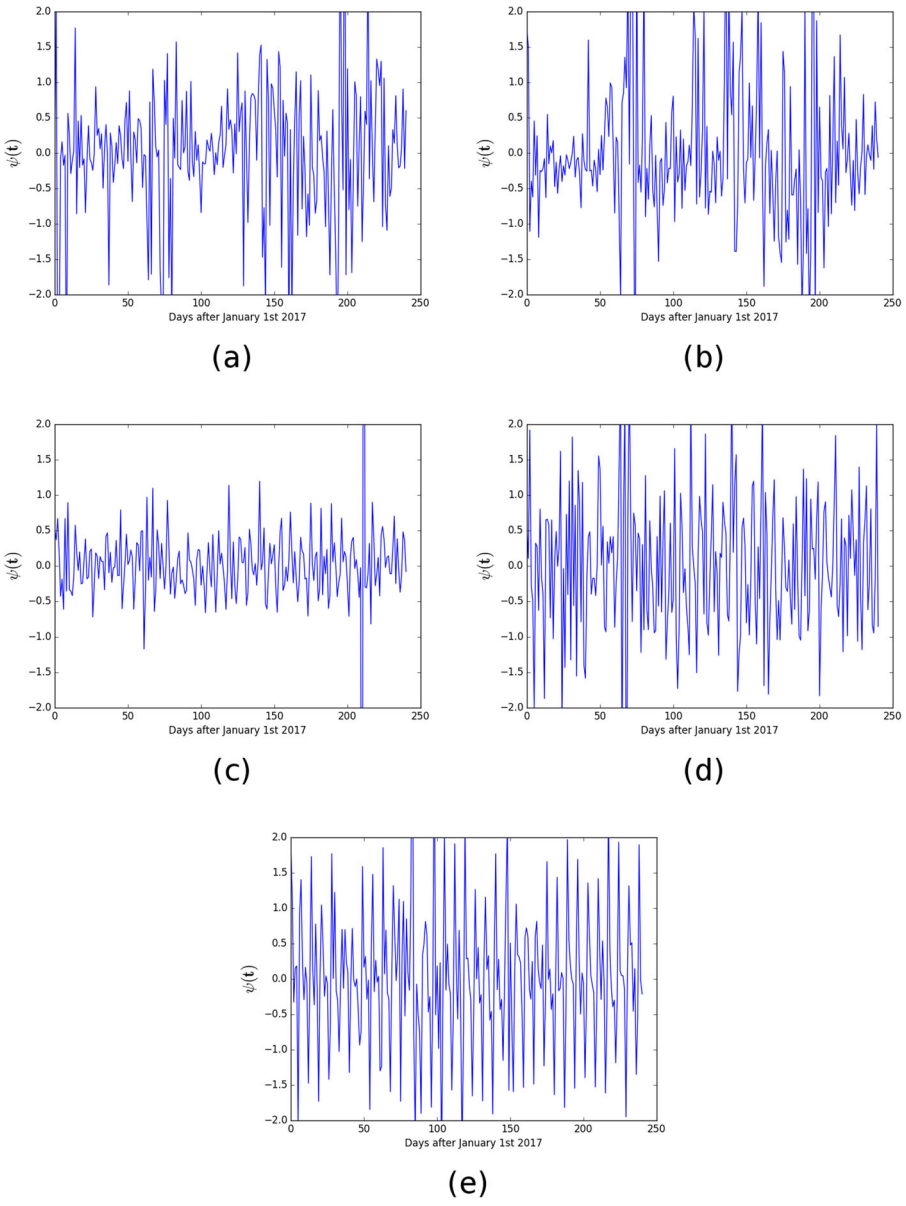


**Fig. 8** The time series of noise variables for (**a**) Bitcoin, (**b**) Ethereum, (**c**) Twitter, (**d**) Reddit, and (**e**) GitHub sampled from 242 days

series. We show the various permutations in Fig. 9. Here, we can see peeks in temporal correlation for social media at 7, 14, 21, etc. days. Also interesting to note, there are many instances of correlations being much more extreme than the “coincidental" correlations for small displacements, indicating a possible correlation for small changes in time.

While our equations operate on normal random variables, there is evidence to suggest other distributions may be a more appropriate fit (Mantegna and Stanly 2000). To test this, we first estimate the parameters using MLE and then extract and measure the noise variables from each time series. Figure 10 shows the distribution of noise variables for

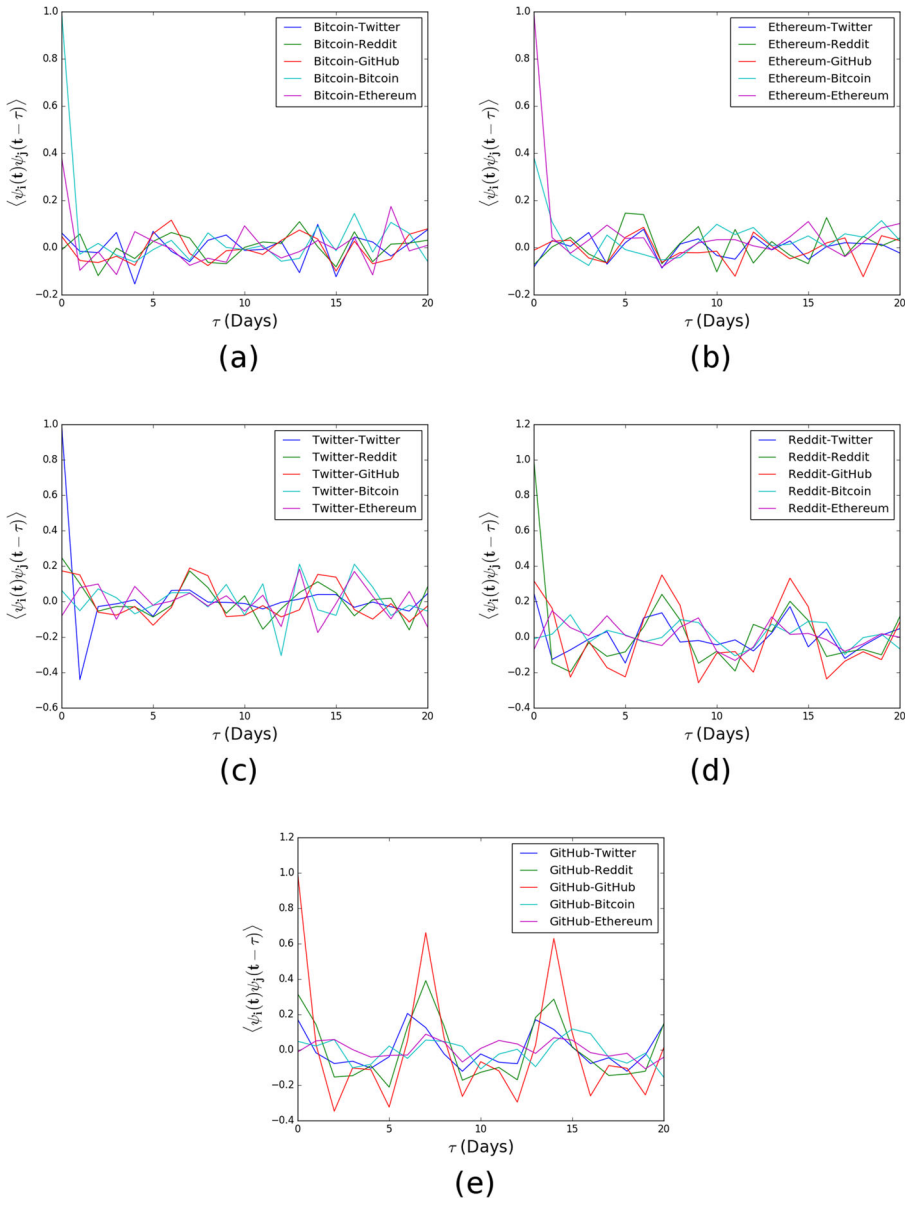


**Fig. 9** The temporal correlation of noise variables for fixed (**a**) Bitcoin, (**b**) Ethereum, (**c**) Twitter, (**d**) Reddit, and (**e**) GitHub using 242 days. Here, each time series is shifted by a number of days $\tau$ and the correlation coefficient is measured

some of our time series. As can be seen, there are various abnormalities associated with each distribution. This may be due to time varying parameters, however breaking time periods into smaller periods to examine this can create a potential subjective bias. Also, a few outliers [which can be directly observed in the original time series, e.g., for Twitter in Fig. 1a] give rise to outliers in the time series and histogram of the corresponding noise [e.g., somewhat visible in Fig. 8c, but not visible in Fig. 10c due to uniform scales used for all noise data for comparison]. These outliers, in turn, can strongly bias the MLE parameter values and produce noticeable deviations between the histogram of the empirical noise and a standard normal distribution [Fig. 10c].

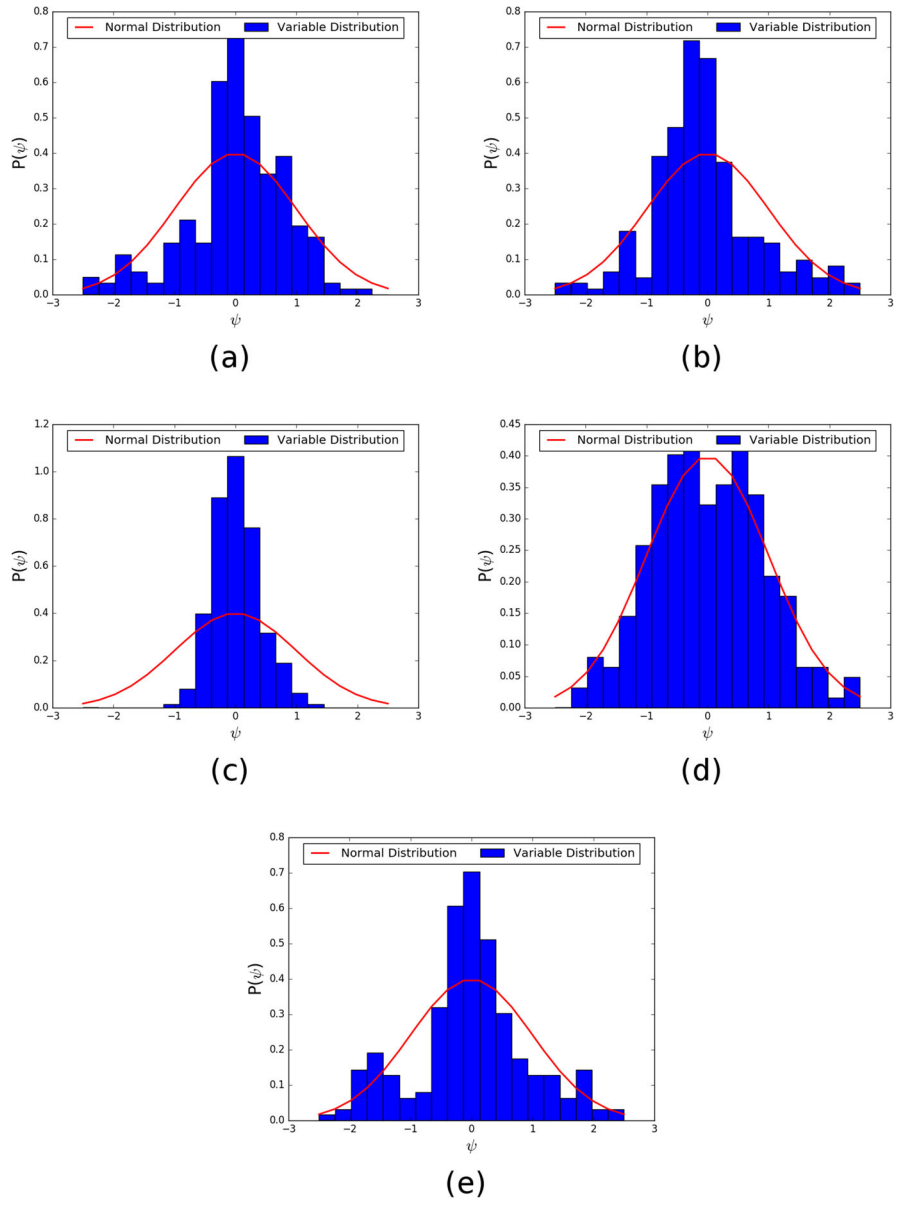


**Fig. 10** The distribution of noise variables for (**a**) Bitcoin, (**b**) Ethereum, (**c**) Twitter, (**d**) Reddit, and (**e**) GitHub sampled from 242 days of data collection compared against a normal distribution with mean zero and unit variance

**Acknowledgements**
This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) and the Army Research Office (ARO) under Contract No. W911NF-17-C-0099, the Army Research Laboratory (ARL) under Cooperative Agreement Number W911NF-09-2-0053 (NS-CTA), and by the Office of Naval Research (ONR) Grant No. N00014-15-1-2640. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies either expressed or implied of the Army Research Laboratory or the U.S. Government.

**Authors' contributions**
SD conceived the research; SD, AC, JF, BKS, and GK designed the research; SD implemented and performed numerical experiments and simulations; SD, AC, JF, BKS, and GK analyzed data and discussed results; SD, AC, JF, BKS, and GK wrote, reviewed, and revised the manuscript. The author(s) read and approved the final manuscript.

**Availability of data and materials**
The data that was used in this study was provided by the SocialSim Program of the Defense Advanced Research Projects Agency. Restrictions apply to the availability of these data, which was used under license for the current study, and so are

not publicly available. Data is however available from the authors upon reasonable request and with permission of the Defense Advanced Research Projects Agency.

**Competing interests**
The authors declare that they have no competing interests.

**Author details**
[1]Department of Physics, Applied Physics, and Astronomy, Rensselaer Polytechnic Institute, 110 8th Street, 12180-3590 Troy, NY, USA. [2]Network Science and Technology Center, Rensselaer Polytechnic Institute, 110 8th Street, 12180-3590 Troy, NY, USA. [3]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, 110 8th Street, 12180-3590 Troy, NY, USA. [4]Department of Computer Science, Rensselaer Polytechnic Institute 110 8th Street, 12180-3590 Troy, NY, USA.

Received: 29 July 2019 Accepted: 17 February 2020
Published online: 11 March 2020

**References**

Bassler KE, McCauley JL, Gunaratne GH (2007) Nonstationary increments, scaling distributions, and variable diffusion processes in financial markets. Proc Nat Acad Sci 104(44):17287–17290. https://doi.org/10.1073/pnas.0708664104
Black F., Scholes M. (1973) The pricing of options and corporate liabilities. J Polit Econ 81(3):637–654
Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. J Comput Sci 2(1):1–8. https://doi.org/10.1016/j.jocs.2010.12.007
Bouchaud J-P, Potters M (2000) Theory of Finacial Risks. Cambridge University Press, Cambridge, UK
Chaim P, Laurini MP (2019) Nonlinear dependence in cryptocurrency markets. N Am J Econ Finance 48:32–47. https://doi.org/10.1016/j.najef.2019.01.015
Cox MG, Hammarling S (1990) Reliable Numerical Computation. Oxford University Press, Oxford
Cretarola A, Figà-Talamanca G (2019a) Bubble regime identification in an attention-based model for bitcoin and ethereum price dynamics. Econ Lett:108831. https://doi.org/10.1016/j.econlet.2019.108831
Cretarola A, Figà-Talamanca G (2019b) Detecting bubbles in bitcoin price dynamics via market exuberance. Ann Oper Res. https://doi.org/10.1007/s10479-019-03321-z
Cretarola A, Figà-Talamanca G, Patacca M (2019) Market attention and bitcoin price modeling: theory, estimation and option pricing. Decisions Econ Finan. https://doi.org/10.1007/s10203-019-00262-x
Cretarola A, Figà-Talamanca G (2019) Modeling bitcoin price and bubbles. In: Salman A, Razzaq MGA (eds). Blockchain and Cryptocurrencies. Chap. 1. IntechOpen, Rijeka. https://doi.org/10.5772/intechopen.79386
DARPA Computational Simulation of Online Social Behavior (SocialSim). https://www.darpa.mil/program/computational-simulation-of-online-social-behavior. Accessed 25 Oct 2019
Franco JCG (2003) Maximum likelihood estimation of mean reverting processes. http://www.investmentscience.com/Content/howtoArticles/MLE_for_OR_mean_reverting.pdf. Accessed 12 Feb 2020
Fry J (2018) Booms, busts and heavy-tails: The story of bitcoin and cryptocurrency markets? Econ Lett 171:225–229. https://doi.org/10.1016/j.econlet.2018.08.008
Gardiner CW (1985) Handbook of Stochastic Methods, 2nd edn. Springer, New York City
Hurn AS, Lindsay KA, Martin VL (2003) On the efficacy of simulated maximum likelihood for estimating the parameters of stochastic differential equations. J Time Ser Anal 24. https://doi.org/10.1111/1467-9892.00292
Itô K (1944) Stochastic integral. Proc Imp Acad 20(8):519–524. https://doi.org/10.3792/pia/1195572786
Johnson RA, Wichern DW (2002) Applied Multivariate Statistical Analysis, 5th Edition. Prentice Hall, Upper Saddle River
Kim YB, Kim JG, Kim W, Im JH, Kim TH, Kang SJ, Kim CH (2016) Predicting fluctuations in cryptocurrency transactions based on user comments and replies. Plos One 11. https://doi.org/10.1371/journal.pone.0161197
Kreuser JL, Sornette D (2018) Bitcoin bubble trouble. Wilmott Vol. 95. pp 30–39
Lamon C, Nielsen E, Redondo E (2017) Cryptocurrency Price Prediction Using News and Social Media Sentiment. CS229 Final Project Report, Stanford University. http://cs229.stanford.edu/proj2017/final-reports/5237280.pdf
Mai F, Shan J, Bai Q, Wang S, Chiang R (2018) How does social media impact bitcoin value? a test of the silent majority hypothesis. J Manag Inf Syst 35:19–52. https://doi.org/10.1080/07421222.2018.1440774
Mantegna RN, Stanly HE (2000) An Intorduction To Econophysics. Cambridge University Press, Cambridge, UK
Maruddani DAI, Trimono (2018) Modeling stock prices in a portfolio using multidimensional geometric brownian motion. Journal of Physics: Conference Series 1025:012122. https://doi.org/10.1088/1742-6596/1025/1/012122
Máté G, Néda Z (2016) The advantage of inhomogeneity: Lessons from a noise driven linearized dynamical system. Phys A Stat Mech Appl 445:310–317. https://doi.org/10.1016/j.physa.2015.11.011
Mejía Vega CA (2018) Calibration of the exponential ornstein–uhlenbeck process when spot prices are visible through the maximum log-likelihood method. example with gold prices. Adv Differ Equ 2018(1):269. https://doi.org/10.1186/s13662-018-1718-4
Merton R. C. (1971) Optimum consumption and portfolio rules in a continuous-time model. J Economic Theory 3(4):373–413. https://doi.org/10.1016/0022-0531(71)90038-X
Merton RC (1973) The theory of rational option pricing. Bell J Econ Manage Sci 4:141–183
Øksendal B (2003) Stochastic Differential Equations An Introduction with Applications, 6th edn.. Springer, Berlin, Heidelberg
Onnela J-P, Chakraborti A, Kaski K, Kertész J, Kanto A (2003) Dynamics of market correlations: Taxonomy and portfolio analysis. Phys Rev E 68:056110. https://doi.org/10.1103/PhysRevE.68.056110
Phillips RC, Gorse D (2017) Predicting cryptocurrency price bubbles using social media data and epidemic modelling. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). pp 1–7. https://doi.org/10.1109/SSCI.2017.8280809
Plerou V, Gopikrishnan P, Rosenow B, Amaral LAN, Guhr T, Stanley HE (2002) Random matrix approach to cross correlations in financial data. Phys Rev E 65:066126. https://doi.org/10.1103/PhysRevE.65.066126

Rayner JCW (1985) Maximum likelihood estimation of $\mu$ and $\sigma$ from a multivariate normal distribution. Am Stat 39(2):123–124. https://doi.org/10.1080/00031305.1985.10479410
Reddy K, Clinton V (2016) Simulating stock prices using geometric brownian motion: evidence from australian companies. Aust Account Bus Financ J 10:23–47. https://doi.org/10.14453/aabfj.v10i3.3
Rosati P, Fox G, Lynn T (2018) Bitcoin and the Role of Social Media: An Empirical Analysis of Firm Level Legitimation Strategies. In: Proceeding of the 2018 British Academy of Management Conference (BAM 2018), Bristol
Saha K (2018) An investigation into the dependence structure of major cryptocurrencies. SSRN. https://doi.org/10.2139/ssrn.3241216
Sándor B, Néda Z (2015) A spring-block analogy for the dynamics of stock indexes. Phys A Stat Mech Appl 427:122–131. https://doi.org/10.1016/j.physa.2015.01.079
Sauer T (2013) Computational solution of stochastic differential equations. Wiley Interdiscip Rev Comput Stat 5(5):362–371. https://doi.org/10.1002/wics.1272
Schwartz ES (1997) The stochastic behavior of commodity prices: Implications for valuation and hedging. J Finance 52(3):923–973. https://doi.org/10.1111/j.1540-6261.1997.tb02721.x
Singh R, Ghosh D, Adhikari R (2018) Fast bayesian inference of the multivariate ornstein-uhlenbeck process. Phys Rev E 98:012136. https://doi.org/10.1103/PhysRevE.98.012136
Tang CY, Chen SX (2009) Parameter estimation and bias correction for diffusion processes. J Econ 149(1):65–81. https://doi.org/10.1016/j.jeconom.2008.11.001
Tarnopolski M (2017) Modeling the price of bitcoin with geometric fractional brownian motion: a monte carlo approach arXiv preprint arXiv:1707.03746
Teng L, Ehrhardt M, Günther M (2016) Modelling stochastic correlation. J Math Ind 6(1):2. https://doi.org/10.1186/s13362-016-0018-4
Wilmott P, Howison S, Dewynne J (1995) The Mathematics of Financial Derivatives. Cambridge University Press, Cambridge, UK
Wu K, Wheatley S, Sornette D (2018) Classification of cryptocurrency coins and tokens by the dynamics of their market capitalizations. R Soc Open Sci 5(9):180381. https://doi.org/10.1098/rsos.180381
Yao S, Hao Y, Liu D, Liu S, Shao H, Wu J, Bamba M, Abdelzaher T, Flamino J, Szymanski B (2018) A predictive self-configuring simulator for online media. In: 2018 Winter Simulation Conference (WSC). pp 1262–1273. https://doi.org/10.1109/WSC.2018.8632412

## Publisher's Note