Chapter 18
Dynamic Topic Modelling for Cryptocurrency Community Forums

M. Linton, E.G.S. Teo, E. Bommes, C.Y. Chen and Wolfgang Karl Härdle

Abstract Cryptocurrencies are more and more used in official cash flows and exchange of goods. Bitcoin and the underlying blockchain technology have been looked at by big companies that are adopting and investing in this technology. The CRIX Index of cryptocurrencies http://hu.berlin/CRIX indicates a wider acceptance of cryptos. One reason for its prosperity certainly being a security aspect, since the underlying network of cryptos is decentralized. It is also unregulated and highly volatile, making the risk assessment at any given moment difficult. In message boards one finds a huge source of information in the form of unstructured text written by e.g. Bitcoin developers and investors. We collect from a popular crypto currency message board texts, user information and associated time stamps. We then provide an indicator for fraudulent schemes. This indicator is constructed using dynamic topic modelling, text mining and unsupervised machine learning. We study how opinions and the evolution of topics are connected with big events in the cryptocurrency universe. Furthermore, the predictive power of these techniques are investigated, comparing the results to known events in the cryptocurrency space. We also test hypothesis of self-fulling prophecies and herding behaviour using the results.

M. Linton (✉)
University of York, Unter den Linden 6, Heslington York YO10 5DD, UK
e-mail: msl508@york.ac.uk

E.G.S. Teo
School of Business, Singapore Management University, 50 Stamford Road, Singapore 178899, Singapore
e-mail: elisabeth.bommes@googlemail.com

E. Bommes · C.Y. Chen
Ladislaus von Bortkiewicz Chair of Statistics, C.A.S.E. - Center for Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
e-mail: elisabeth.bommes@googlemail.com

C.Y. Chen
e-mail: chencath@hu-berlin.de

W.K. Härdle
C.A.S.E.-Center of Applied Statistics and Economics, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany
e-mail: haerdle@hu-berlin.de

W.K. Härdle et al. (eds.), *Applied Quantitative Finance*, Statistics and Computing,
DOI 10.1007/978-3-662-54486-0_18

18.1 Introduction

Cryptocurrencies such as Bitcoin have become more mainstream over the years with big companies adopting and investing in the technology. Once seen to be the domain of technophiles and radicals, cryptocurrencies are now widely traded on many exchanges throughout the world. Governments have also discussed the possibilities of adopting cryptocurrencies as a means to offer digital currency. The underlying network (called the blockchain) of cryptocurrency is decentralised, unregulated and highly volatile, making its situation at any given moment difficult to assess. On the other hand, an almost bottomless source of information can be found in the form of unstructured text written by cryptocurrency users on the internet. Crowd wisdom found in such networks can be a powerful indicator of major events affecting cryptocurrencies. We attempt to take advantage of this to analyse and assign quantitative meaning to such resources.

Early academic statistical analysis of Bitcoin includes Cheah and Fry (2015) and Cheung et al. (2015), both looked at speculative bubbles using Bitcoin price data. More related to this paper are works that looked at social media information and search engine data such as Kristoufek (2013), Mai et al. (2015) and Matta et al. (2015).

Utilizing techniques from dynamic topic modelling (DTM), text mining and machine learning, we pull data from a popular cryptocurrency forum and attempt to detect events such as new trends in currencies, fraudulent schemes or legal and economic issues. The DTM technique, as a type of unsupervised learning, is demanded when the taxonomy is unclear. Some important topics may be left out if one does a subjective judgement for taxonomy. The DTM is designed for summarizing the unknown but important features in the world. In addition to "discover" and "quantify" the hidden topics, the DTM is able to characterize the evolution of the hidden topics, which may be useful for evaluating the importance and persistence. Specifically, we collect user information and text associated with time stamps and apply unsupervised dynamic topic modelling, studying how opinions and the evolution of topics are connected with big events in the cryptocurrency universe. Furthermore, the predictive power of these techniques are investigated, comparing the results to known events in the cryptocurrency space. We also test hypothesis of self-fulfilling prophecies and herding behaviour using the results. For example, Smailović et al. (2013) were able to improve predictive power for stock markets by using sentiment derived from Twitter feeds. Cryptocurrency discussion forums tend to be very responsive and sensitive to events; this makes it a suitable candidate to test the predictive ability of dynamic topic modelling.

18.2 Data

A good, consistent and representative source of information regarding the cryptocurrency community can be found on talk forums such as http://bitcointalk.org. Acquiring the data from this platform requires deploying a web scraper to download the relevant html pages from the server and extract the embedded information. Good practices of web scraping were used to ensure there was no risk of overloading servers such as waiting fifteen seconds between each request and respect for the robots.txt protocol. Information regarding thread ids, post ids, usernames, time stamps, post titles, post texts, quotes of other posts and links were collected and stored in a database. There are three main discussion boards which were used in this study, they are "Bitcoin", "Economy" and "Alternative Cryptocurrencies". The two remaining discussion boards were "Other" which was discarded as it mainly deal with non-related topics and "Local" which is also discarded as discussions are in local languages. Each of the main discussion boards were divided into subforums such as "Trading Discussions" and "Scam Accusations". In total there were little under 200 subforums, half a million different threads with over 15 million posts (including local discussion). For the purpose of our study, we concentrate on the Bitcoin discussion subforum.

Knowledge is power so the more information we have, the better. Aside from this, the main motivations behind collecting these bits of information are as follows: Thread ids and post ids are used to uniquely identify posts and the thread they come from; usernames are used to associate each post with an agent in order to create a graph for herding and social network analysis; time stamps are used to classify posts into time slices for the dynamic topic model; post titles and post texts are used in conjunction to form a document for the dynamic topic model; links and quotes are used in order to analyse how posts relate to each other and other websites which is useful for herding and social network analysis.

18.3 Topic Modelling

We apply topic modelling to these forums in order to model trends in the community and to see how real life events effect the topics discussed and vice versa. The most commonly used model to model topics in machine learning is LDA (Latent Dirichlet Allocation) by Blei et al. (2003).

This model, however, makes the assumption that all documents modelled are exchangeable and therefore the aspect of time is completely lost and the idea of detecting events becomes pointless. Therefore, the model we use is the dynamic topic model proposed by Blei and Lafferty (2006), which is a variant of LDA that analyses documents in a set of predetermined discrete time slices and assumes topics evolve smoothly from slice to slice with Gaussian noise.

LDA is a generative probabilistic model for text, however it has also been applied successfully to other types of discrete data sets such as images. This model differs from most as it is completely unsupervised, therefore removing the bottleneck of having to acquire a trained model, and the problem it tries to solve is not classification into topics, but rather assigning topic distributions to documents. These properties mean that it is ideal to apply to large quantities of unstructured text where it would be impossible to obtain reliable training data to produce a model and simply classifying documents into topics would produce confusing and unrealistic results. Bao and Datta (2014) apply the LDA method to extract the risk types (meaningful topics) in Security Exchange Commission 10-K forms, and find many plausible and meaningful risk types that have been left out in a supervised learning scheme proposed by Huang and Li (2011). The inferred topics from a supervised learning only cover 78% of topic pools.

The Dirichlet distribution is defined on a $(k-1)$ dimensional simplex

$$\Delta_k = \left\{ q \in \mathbb{R}^k : \sum_{i=1}^{k} q_i = 1, q_i \geq o, i - 1, 2, \ldots, k \right\}. \tag{18.1}$$

It can be thought of as a distribution of random probability mass/density functions (pdf). An excellent example based introduction can be found in Frigyik et al. (2010).

Definition 18.1 Let Q be a real value in Δ_k and suppose that $\alpha \in \mathbb{R}^k, \alpha_i > 0$ and define $\alpha_0 \stackrel{\text{def}}{=} \alpha^T 1$. Then Q has a $Dir(\alpha)$ distribution with pdf $f(q;\alpha) = \frac{\Gamma(\alpha_0)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} q_i^{\alpha_i - 1}$.

Density plots are given in Fig. 18.1 for different α. Given a document with a certain word distribution, the task is obviously to determine α from the set of documents.

The gamma function is a generalization of the factorial function, $\Gamma(s) = s\Gamma(s-1)$ with $\Gamma(1) = 1$. The mean of a $Dir(\alpha)$ random variable is $EQ = \alpha/\alpha_0$. Note that α determines the "location" of words in documents, a "small" α creates sharp peaks on defined locations. You may think of the document that has been written by the poet in the flim "Shining", in the described $Dir(\alpha)$ framework, there is just one "big" peak of the words at "all work and no play makes Jack a dull boy". With just $k = 2$ words in a document the $Dir(\alpha)$ reduces to the Beta distribution with pdf

$$f(x; a, b) = \frac{\Gamma(a+b)}{\Gamma(a) + \Gamma(b)} x^{a-1}(1-x)^{b-1}. \tag{18.2}$$

For $\alpha = (a, b)^T$ with $Q = (X, 1-X) \sim Dir(\alpha)$ for $X \sim Beta(a, b)$.

In a Bayesian context, employed here entirely for numerical and computational reasons, one finds that the multinomial distribution with pdf

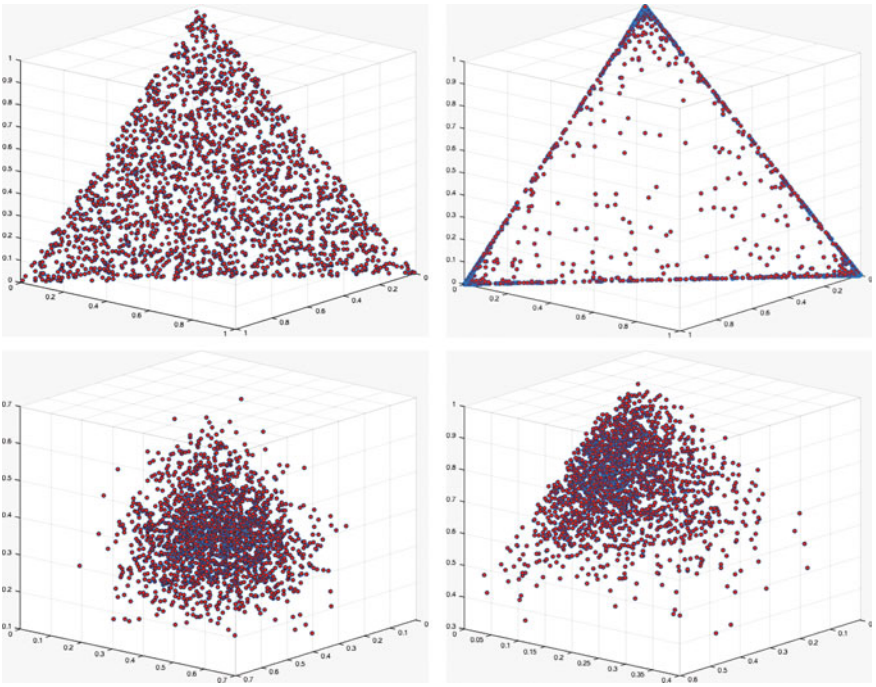

Fig. 18.1 Plots of sample pmfs drawn from Dirichlet distributions for various values of α. XFGtdmDirichlet

$$f(x; n, q) = \frac{n!}{\prod_{i=1}^{k} x_i!} \prod_{i=1}^{k} q_i^{x_i}, x, q \in \mathbb{R}^k \tag{18.3}$$

is a so called conjugate prior.

As the binomial distribution (for $k = 2$) is the conjugate prior for the Beta distribution, one finds that if $(X \mid q) \sim Mult_R(n, q)$ and $Q \sim Dir(\alpha)$, then $(Q \mid X = x) \sim Dir(\alpha + x)$. Again we refer for a proof of this to Frigyik et al. (2010).

The basic idea of a static Topic Model (TM) is to take a document as a sample of words generated by a $Dir(\theta)$ distribution, where θ represents the topic. More precisely it is assumed that a document is generated via the following imaginary random process:

1. For each topic k, draw a distribution over words $\vec{\beta}_k \sim Dir_v(\eta)$

2a. For each document d, draw topic proportions θ_d from over the $(k-1)$ simplex

2b. For each word $W_{d,n}$ within the document:

i. Draw a topic assignment $Z_{d,n} \sim Mult(\vec{\theta}_d)$, $Z_{d,n} \in \{1, \ldots, k\}$

ii. Draw a word $W_{d,n} \sim Mult(\vec{\beta}_{z_{d,n}})$, $W_{d,n} \in \{1, \ldots, V\}$

Table 18.1 Most frequent words used in NASDAQ articles

Word	Freq. (in k)	Freq. for top 5 sectors
Free	649	10
Well	238	9
Gold	235	1
Best	207	9
Fool	200	5
Strong	196	5
Like	172	5
Top	167	3
Better	162	0
Motley	152	2

β_z is a vector of β, one for each topic. β is a matrix of word|topic parameters.

The number of topics is assumed known beforehand though determining the number of topics (clusters) is rather challenging in unsupervised learning. One can easily find some methods being proposed for estimating the number of topics automatically, but one has to be aware of several restrictions. Firstly, Wallach et al. (2010) find that the estimated numbers of topics are strongly model-dependent. Besides, merely using fit statistics such as perplexity may be problematic due to a negative relation between the best fitted model and the substantive fit (Chang et al. 2009). To balance the substantive fit and statistical fit, Bao and Datta (2014) propose strategic procedures - Firstly, employing statistical fit to reduce the set of candidate models with different numbers of topics. Relying on the predefined perplexity, one can optimize the predictive power of model. In their case, the numbers can be chosen as 30, 40 and 50 in terms of perplexity and a converge in the range [30, 50] is shown. Secondly, the substantive fit for semantic coherence is compared among the competing models. To be specific, the model precision in word intrusion task is evaluated. It's so called "semantic validation". The semantic coherence of topics perhaps is the most useful indicator w.r.t the quality of topics, reflecting to how well the topic matches a human concept through a list of keywords. The number, 30, is therefore chosen due to its best semantic coherence performance.

Let us provide an example that sheds some light on this generation mechanism. Suppose that the "word universe" corresponds to the most frequent words in the NASDAQ analysis study by Zhang et al. (2016) and Bommes et al. (2017), as given in Table 18.1.

The idea is now that different topics have different word distribution as given by $Mult(\beta_z)$. Suppose there were $k = 2$ topics/sectors, corresponding to "finance" and "IT" and further suppose that the distribution of words over topics are generated by $Dir(\theta)$. To be precise, for $k = 2$, the Dirichlet distribution boils down to a $Beta(\theta)$ distribution. It could be the case that for the topic "finance", the third most frequent word "gold" is more concentrated. Whereas, for the topic "IT", concentration would be more around the words "fool" and "motley". See Fig. 18.2 below for an illustration

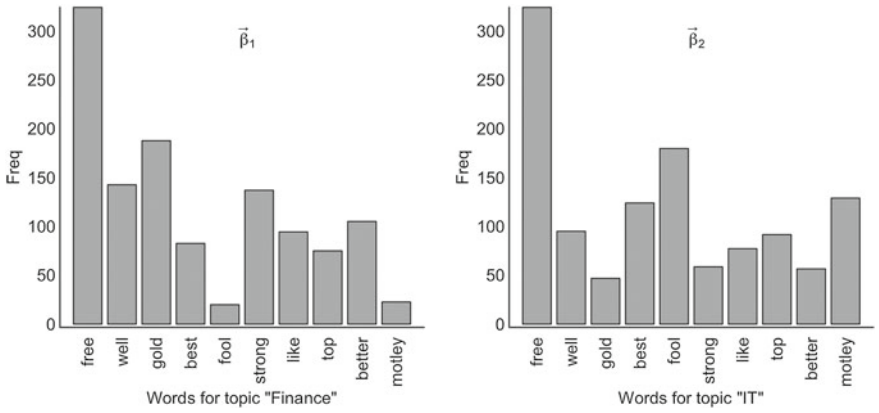

Fig. 18.2 Distribution of words by topic ($\vec{\beta}_1$ and $\vec{\beta}_2$). XFGdtmWDistr

that shows the random outcomes $\vec{\beta}_1$ and $\vec{\beta}_2$. In such as scenario, we would prefer a different word distribution for each these topics.

Step 2bi. now refers to the random mechanism that a word to be written down is drawn from $\vec{\beta}_1$ or $\vec{\beta}_2$. Suppose that the first has to be drawn from $\vec{\beta}_1$ since $Z_{1,1} = 1$, for $d = 1$ (1st document) and $n = 1$ (first word). So a random outcome as described in Step 2bii. could be the word $W_{1,1} =$ "gold" (the word with the second highest frequency in $\vec{\beta}_1$. For the next word ($n = 2$), $Z_{1,2}$ could take the value 1 again and now $W_{1,2} =$ "strong" could be the outcome. A third word could be via $Z_{1,3} = 2$, $W_{1,3} =$ "free", and so on. The task of TM is now to invert this mechanism and calibrate the observed documents to the parameters of the Dir and $Mult$ distributions.

The problem of static TM though is that there is no timeline, an issue that is of course necessary for the questions we would like to study here. The dynamic topic model, on the other hand models each time slice with LDA, but its parameters β and α are chained together in a state space model which evolves with Gaussian noise:

$$\beta_{t,k}|\beta_{t-1,k} \sim N(\beta_{t-1,k}, \sigma 2I) \tag{18.4}$$

$$\alpha_{t,k}|\alpha_{t-1,k} \sim N(\alpha_{t-1,k}, \delta 2I) \tag{18.5}$$

Like this we get a smooth evolution of topics from slice to slice. The state space diagram describes the model well.

Due to the nonconjugacy of the Gaussian and multinomial distributions, exact inference is intractable so the authors present two methods for approximate inference using variational methods: variational Kalman filtering and variational wavelet regression (Fig. 18.3).

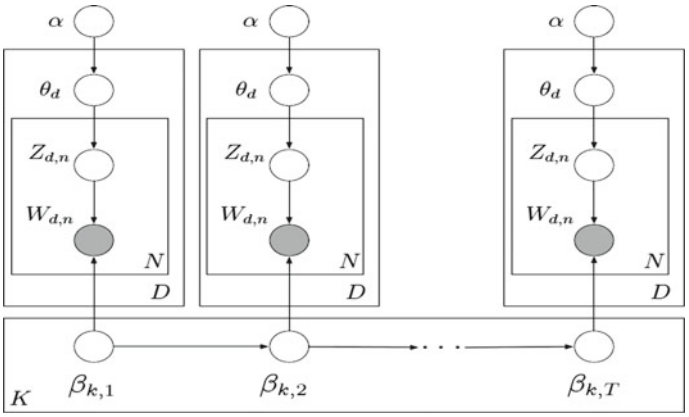

Fig. 18.3 State space diagram of the dynamic topic model

18.4 Preprocessing

Preprocessing steps make a big difference to the outcome of topic models. Especially when working in the domain of a forum where thousands of users post everyday, most likely without looking words up in the dictionary or worrying about the correctness of their grammar, we will find many spelling mistakes, slang and proper names that aren't going to be simple to handle. Therefore, a natural approach to preparing the data appropriately would be to use a POS tagging algorithm coupled with a tokeniser to infer from context what words have which function. Stop words will appear multiple times in each sentence without conveying any meaning and therefore are removed and so are functional words, verbs, adjectives and adverbs leaving us only with nouns, proper nouns and foreign words. In this way we have all the most important information from each post without losing out on non-standard vocabularies that arise in the community. To combat typos, the words occurring in fewer than 10 documents were removed and to get rid of generic words, the words appearing in more than 10% of the documents were also removed. In the end, from a dictionary of 500,000 words, we obtained one of 10,000 meaningful words. Once we had the cleaned text, the preparation for the dynamic topic model (code by Sean M. Gerrish) consisted of converting the corpus to a sparse matrix representation whereby each line represented a document and was in the following form:

$N\_unique\_words\ word\_id : word\_count\ word\_id : word\_count.....$

Also a file containing information about the time slices was prepared of the following format:

N_time_slices
$N\_docs\_slice\_1$
$N\_docs\_slice\_2$
...

Where N denotes number of documents in the corresponding slice. On top of these necessary files, for each corpus a file containing metadata, a dictionary file and a vocabulary file were also produced. The metadata file contains a header describing the fields and then each line represents a document with the following pieces of information: thread id, post id, date time, username, post text, post quotes and post links. This will come in handy for information retrieval and herding analysis. The dictionary file is a python dictionary object which maps ids to words and contains word count information. The vocabulary file is a human readable file where each line is a word from the dictionary and its position maps to its key.

18.5 Trends

As mentioned in the introduction, the data acquired from the forum was divided into subforums. The main subforums by posting volume are: 'Economics', 'Bitcoin Discussion', 'Altcoin Discussion' and 'Speculation'. The dynamic topic model was run on these subforums and in addition also with the subforum 'Scam Accusations'. The commonly used $50/k$ heuristic by Griffiths and Steyvers (2004) for the alpha parameter was chosen and a varying number of topics were modelled. All models were run with weekly data over the 2009/11/22 (when the forum was created) to 2016/08/06 period.

Each topic in the hidden structure is represented as a distribution over words and therefore the most human interpretable way of understanding what a topic is about is to look at the most probable words in each distribution. An example representation can be found in Table 18.2 in which some topics are shown for the last time slice in the Bitcoin Discussion subboard. Each time slice will have it's own similar representation. While the words may change over time as new trends emerge and fall, the topic will intuitively remain the same. For example, in the table shown we can see that topic 50 is about Bitcoin mining, but the top words in the first time slice are rather different even though we would still assign the same topic label to it; cpu, difficulty, proof, mining, adjustment, proof-of-work, power, attack were the top words in 2009 in topic 50, demonstrating how Bitcoin mining has evolved to cope with the increasing mining difficulty. In fact we can directly compare different mining hardware and how they were relevant over different periods of time in Fig. 18.4.

As we can see, in topic 50 the word CPU was very prominent initially and all the others were non-existent. Then when the network grew to an extent that the quantity of Bitcoins produced by CPU mining were worth less than what it cost to operate, GPU mining came into play. Another stride in mining hardware was the usage of application specific integrated circuits (asic). The first asic mining hardware project called the 'Avalon Project' was announced in 2012 on the forum and the peak in the third plot in January 2013 corresponds to the release of their first chip. In the fourth plot we see the timeline of Antminer, a brand of asics considered to be the current top of the line. As expected we can see a positive trend over the last years with peaks in discussion around releases of new models.

Table 18.2 Notable topics from 50 topic model on Bitcoin Discussion subforum from 2016/07/31 to 2016/08/06

Topic number	Most probable words
1	Value, gold, bar, dollar, rate, demand, interest, asset
2	Business, casino, house, trust, gambling, run, strategy, player
5	Government, control, criminal, law, study, regulation, state, rule
7	Use, service, option, cash, good, spend, fiat, convert
12	Account, payment, fund, card, paypal, party, merchant, credit
18	Score, online, pay, shop, bill, product, purchase, phone
20	Wallet, key, paper, computer, storage, code, data, secure
23	Price, trade, market, trader, drop, volume, sell, stock
24	Trading, term, hold, buy, pump, dump, earn, gamble
30	Exchange, bitfinex, lesson, cryptocurrency, crash, platform, altcoins, popularity
32	Investment, risk, invest, aim, impact, salary, making, way
33	Year, altcoins, end, today, adoption, prediction, happen, trend
35	Transaction, block, fee, chain, confirmation, hour, minute, hardfork
38	Altcoin, company, loss, hack, scam, hacker, scammer, road
42	Bank, system, security, fiat, banking, role, function, institution
45	Ethereum, split, advantage, issue, side, change, fork, core
48	Forum, post, topic, member, bitcointalk, thread, index, php
50	Mining, miner, network, power, pool, cost, reward, electricity

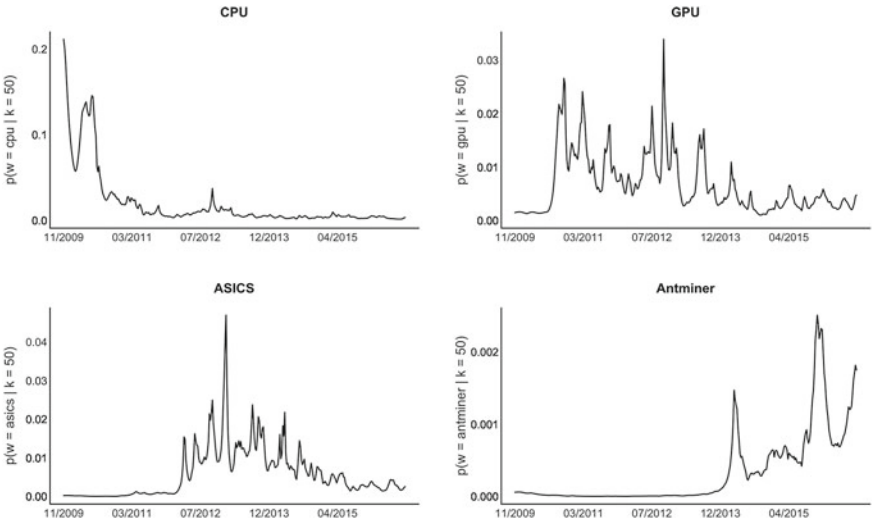

Fig. 18.4 Comparison of word evolution for different mining technologies 22/11/2009–06/08/2016.
XFGdtmMining

As an up and coming and fast growing technology, Bitcoin has had its fair share of issues. In fact, due to its unregulated nature and uncertainty of legality or legitimacy as currency in most corners of the world, the cryptocurrency history is laden with high profile hacks, ponzi schemes and scam websites. Many of these go undetected for months until a certain point where gradually complaints start to stack up and a realisation or confirmation of the events takes place.

Probably the biggest example of such an event in Bitcoin history is the insolvency of the MtGox Bitcoin exchange in 2014. MtGox originally started off in 2007 as a platform for trading Magic: The Gathering Online trading cards which is where it got its name (Magic: The Gathering eXchange). In 2010, however, it was rebranded as one of the first exchanges where people could buy and sell Bitcoins. The exchange grew gradually and watched the price of Bitcoin go from less than 0.1 USD in 2010 to parity with the US dollar in 2011. At this point however, the owner of MtGox decided to sell the exchange in order to dedicate himself to 'other projects'. An internal email dating back from after the sale of the exchange revealed that already 80,000 Bitcoins (worth over $60,000 at the time) had already been missing before any of the public fiascos had occurred and had never been recovered. However, it was only three months later that a major event occurred. 60,000 accounts were exposed publicly and a compromised MtGox auditors account was used to create huge sell orders and crash the Bitcoin price from $17.51 to $0.01. As a result of this event the site was down for a week and many of the exposed accounts were used to steal coins from other bitcoin services due to password reuse. However, unlike many other Bitcoin services, MtGox managed to recover its reputation and became the largest Bitcoin exchange, handling 70% of all trades worldwide. Fast forwarding to 2013, when their real problems began, in June withdrawals of US dollars were suspended and even though a couple of weeks later in July it had been announced that withdrawals had fully resumed, as of September few withdrawals had successfully been completed. Complaints piled up over the next few months and on 7 February 2014 all Bitcoin withdrawals had been suspended for good. On the 24th of February all activities had halted, the website went offline and a leaked internal crisis management document claimed that 744,408 Bitcoins (worth almost half a billion dollars) had been lost and the company was insolvent.

As we can see, MtGox has had a roller coaster of a past with repeated security issues and poor management and has therefore been a major topic of discussion among users of the main Bitcoin forum. The main topics in which MtGox arises are predictively topic 23 about Bitcoin trading and markets and topic 38 about scams and hacks. Naturally the word/topic probability plot in Fig. 18.5 reflects this and we can see peaks corresponding to the main events. In topic 38 there is a clear peak in mid 2011 during the first hack and in February 2014 also. Meanwhile in topic 23 there is a gradual peak starting in mid 2013 when the transaction issues first occurred and trailing off at the same time MtGox starts to gain momentum in topic 38.

MtGox is only one example of the many scams and hacks resulting in huge losses that have occurred over the years and it is because of this that cryptocurrencies get a bad rap. Many services have come and gone, but none quite so spectacularly as MtGox.

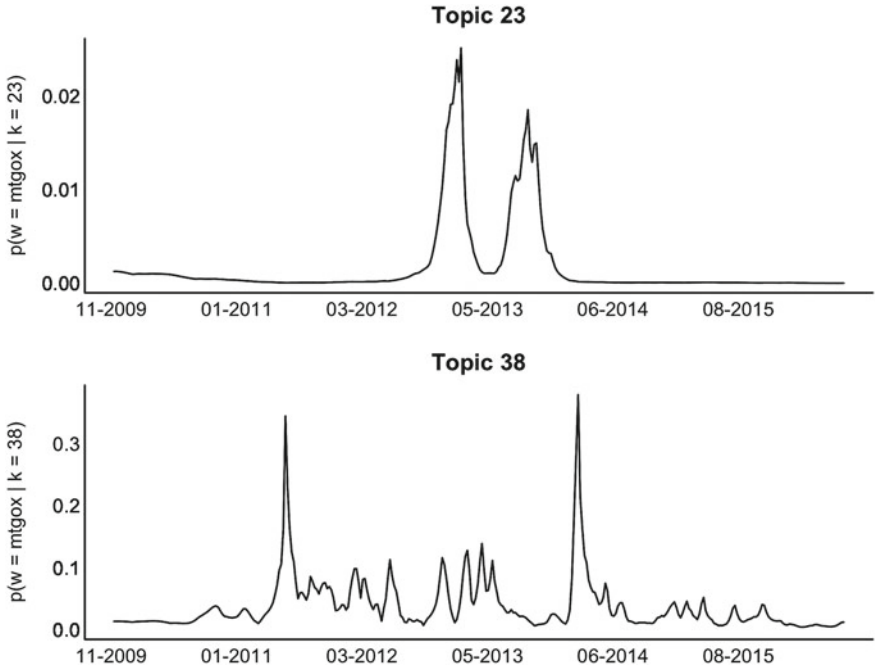

Fig. 18.5 MtGox word evolution 22/11/2009–06/08/2016. XFGdtmMtGox

Currency exchanges, mining hardware manufacturers, technology startups, mining pools and many other cryptocurrency related services have almost infallibly been victims of hacks and inside jobs, revealed as ponzi schemes, virus promoters etc. As soon as such events occur or are discovered, we would expect there to be gradual buildups or sudden explosions of discussion on the forum depending on the situation. In general, we would expect any event in the Bitcoin universe to be discussed on the forum and therefore be a part of the inferred generative process of the topic structure.

We want to evaluate the effectiveness of topic models in discerning these types of events. In our MtGox example, the word probabilities over time are characterised by relatively flat probabilities in general and spikes at the time of events. We can take advantage of this structure and hypothesise that it extends to other events. First we must validate this against other events. A curated list of Bitcoin services which have been victims of hacks or perpetrators of scams have been compiled over the years in a thread on http://bitcointalk.org (https://bitcointalk.org/index.php?topic=576337.0). This list will form our basis for event discovery validation. This could be done for other types of events however the most complete information can be found regarding scam/hack events since they are of relevance and interest to all involved with Bitcoin. We look at the topic *prominence* for this set of words and see if the model correctly partitions them in a scam/hack topic.

18.6 Choosing K and Analysis

The choice of the number of topics has been an issue ever since topic models were first introduced in 2003. For this particular study, we used the Umass coherence metric by Mimno et al. (2011) to evaluate which number of topics was optimal. This method involves taking the top N words for each topic and taking measures of their occurrences and co-occurrences in the corpus. Formally it is defined as:

$$\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{D(w_i, w_j) + \epsilon}{D(w_i)} \tag{18.6}$$

where w_i and w_j are the ith and jth ranked words in a given topic respectively and $D(w)$ is the number of documents in which word w occurs. We set $N = 20$.

It has been shown to correlate well with human interpretations of what constitutes a coherent topic. In addition, the metric does not require external validation, simplifying the procedure and making it more versatile. To make the repeated training of models viable, we calculated Umass coherence on a subsample of 100 weeks of data. In Table 18.3 we can see the results of the coherence evaluation. We have taken the arithmetic mean and standard deviation of the output values over the 100 chained LDA models; higher values mean more human understandable topics. Clearly our model is optimal when we choose 30 for the k parameter since on average the topics are more coherent and stable over time. We also observe that lower numbers of k are more coherent than higher values, but are also less stable over time. While this method does a good job at finding the number of topics more attuned to human intuition, we would also like to study how this effects event detection.

The generative process described now gives us a multi-layer interpretation of the data. We have K topics with D documents and W words. Each topic can be described by a vector of length W of word/topic probabilities. Each document can be described by a vector of length K of topic/document probabilities. Each topic changes over each of the T time slices and therefore each topic/document distribution acquires a different meaning depending on where it is in the timeline.

Say we have a particular word w in our vocabulary we would like to learn something about. The best way to do this is to look at the word probabilities over a certain

Table 18.3 Topic coherence statistics

Number of topics k	μ	σ
10	–185.74	66.62
20	–204.28	65.57
30	–176.46	52.80
40	–202.10	68.99
50	–205.83	63.17

time slice in the topics. We can call this concept the word *prominence* and we would like to maximize this in order to find the most relevant topic.

$$\arg\max_{k} \frac{1}{t_j - t_i} \sum_{t=t_i}^{t_j} p(w|k, t_i) \tag{18.7}$$

Once we have found this topic (or topics if we want to find several), looking at the topics top words will allow us to discover in which context this term is discussed the most. We can also plot the evolution of the probability over time of this particular word in this topic and see when it was most used, when it came into use or passed out of use. Quite often words with same spelling but different meaning (homonyms) occur or words that can be discussed in different contexts (for example *price* could be present in a stock market topic or in a groceries topic). Whereas usually it wouldn't be a simple task to discern these words, topic models account for them very nicely and provide a useful perspective.

In addition to analysing the word/topic distribution we can also take a look at the topic/document distributions and determine in which time slice which topics were 'hotter' and which were 'colder' and identify trend starters. The hotter a topic k at time t, the more documents are going to exhibit higher mixtures of the topic. The inverse is true for colder topics. We can define the topic *temperature* as follows by Hall et al. (2008):

$$\sum_{d:t_d=t} p(k|d)\,p(d|t) = \frac{1}{D_t} \sum_{d:t_d=t} p(k|d) \tag{18.8}$$

where D_t is the number of documents in time slice t and t_d is the date document d was written.

18.7 Detection

From the list of events acquired from the forum, all those solely concerning individuals or causing losses of fewer than 1000 Bitcoins were removed. As a consequence of this procedure, we were left with 33 different Bitcoin services (and 37 different events). For each word we determine which topics the word achieves a topic *prominence* larger than a certain threshold. Typically, any given word will only appear in a handful of topics and most in just 1 or 2. Even though a certain topic may not have anything to do with a chosen word, topic models have the property that the probability of a word occurring in a topic is never 0, albeit negligible. Therefore we use a very low empirically tested threshold to determine which topics to test and discard the noisy ones. Then we analyse the topic *prominence* of the words conditioned on topics through time and determine an event occurring to be when its upper control limit is breached. I.e. when:

Table 18.4 Events in chronological order, an asterisk means undetected in the 50 topic model

Event	Dates	Topic
Ubitex* (1,138b)	2011-04 to 2011-07	None
Allinvain	2011-06-13	23
MtGox	2011-06-19	23
Mybitcoin	2011-06-20, 2011-07	23
Bitomat	2011-07-26	23
Mooncoin	2011-09-11	23
Bitscalper	2012-01 to 2012-03	23
Linode	2012-03-01	23
Betcoin* (3,171b)	2012-04-11	None
Bitcoinica	2012-04-12, 2012-07-13	23
Btc-e	2012-07-13	12
Kronos	2012-08	23
Bitcoin Savings and Trusts	2012-08-28	23
Bitfloor	2012-09-04	23
Btcguild* (1,254b)	2013-03-10	None
OkPay (main victim of 2013 Fork)	2013-03-11	30
Ziggap* (1,708b)	2013-02 to 2013-04	None
Just-Dice	2013-07-15	23
Basic-Mining* (2,131b)	2013-10	None
Silkroad2	2013-10-02	23
Vircurex* (1,454b)	2013-10-05	None
GBL	2013-10-26	12
Bips* (1,294b)	2013-11-17	None
Picostocks* (5,896b)	2013-11-29	None
MtGox	2014-02-24	23
Flexcoin	2014-03-02	23
Cryptorush	2014-03-11	23
Mintpal	2014-10-14	23
Silkroad2	2014-11-06	23
Bitstamp	2015-01-04	23, 25
Bter	2015-02-14	23
Cryptsy	2016-01-01	23
Shapeshift	2016-04	23
Gatecoin*	2016-05-13	None
Bitfinex	2016-08-03	12

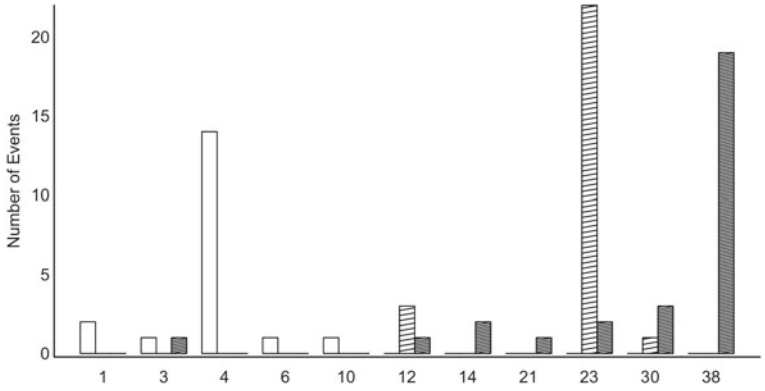

Fig. 18.6 Event partitioning over varying k parameters, 10 topics (*no filling*), 30 topics (*dashed filling*), 50 topics model (*densely dashed filling*). XFGdtmEvents

$$p(w|k, t_{i+1}) > \mu(p(w|k, t_{1:i})) + 3\,\sigma(p(w|k, t_{1:i})) \tag{18.9}$$

Table 18.4 contains the information regarding our events and the dates they occurred. We compared these events against those detected in our model using the method described and have marked with an asterisk those that went undetected.

Most of the events causing losses of circa 2000 Bitcoins and under (indicated) went undetected and almost all of those causing larger losses were identified. As hypothesized in the previous section, the large majority of these events were found to be in a single topic (topic 38), demonstrating the effectiveness of topic models in discriminating event types and providing an indicator for future such events.

This event detection algorithm was also run on our 10, 30 and 50 topic models. For the varying number k we can see what effect it has on our event distribution in Fig. 18.6. With the number of topics considered to be most coherent, our events are grouped mainly into a single topic. On the other hand, the less coherent topics are composed of many junk topics in the higher k case, or more general topics in the lower, therefore resulting in inconsistency in the experiment. A lower k results in fewer detections as our topics will each be less relevant and a higher k results in many junk topics and detections across more topics.

In addition, for each event we can observe the impact it has on the topic structure by measuring the deviation of the topic *temperature* from the mean at the time in which it occurred. Since our timeline and number of time slices is large and we are using a symmetric Dirichlet prior, our topics are going to be rather general and fixed through time and the change in temperature between different times won't be significant. However, one can note in Fig. 18.7 that all values are positive at the times the events occurred and appreciate the event hierarchy that follows.

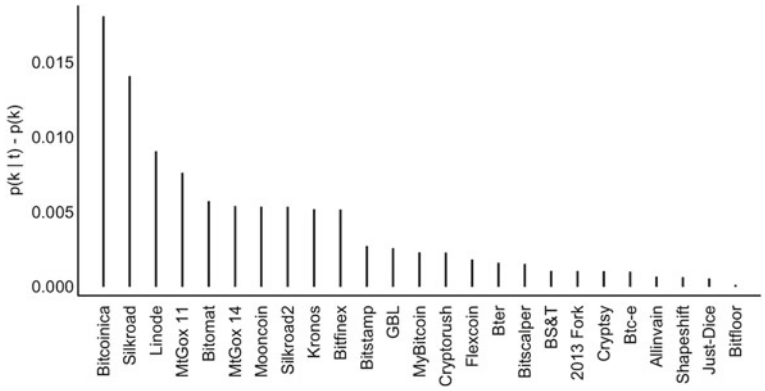

Fig. 18.7 Plot of ordered topic *temperature*s at time of event with k being the event topic and t being the time of the event XFGdtmTemperature

18.8 Conclusion

In the above piece of work we have introduced and explained topic models. A dataset has been created from user posts on http://bitcointalk.org by using web scraping; then text-mining techniques were used to prepare the data for dynamic topic modelling and consequently a walk through of all the steps for constructing such a model has been provided. We have presented a study and exploration of the popular cryptocurrency forum in this framework and employed an event detection technique to capture the effect of high profile scamming and hacking on the community. The number of topics parameter has been shown to be optimal for event detection when it accords with a measure of topic coherence. In addition, the constructed model partitions almost all of the events above a certain severity in a single topic.

References

Bao, Y., & Datta, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science*, *60*(6), 1371–1391.

Blei, D., Ng, A. Y., Jordan, M. I., & Lafferty, J. (2003). Latent Dirichlet allocation; *Journal of Machine Learning Research, 3*, 993–1022.

Blei, D., & Lafferty, J. (2006). Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning (AMC)*.

Bommes, E., Chen, C. Y., Härdle, W. K. (2017). Textual sentiment and sector-specific reaction. *Forthcoming*.

Chang, J., Boyd-Graber, J. L., Wang, C., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 288–296.

Cheah, E. T., & Fry, J. (2015). Speculative bubbles in Bitcoin markets? An empirical investigation into the fundamental value of Bitcoin. *Economics Letters*, *130*, 32–36.

Cheung, A., Roca, E., & Su, J. J. (2015). Crypto-currency bubbles: An application of the Phillips-Shi-Yu (2013) methodology on Mt. Gox bitcoin prices. *Applied Economics*, *47*(23), 2348–2358.

Frigyik, B. A., Kapila, A., & Gupta, M. R. (2010). Introduction to the Dirichlet distribution and related processes. *Technical Report*, Department of Electrical Engineering, University of Washington.

Griffiths, T., & Steyvers, M. (2004). Finding Scientific Topics. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(Suppl1), 5228–5235.

Hall, D., Jurafsky, D., & Manning, C. (2008). Studying the history of ideas using topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 363–371.

Huang, K. W., & Li, Z. L. (2011). A multilable text classification algorithm for labeling risk factors in SEC form 10-K. *ACM Transactions on Management Information Systems (TMIS)*, *2*(3), 18.

Kristoufek, L. (2013). BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports*, *3*, 3415.

Mai, F., Bai, Q., Shan, Z., Wang, X. S., & Chiang, R. H. (2015). The impacts of social media on Bitcoin performance. In *Proceedings of the Thirty Sixth International Conference on Information Systems (ICIS 2015)*.

Matta, M., Lunesu, I., & Marchesi, M. (2015). Bitcoin spread prediction using social and web search media. *Proceedings of DeCAT*.

Mimno, D., Wallach, H. M., Talley E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272.

Smailović, J., Grčar, M., Lavrač, N., & Žnidaršič, M. (2013). Predictive sentiment analysis of tweets: A stock market application. In *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data* (pp. 77–88). Berlin: Springer.

Wallach, H. M., Jensen, S. T., Dicker, L. H., & Heller, K. A. (2010). An alternative prior process for nonparametric Bayesian clustering. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, *9*, 892–899.

Zhang, J. L., Härdle, W. K., Chen, C. Y., & Bommes, E. (2016). Distillation of news flow into analysis of stock reactions. *Journal of Business and Economic Statistics*, *34*, 547–563.