Knowledge Discovery Approach from Blockchain, Crypto-currencies, and Financial Stock Exchanges

Sofiane Lagraa, Jeremy Charlier, Radu State
SnT, University of Luxembourg
Luxembourg
firstname.lastname@uni.lu

ABSTRACT

Last few years have witnessed a steady growth in interest on crypto-currencies and blockchains. They are receiving considerable interest from industry and the research community, the most popular one being Bitcoin. However, these crypto-currencies are so far relatively poorly analyzed and investigated. Recently, many solutions, mostly based on ad-hoc engineered solutions, are being developed to discover relevant analysis from crypto-currencies, but are not sufficient to understand behind crypto-currencies.

In this paper, we provide a deep analysis of crypto-currencies by proposing a new knowledge discovery approach for each crypto-currency, across crypto-currencies, blockchains, and financial stocks. The novel approach is based on a conjoint use of data mining algorithms on imbalanced time series. It automatically reports co-variation dependency patterns of the time series. The experiments on the public crypto-currencies and financial stocks markets data also demonstrate the usefulness of the approach by discovering the different relationships across multiple time series sources and insights correlations behind crypto-currencies.

KEYWORDS

Time series, crypto-currency, blockchain, gradual pattern mining, graph mining, imbalanced data, knowledge discovery

1 INTRODUCTION

The number of crypto-currencies has increased frequently on a monthly basis. As of 7 December 2017, more than 1300 were available. This number is still increasing [1]. The most popular decentralized crypto-currency is the *Bitcoin* (BTC) [4, 16], representing more than 180 GB as of February 2018. It can be seen as a large ledger of transactions, called *blockchain* that represents transfers of bitcoins. The blockchain is maintained by a peer-to-peer network of nodes, and a consensus protocol ensures it can only be updated consistently. The transactions of the blockchain cannot be altered, removed or deleted if already-published. Each transaction contains transaction id, sender, receiver, value (in BTC), and timestamp. The same transaction can involve multiple sender addresses and multiple receiver addresses. Furthermore, multiple addresses can belong to the same user. Users are also anonymous, there is no information associated with a given user. Public blockchains with their crypto-currencies constitute an unprecedented research dataset of financial transactions.

In 2017, Bitcoin price surged from 1,000 USD to 20,000 USD. In January 2018, it lost half of its value. Many critics, mostly inherited from finance, raised against the strong volatility related to crypto-currency markets and their absence of regulation. For a better understanding of the crypto-currency market movements, we address combinatorial analysis for crypto-currencies and stock markets. More particularly, we focus on the following 13 crypto-currencies, Bitcoin, Bitcoin Cash, Litecoin, Dash, PIVX, Monero, Dogecoin, Decred, XEM, Ethereum, Ethereum Classic, ZCash, and Vertcoin. In parallel, we select 6 major financial stocks index, SX5E, HSCEI, NIKKEI, NASDAQ and DOW JONES. They represent the stock index of Eurozone stocks, Hang Seng China Enterprises Index, Tokyo stock exchange, American stock exchange of technology companies and American stock exchange of industrial companies. All time series have a different starting date of quotation. Thereby, we can ask the following questions:

- Does a strength relation exist between a financial crypto-currency and its blockchain data?
- Does relationship exist between multiple crypto-currencies in different blockchains and financial currencies?
- Are there strong relationships between financial crypto-currencies and stock exchanges?

To answer these questions, we leverage on co-occurrences of co-variations of several features from crypto-currencies and stock markets. One example can be the pattern "*the higher the bitcoin price, the higher the transaction count, the lower the fees*". Our solution is based on the combination of gradual pattern mining and graph pattern mining applied to data mining on financial data. This state of the art combinatorial algorithm tackles the difficulty of finding dependencies between crypto-currencies and between financial stock exchanges.

This paper has the following major contributions:

- We propose a knowledge discovery approach for mining imbalanced crypto-currencies and financial stock exchange time series. Our solution is a combination of gradual pattern mining and graph pattern mining algorithms on time series.
- We propose a graph-based model for modeling of gradual patterns into undirected graphs. The benefits of graph-based model include: (1) Summarization and reduction of discovered patterns volume and storage in each crypto-currency. (2) Useful for mining crypto-currencies in order to discover the common and frequent trend across crypto-currency features.

[1] https://coinmarketcap.com

MiLeTS'18, 20 August, 2018, London, United Kingdom

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

- We propose a graph-based visualization for the discovered gradual patterns. The new technique visualizes co-variation patterns rules using vertices and edges where vertices typically represent co-variation features and edges indicate relationship in co-variation patterns.

Experiments results on multiple crypto-currencies across financial stocks demonstrate how our approach is useful for discovering insight co-variation patterns.

The rest of the paper is organized as follows. Section 2 presents an overview of dataset used for mining with their associated challenges and highlights research questions. Section 3 presents a new knowledge discovery approach composed with different steps. Section 4 presents experimental studies using blockchain of crypto-currency data and financial stock exchange time series. Section 5 discusses analysis techniques previously applied to a blockchain and crypto-currency data. In Section 6, we briefly discuss results, and future work.

2 DATASET OVERVIEW AND RESEARCH QUESTIONS

Coinmetrics tools [2] [3] have been used for crypto-currencies data collection and parsing. The tools are composed of a coin crawler and parser programs. Data collection across multiple sources is determinant to avoid biases in the results. Therefore, data from Bitcoin, Bitcoin Cash, Litecoin, Dash, PIVX, Monero, Dogecoin, Decred, XEM, Ethereum, Ethereum Classic, ZCash, and Vertcoin have been gathered since their creation. It is noteworthy to remind the imbalanced structure of the dataset since all time series have a different beginning date. Table 1 shows an example of two imbalanced time series of two crypto-currencies Bitcoin and Xem. The Bitcoin crypto-currency starts two years before the Xem crypto-currency.

Our problem of imbalanced time series is different from the class-imbalance problem in the sense that classification categories are not equally represented. In this context, the imbalanced crypto-curencies sources make the correlation and co-variation discovery more challenging.

Due to the high volatilities of crypto-currencies markets, our aim is to define metrics to better understand their movements. More especially, the following questions are addressed:

- How to extract correlation and co-variations patterns from multiple imbalanced time series?
- Are co-variations across all imbalanced time series equal?
- How to mine multiple co-variations metrics of crypto-currencies?
- How to represent and visualize co-variations metrics?

We propose an intra and inter crypto-currencies analysis approach based on pattern mining approaches.

3 KNOWLEDGE DISCOVERY APPROACH

In this section, we present our knowledge discovery approach. Our goal is twofold: (1) discovering frequent co-variations correlation for each crypto-currency and across multiple crypto-currencies. (2) discovering dependencies and frequent co-variations correlation between crypto-currency markets and stock exchanges.

[2]https://coinmetrics.io
[3]https://github.com/whateverpal/coinmetrics-tools

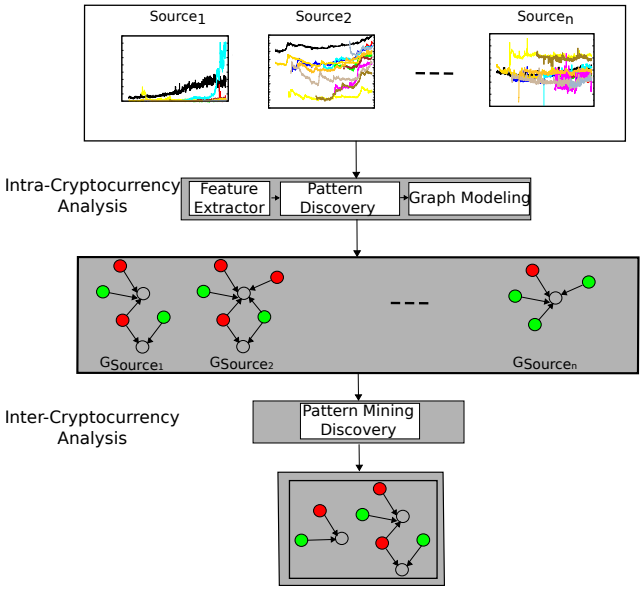


Figure 1: Overview of pattern discovery process from imbalanced multiple time series

3.1 Overview

Figure 1 represents an overview of the different components of the knowledge discovery process. Due to the imbalanced time series of crypto-currencies, our methodology is composed of two blocks of mining steps: intra and inter-crypto-currency. For each block, co-variations correlation are computed.

3.2 Feature extractor

The crypto-currency dataset is structured in 8 columns: date, transaction volume, transaction count, generated coins, and fees plus financial extra-metrics such as the market capitalization (USD), exchange volume (USD), and price (USD). We further explain the meaning of each of the features:

- Date: daily crypto-currency data.
- Transaction volume (USD): the sum of all transaction outputs belonging to the blocks mined on a given day.
- Transaction count: number of transactions in all blocks.
- Crypto-currency Market Capitalizations (USD): the market crypto-currency value in US dollar at a given day . It's calculated by multiplying the *Price* by the *Circulating Supply*. Circulating Supply is the best approximation of the outstanding number of coins.
- Price (USD): The price of a crypto-currency in US dollar. Price is calculated by taking the volume weighted average of all prices reported by each exchange platform.
- Exchange volume (USD): volume is defined as the amount of traded bitcoin on the exchange. For example, if 10 bitcoins is transacted between a maker and a taker, then the total exchange volume is 10 bitcoins.
- Generated coins: number of created coins.
- Fees: the total fees of transaction data.

	Bitcoin				...	Xem			
	Values		Blockchain		...	Values		Blockchain	
Time	Price(USD)	Fees	TxCount	GeneratedCoins		Price(USD)	Fees	TxCount	GeneratedCoins
2013-04-28	134.21	41.67	47031	3750	...				
2013-04-29	144.54	31.67	40035	4200	...				
...	...	...	...	...	...				
2015-04-01	247	15.35	110634	3750	...	0.000242	23721.28	533	0
2015-04-02	253	17.36	121155	3475	...	0.000314	30939.69	719	0

Table 1: Example of imbalanced multi-source data

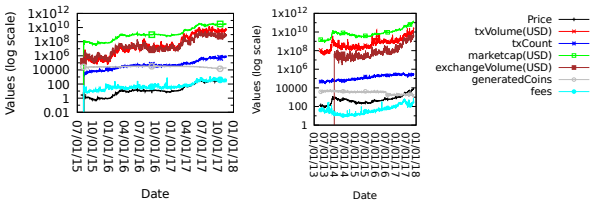
Figure 2 highlights the different features of imbalanced crypto-currencies. These features are the combination of the characteristics of blockchain crypto-currencies and their market values. The idea is to underline the dependency between the characteristics of blockchain and, its market, each individual crypto-currency and all combined crypto-currencies.

3.3 Intra crypto-currency correlation discovery

Due to the imbalanced crypto-currencies data, each crypto-currency is mined independently in order to discover co-variations correlation between their features.

3.3.1 ***Mining the frequent co-variation cryto-currency features.*** A dataset is a set of tuples D defined over the schema $S = \{I_1, \ldots, I_n\}$. Table 1 shows an example dataset where the schema of a bitcoin values and its blockchain is $S = \{Price(USD), Fees, TxCount, GeneratedCoins\}$.

A gradual item is a pair $(item, variation)$ of an item (attribute) $item \in S$ and a variation $variation \in \{\uparrow, \downarrow\}$, where $\uparrow$ stands for a positive (ascending) variation and $\downarrow$ for a negative (descending) variation. A gradual itemset is defined as a non-empty set of gradual attributes. For instance, the gradual itemset $I_1 = \{(Price(USD), \uparrow), (TxCount, \uparrow)\}$ means "*the higher the bitcoin price, the lower the fees*". $I = \{(item_{k_1}, variation_{k_1}), \ldots, (item_{k_i}, variation_{k_i})\}$ where all distinct $\{k_1, \ldots, k_i\} \subseteq \{1, \ldots, n\}$. Two tuples t and t' can be ordered with respect to I if all values of the corresponding items from the gradual itemset can be ordered to respect the respective variation: $t.item_{k_i} \leq t'.item_{k_i}$ if $variation_{k_i} = \uparrow$ and $t.item_{k_i} \geq t'.item_{k_i}$ if $variation_{k_i} = \downarrow$ where t precedes t' in the order induced by I. For instance, from Table 1, it can be seen that at time $t_1 = 2013-04-28$, and $t_2 = 2013-04-29$ can be order with respect to I_1 as $t_1.Price(USD) \leqslant t_2.Price(USD)$ AND $t_1.PriceTxCount \leqslant t_2.TxCount$.



(a) ETH **(b) BTC**

Figure 2: Crypto-currency features

Let $T = \{t_1, \ldots, t_n\}$ be a list of tuples from D and I be a gradual itemset. T respects I if $\forall\, i \in [1, n-1]$ we have $t_i.time < ti + 1.time$. Let T_I be the set of lists of tuples that respect I. The formal definition of the support (number of co-variations) of I is $support(I) = \frac{max_{T \in T_I(|T|)}}{|D|}$. This support means the size of the longest list of tuples that respect I. We note that the support of an itemset with a single item is always 100% as it is always possible to order all the tuples by one column [8]. We define a gradual pattern P is a itemset-value pair $(itemset, support(itemset))$.

For instance, in Table 1 the longest sequences of time series in bitcoin dataset that can be ordered according to I_1 are $\{2013-04-28, 2013-04-29, 2015-04-02\}$. Thus, $support(I_1) = \frac{3}{4} = 75\%$ meaning that 75% of the input time series can be ordered consecutively according to I_1.

Definition 3.1 (Frequent gradual patterns). Let a gradual pattern p is said to be frequent if $support(p) \geq min\_support$ where $min\_support$ is a minimum support threshold defined by a user.

The number of extracted frequent gradual patterns can be very huge, in order to reduce the number of patterns without loss of information, many studies define and use the *closed patterns*.

Definition 3.2 (Closed patterns). A gradual pattern p is said to be closed if there does not exist any p' such that $p \subset p'$ and $support(p) = support(p')$.

Table 2 shows an example of a crypto-currency dataset where $S = \{Time, Price(USD), Fees, GeneratedCoins\}$ and $T = \{t_1, t_2, t_3, t_4\}$ in the classical form of a database table. The transactions are the rows and the attributes the columns. In this example, S is taken from a set of crypto currency features used in experiments: date, transaction volume (USD), transaction count, crypto-currency market, capitalizations (USD), price (USD), exchange volume (USD), generated coins, fees.

Then, a state of the art closed frequent gradual itemset mining algorithm is applied. We use Para-Miner [17], a generic pattern mining algorithm for multi-core architectures. Para-Miner includes itemset, gradual itemset and graph mining algorithms. The resulting closed frequent co-variation patterns of each individual crypto-currency source are the patterns we are looking for.

Table 3 shows the frequent closed gradual patterns of a crypto-currency. The frequency has been fixed to a minimum support threshold of three transactions ($min\_support = 3$).

Due to imbalanced time series, we run multiple instance of gradual pattern mining algorithm on different crypto-currency sources.

TID	Time	Price (USD)	Fees	TxCount	Generated Coins
1	2013-04-28	134.21	41.67	47031	3750
2	2013-04-29	144.54	31.67	40035	4200
3	2015-04-01	247	15.35	110634	3750
4	2015-04-02	253	17.36	121155	3475

Table 2: Example dataset of a crypto-currency

Pattern ID	Pattern (Frequency)
1	~~Time↑ Price↑ (4/4)~~
2	~~Fees↑ (4/4)~~
3	~~TxCount↑ (4/4)~~
4	~~GenCoins↑ (4/4)~~
5	Time↑ Price↑ GenCoins↓ (3/4)
6	Time↑ Price↑ Fees↓ (3/4)
7	Time↑ Price↑ TxCount↑ (3/4)
8	TxCount↑ GenCoins↓ (3/4)
9	Time↑ Price↑ TxCount↑ GenCoins↓ (3/4)

Table 3: Frequent closed gradual patterns from Table 2.

Each crypto-currency source $crypto_i$ has input of a gradual pattern mining algorithm instance $inst_i \forall i \in [0, CS]$ where CS is the number of crypto-currency sources. Each instance $inst_i$ discovers gradual patterns $patterns_i = \{p_{i_1}, \cdots, p_{i_j}\}$ where j is the number of patterns in an instance $inst_i$. Thus, we obtain a set of frequent closed gradual pattern from each crypto-currency source.

3.3.2 ***Graph-based patterns modeling and visualization.*** Given a set of patterns from a crypto-currency source, the goal of this step is to propose a graph model for representing the discovered patterns for mining and visualization. The graph is used for visualization of the frequent gradual patterns results of each crypto-currency instance as well as for mining multiple crypto-currency sources. The main advantage of using graph data structure is its ability to reduce the number of redundant items in patterns. The graph uses an item as a node exactly once rather than two or more times as was done in output gradual mining algorithm. There are two main steps to construct co-variations graph for frequent gradual patterns discovered and generated from an instance: the post-processing step of an instance, and the co-variations graph construction step.

Post processing. This step consists in filtering the discovered patterns based on constraints provided as inputs. The filter is to remove individual patterns with a single item because their supports are always 100% [8]. Then, we focus on patterns with two or more items.

Co-variations graph. To characterize dependency relations between an item and the support in a pattern, we introduce a notion of *item-dependency rule*. We also introduce *co-variations graph* as an intuitive graph representation for item-dependency rules.

Definition 3.3 (Item-dependency rule). Let P={$itemset$, $support$ ($itemset$) } be a gradual pattern where $itemset = \{(item_{k_1}, variation_{k_1}), \ldots, (item_{k_i}, variation_{k_i})\}$ where all distinct $\{k_1, \ldots, k_i\} \subseteq \{1, \ldots, n\}$. An item-dependency rule $(item_{k_i}, variation_{k_i}) \rightarrow support(itemset)$ denotes a dependency on $support(itemset)$ of pattern P.

Intuitively, an item-dependency rule indicates that the occurrence of $(item_{k_i}, variation_{k_i})$ in pattern P is $support(itemset)$ if and only if $(item_{k_i}, variation_{k_i})$ occurs in P.

Table 4 shows an example of item-dependency rules. Every item in a frequent pattern is converted into an item-dependency rule with keeping the pattern identifier associated to a rule identifier. The identifiers are useful for maintaining the traceability of the item in the pattern.

Rule ID	Pattern ID	Item-dependency rule
1	5	Time↑ → (3/4) Price↑ → (3/4) GenCoins↓ → (3/4)
2	6	Time↑ → (3/4) Price↑ → (3/4) Fees↓ → (3/4)
3	7	Time↑ → (3/4) Price↑ → (3/4) TxCount↑ → (3/4)
4	8	TxCount↑ → (3/4) GenCoins↓ → (3/4)
5	9	Time↑ → (3/4) Price↑ → (3/4) TxCount↑ → (3/4) GenCoins↓ → (3/4)

Table 4: Item-dependency rules of Table 3

Definition 3.4 (Star of rules). Let P=itemset, support(itemset) and $R = \{(item_{k_1}, variation_{k_1}) \rightarrow support(itemset), \ldots, (item_{k_i}, variation_{k_i}) \rightarrow support(itemset)\}$ be a gradual pattern and its item-dependency rules, respectively. A star of rules SR is an attributed, single-level, rooted tree which can be represented by a 3-tuple $SR = (r, \delta, alpha)$, where r is the root vertex that represents the $support(itemset)$, δ is the set of leaves $(item_{k_i}, variation_{k_i})$, and α is a labeling function of a root r where each rule has its own root r. Edges exist between r and any vertex in δ and no edge exists among vertices in δ.

Thus, a pattern is represented by star of rules. The union of two stars of rules $SR = (r, \delta, \alpha)$ and $SR' = (r', \delta', \alpha')$ is the union of their vertex sets and their edge. Which means $SR \cup SR' = (r \cup r', \delta \cup \delta', \alpha \cup \alpha')$. The union of stars of rules is called *co-variations graph.*

Table 5 shows stars of rules constructed from Table 4. The circle vertex with the green color represents an item with its the symbol variation ↑ having a positive (ascending) variation. In contrast, the vertex with the red color represents an item with is symbol variation ↓ having a negative (descending) variation, and the rectangle vertex represents the support threshold of a pattern linked with circle vertices.

Definition 3.5 (Co-variations graph). Let SR be a set of stars of rules. A co-variations graph over a set of items with their variations

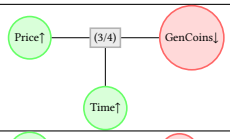
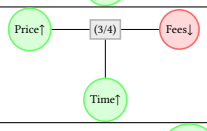
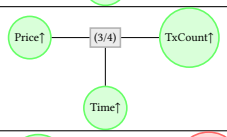
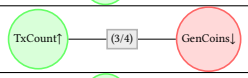
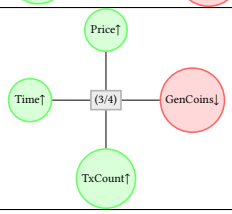
Rule ID	Pattern ID	Stars of rules
1	5	
2	6	
3	7	
4	8	
5	9	

Table 5: Stars of rules

V is a undirected graph $G = (V, E, \beta) = \bigcup_{i=1}^{\Omega} SR_i$ where Ω is the number of patterns discovered in an instance:

- V is the set of vertices in G, where each vertex $v \in V$ is an item and a support of a pattern from V. The same items having the same labels share the same vertex identifier. For each support of a pattern has a unique vertex identifier.
- E is a set of edges in G. Let $u = (item_{k_i}, variation_{k_i})$ and $v = support(itemset)$ be two vertices. There is an edge $(u, v) \in E$ if and only if there exists a item-dependency rule $u \rightarrow v$ in a star of rules.
- β is a labeling function of items and supports of patterns.

Figure 3 shows an example of a co-variations graph. All vertices sharing the same labels are merged int the same vertex except the supports of patterns, where each support has its own vertex identifier. This condition is necessary for keeping a traceability of patterns and do not change the patterns interpretation.

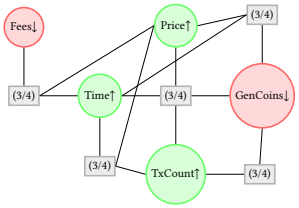


Figure 3: Co-variations graph

3.4 Inter crypto-currency correlation discovery

Every crypto-currency source has its own co-variations graph which constitute a graph database of co-variations graphs. Through graph mining, the frequent co-variations subgraphs across crypto-currencies are discovered. The frequent co-variations subgraphs allow to obtain the common co-variations features and correlation among a set of crypto-currencies. Mining these graphs allows to extract the frequent subgraph patterns describing the common co-variation in crypto-currencies. The advantage is to mine co-variation of gradual patterns discovered from individual imbalanced crypto-currency sources. The data mining technique used to discover the frequent subgraphs is called *frequent subgraph mining algorithm* [22, 23]. The frequent subgraphs represent the frequent co-variations in crypto-currencies. Given D_g is the co-variations graph database, the frequent subgraphs mining aims to mine co-variations graphs with more support value in comparison with predefined minimum support threshold [22, 23].

Definition 3.6 (Subgraph). Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs. G_2 is a subgraph of G_1 iff $V_2 \subseteq V_1$ and $E_2 \subseteq E_1$. In such a case, we also say that G_1 is a a supergraph of G_2.

To determine the frequency of a subgraph in a graph dataset, it is necessary to mine subgraphs that are isomorphic with an existing graphs in a graph dataset.

Definition 3.7 (Subgraph Isomorphism). Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs. A subgraph isomorphism from G_1 to G_2 is a function $f : V_1 \rightarrow V_2$ such that if $(u, v) \in E_1$, then $(f(u), f(v)) \in E_2$ and the labels should be added to the mapping. In addition, $(u, v) \notin E_1$, then $(f(u), f(v)) \notin E_2$. f is an induced subgraph isomorphism.

Definition 3.8 (Induced subgraph). Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ be two graphs. A subgraph $G_1 \subseteq G_2$ is an induced subgraph, if $E_1 = E_2 \cap E(V_1)$. In this case G_1 is induced by its set V_1 of vertices.

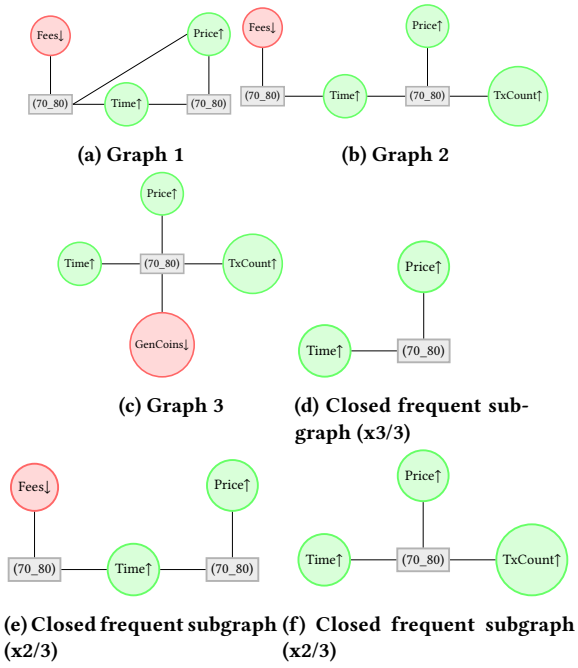
Definition 3.9 (Frequent Subgraph). Let $D_g = G_1, G_2, \ldots, G_m$ and $min\_support_{D_g}$ be a labeled undirected graph dataset and minimum support threshold in graph dataset D_g. A frequent subgraph is a graph whose support is not less than a minimum user-specified support threshold $min\_support_{D_g}$. The support or frequency of a subgraph G_s is the percentage or number occurrences of subgraphs in D_g.

The graph support G_s is denoted by $Sup(G_s)$ and is given as:

$$Sup(G_s) = \frac{\sum_{i=1}^{n} G_i}{|D_g|} \tag{1}$$

The support of frequent subgraph G_s is the number of graphs in D_g induced G_s by the total number of graphs in D_g.

Figure 4 shows an example of a co-variation graph dataset D and a closed frequent co-variation subgraph with a minimum support $min\_support_{D_g} = 2/3$. We discretized continuous numeric supports into bins of numeric frequency intervals in order to regroup co-variation patterns having close frequency values. As an example, the frequency values of patterns are discretized by bins of 10 from 0% to 100%. It means all frequency values between 0% and 10% are represented by the bin 0_10.



(a) Graph 1 (b) Graph 2

(c) Graph 3 (d) Closed frequent subgraph (x3/3)

(e) Closed frequent subgraph (x2/3) (f) Closed frequent subgraph (x2/3)

Figure 4: Co-variations graphs (4a, 4b, 4c) and closed frequent co-variation subgraphs 4d, 4e, 4f among them.

After building the graphs collection from different crypto-currency sources, we can use a state of the art frequent subgraph mining algorithm. The frequent subgraph mining has attracted much attention in the graph mining community with various efficient algorithms developed such as FSG [12], gSpan [22], CloseGraph [23], Gaston [1], and ParaMiner [17]. These techniques aim to identify the frequently occurring subgraph patterns from a given collection of graphs. For discovering the frequent co-variations of multiple crypto-currency sources, we use ParaMiner [17] as the intra crypto-currency methodology.

4 EXPERIMENTAL RESULTS AND ANALYSIS

In this section, we show four sets of experimental studies using blockchain of crypto-currency data and financial stock exchange data. With the aim to fully discover knowledge, the goal is to answer to the questions asked in Section 1: (1) In the first set of experiments, we extract patterns related to each crypto-currency using the intra-crypto-currency correlation discovery method. (2) In the second set of experiments, we extract patterns by mining co-variation patterns across crypto-currencies using the inter crypto-currency correlation discovery method. (3) In the third set of experiments, we mine the prices of crypto-currencies with financial stock exchanges: SX5E, HSCEI, NIKKEI, NASDAQ, DOW JONES. (4) In the fourth set of experiments, we mine the prices of two popular crypto-currencies with the results of a search on Google Trends [4] with the keywords Bitcoin and Ethereum. The objective here is to highlight the relationships between Bitcoin and Ethereum prices and their popularity on the web and social media worldwide.

[4]https://trends.google.com/trends

4.1 Experimental Setup

We collect two datasets from https://coinmetrics.io [5] and https://finance.yahoo.com for blockchain crypto-currencies and financial stock exchanges which contain daily data from the beginning of each crypto currency until January 01, 2018.

4.2 Discovered intra-crypto-currency results

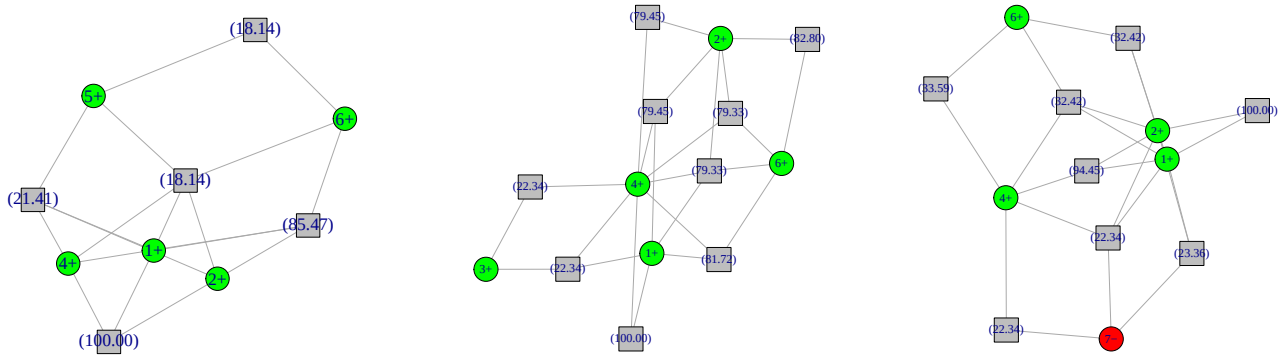
In this set of experiments, we extract the closed frequent gradual patterns with a minimum support threshold $min\_support = 10\%$. The patterns are represented in an encoded graph. We visualize a co-variations graph using three types of vertices as shown in Figure 5. This figure represents a frequent gradual patterns visualization extracted from each crypto-currency source based on the proposed scheme. The circle vertex with the green color and the symbol "+" represents a positive (ascending) variation ↑. The vertex with the red color and the symbol "-" represents a negative (descending) variation ↓. The rectangle vertex represents the support of a pattern linked with circle vertices. Figure 5 shows co-variation frequent patterns of different crypto-currencies. For instance, among the frequent co-variations patterns, we find the following patterns. In Bitcoin, when the transaction volume increases, the exchange volume and crypto-currency market capitalization increase too, occurring in 85.47% of the cases. Another frequent pattern is a high dependency between the transaction volume and crypto-currency market capitalization and occurring daily in 100%. A low dependency between the Bitcoin price and its features: crypto-currency market capitalization, transaction volume, and exchange volume. We find the same patterns between Bitcoin and Ethereum except for the number of co-occurrences which is very low. We notice the majority patterns in Bitcoin and Ethereum have positive (ascending) variation as well as frequent and infrequent co-variation patterns. In XMR crypto-currency, we find a less frequent co-variation patterns with a negative (descending) variation of generated coins where the lower generated coins, the higher crypto-currency market capitalization and transaction volume occurring in 22.34% of the cases.

4.3 Discovered inter-crypto-currency results

In this experiment, based on the discovered co-variations in each crypto-currencies source, we extract the closed frequent subgraphs across co-variations graphs with a minimum support threshold $min\_support_{D_g} = 50\%$.

We mine 8 crypto-currencies co-variation graphs discovered in last experiments 4.2 : ZEC, XMR, VTC, PIVX, LTC, DCR, ETH, BTC. To be able to use the graph mining algorithm, we discretized continuous numeric frequency of patterns in co-variation graph into bins of numeric intervals in order to regroup patterns exhibiting similar values and belonging to the same interval. As an example, the frequencies are discretized by bins of 10 from 0 to 100 *i.e.* all frequencies between 0 and 10 are represented by *(0_10)*. Figure 6 shows a frequent co-variations pattern across crypto-currencies. We find that 92% of crypto-currencies have the same co-variation patterns especially a high dependency between the transaction volume, crypto-currency market capitalizations and exchange volume

[5]https://github.com/whateverpal/coinmetrics-tools



(a) BTC (b) ETH (c) XMR

Figure 5: Co-variation frequent patterns of different crypto-currencies.
1: Date, 2: Transaction volume(USD), 3: Transaction count, 4: Crypto-currency Market Capitalizations(USD), 5: Price(USD), 6: Exchange volume(USD), 7: Generated coins, 8:fees

and a low dependency the price and the rest of features. We notice that the majority of crypto-currencies have the same characteristics and evolve in the same direction. As the last experiment, all variations are positive.

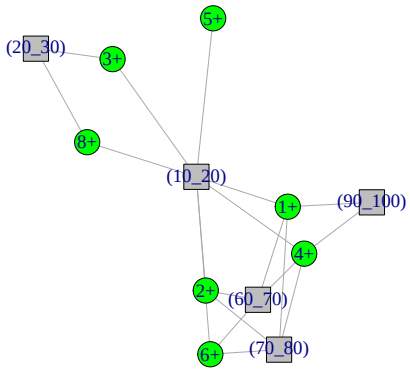


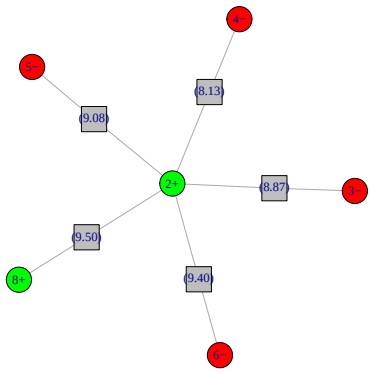
Figure 6: Frequent co-variations pattern across crypto-currencies

4.4 Discovered co-variation patterns between crypto-currencies and financial stock exchanges

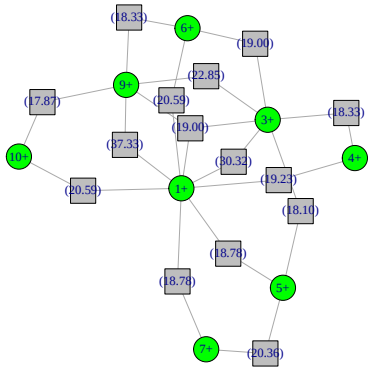
In this experiment, we mine the prices of two popular crypto-currencies Bitcoin and Ethereum with financial stock exchanges. For this, we use the same gradual pattern mining algorithm as experiment 4.2 with a minimum support threshold $min\_support = 8\%$. Figure 7 shows two co-variation patterns across financial stock exchanges and crypto-currency prices. We discover infrequent patterns with a low frequency. In the first pattern, Figure 7a, the higher exchanges euro-dollar, the lower DJI, HSCE, and NDAQ, the higher gold with a frequency between 8% and 9%. In the second pattern, Figure 7b, no high co-variation dependency between financial stock exchanges and Bitcoin and Ethereum prices have been identified. A low dependency exists between the Bitcoin and Ethereum prices (17.87%).

4.5 Discussion

Our knowledge discovery approach from imbalanced time series provides insightful patterns by extracting frequent or infrequent co-variations in crypto-currencies features as well as across the thousand of crypto-currencies and financial stock exchanges. Our tool helps the financial experts to highlight co-variations patterns and the impact on financial markets. In the experiments, we saw, first, the ascendant and positive co-variations of crypto-currencies features: prices, number of transactions, generated coins, market capitalization. Secondly, almost all crypto-currencies have the same positive co-variations features and having the same behavior. Thirdly, knowing that the important feature in cypto-currencies is the price, from our experiments, we have found no high dependency between the price and blockchain features or financial stock exchanges.



(a) Pattern 1



(b) Pattern 2

Figure 7: Co-variation patterns across financial stock exchanges and crypto-currency prices.
1:Date; 2:exchange-EUR-USD; 3:DJI; 4:HSCE; 5:N225; 6:NDAQ; 7:STOXX50E; 8:GOLD; 9:BTC; 10:ETH

Thus, the price is completely disconnected from the rest of features. Interesting challenge is to take into consideration the influence of political events and extra-events on cypto-currencies prices which can be considered as a future work. From our approach and experiments, the financial experts and investors can obtain answers to the questions asked in 1. Moreover, the proposed scheme provides a meaningful visualization even if the number of patterns is very large. The amount of visualized patterns is independent of the number of selected itemsets in input.

5 RELATED WORK

Recently, many studies and works on blockchain data analytics have been published [3], addressing clustering transactions [10, 19], financial frauds [14, 15, 20], predicting interactions transactions [5] or studying denial-of-service attacks [21]. Most of works focus on one or two crypto-currencies analysis such as Bitcoin or Ethereum. However, no multiple analysis have been provided between the crypto-currencies, their blockchains with the different currencies and financial stock markets.

Crypto-currency analysis. In [13], the authors studied the relationship between Bitcoin, and search queries on Google Trends and Wikipedia. They discovered that the search queries and the prices strongly connected; as the price of Bitcoin increases which the interest of investors as well as a general public. In [18], the authors proposed an anomaly detection approach to detect users and transactions are the most suspicious the Bitcoin transaction network. They used unsupervised learning methods based on graphs generated by the Bitcoin transaction network. In [3], the authors proposed a methodology to create custom analytics for the Bitcoin blockchain by constructing a blockchain data as a collection in a NoSQL database and analyzing the collection by using the query language of the database. Other types of blockchain modeling are presented, for instance, in [11], the authors presented GraphSense, a graph-based model for blockchain transactions and crypto-currencies analytics. In [2], the authors studied the characteristics of blockchains by investigating the blockchains activity using statistics tools in order to identify some curiosities in the blockchains.

Time series analysis based on data mining. Many works on time series mining are based on transformation of numeric values into items, and sequences. For instance, in [6], the authors proposed a pattern discovery process from multiple time series based on temporal logic. They transform multiple time series into multiple event sequences and mine the frequent subsequences across the event sequences. In [7] and [9], the authors proposed a review of existing approaches on mining time series. They showed that the existing approaches related to mining time series are generally based on time series transformation for clustering, classification, similarity measures, subsequence matching and subsequence mining. The authors highlighted the issue and the need to handle multi-attributes time series, mining on time series data stream.

The originality of our approach compared to the related works is threefold. Firstly, we address the practical problem of conducting co-occurrence variation mining in a time series for imbalanced data. Secondly, it relies on a completely automatic combination of data mining approaches that find co-variation metrics across multiple currencies and crypto-currencies, and presents the frequency of the co-variations. Thirdly, we provide a graph-based visualization co-variation metrics. All these originalities allow investors, financial experts, scientists, and all public to understand the different process and the behavior of crypto-currencies and blockchain.

6 CONCLUSION AND FUTURE WORK

In this paper, we present a knowledge discovery and extraction approach from crypto-currencies of multiple imbalanced time series.

Our approach is based on gradual pattern mining and graph mining algorithms. It excels to discover co-variations patterns across crypto-currencies and financial stock exchanges. However, to the best of our knowledge, none of existing methods of mining crypto-currencies data and looking for relationships between the discovering crypto-currencies, their blockchain and financial stock exchanges. Our results advocate that such approaches are of interest. It allows financial experts and investors to have deeper analysis of crypto-currencies markets, and a better understanding of its behavior. Experiment results on the public blockchain, crypto-currencies, and financial stock exchanges demonstrate that our approach can provide strong insight co-variation patterns.

REFERENCES

[1] 2005. The Gaston Tool for Frequent Subgraph Mining. *Electronic Notes in Theoretical Computer Science* (2005), 77 – 87. Proceedings of the International Workshop on Graph-Based Tools (GraBaTs 2004).
[2] Luke Anderson, Ralph Holz, Alexander Ponomarev, Paul Rimba, and Ingo Weber. 2016. New kids on the block: an analysis of modern blockchains. *CoRR* (2016).
[3] Massimo Bartoletti, Andrea Bracciali, Stefano Lande, and Livio Pompianu. 2017. A general framework for Bitcoin analytics. *CoRR* abs/1707.01021 (2017).
[4] J. Bonneau, A. Miller, J. Clark, A. Narayanan, J. A. Kroll, and E. W. Felten. 2015. SoK: Research Perspectives and Challenges for Bitcoin and Cryptocurrencies. In *2015 IEEE Symposium on Security and Privacy*. 104–121.
[5] Jérémy Charlier, Sofiane Lagraa, Radu State, and Jérôme François. 2017. Profiling Smart Contracts Interactions Tensor Decomposition and Graph Mining. In *Proceedings of the Second Workshop on MIning DAta for financial applicationS (MIDAS 2017) co-located with the 2017 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD 2017), Skopje, Macedonia, September 18, 2017*. 31–42.
[6] Z. Chen, B. r. Yang, F. g. Zhou, L. n. Li, and Y. f. Zhao. 2008. A New Model for Multiple Time Series Based on Data Mining. In *2008 International Symposium on Knowledge Acquisition and Modeling*. 39–43.
[7] Tak chung Fu. 2011. A review on time series data mining. *Engineering Applications of Artificial Intelligence* (2011), 164 – 181.
[8] Trong Dinh Do, Alexandre Termier, Anne Laurent, Benjamin Negrevergne, Behrooz Omidvar-Tehrani, and Sihem Amer-Yahia. 2015. PGLCM: Efficient Parallel Mining of Closed Frequent Gradual Itemsets. *Knowl. Inf. Syst.* 43, 3 (2015), 497–527.
[9] Philippe Esling and Carlos Agon. 2012. Time-series Data Mining. *ACM Comput. Surv.* (2012), 12:1–12:34.
[10] M. Harrigan and C. Fretter. 2016. The Unreasonable Effectiveness of Address Clustering. In *IEEE Conferences on Ubiquitous Intelligence Computing, (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*. 368–373.
[11] Bernhard Haslhofer, Roman Karl, and Erwin Filtz. 2016. O Bitcoin Where Art Thou? Insight into Large-Scale Transaction Graphs. In *SEMANTiCS*.
[12] Michihiro Kuramochi and George Karypis. 2001. Frequent Subgraph Discovery. In *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM '01)*. 313–320.
[13] Kristoufek L. 2013. BitCoin meets Google Trends and Wikipedia: Quantifying the relationship between phenomena of the Internet era. *Scientific Reports 3* doi:10.1038/srep03415 (2013).
[14] Malte Möser, Rainer Böhme, and Dominic Breuker. 2014. *Towards Risk Scoring of Bitcoin Transactions*. 16–32.
[15] M. MÃűser, R. BÃűhme, and D. Breuker. 2013. An inquiry into money laundering tools in the Bitcoin ecosystem. In *2013 APWG eCrime Researchers Summit*. 1–14.
[16] Satoshi Nakamoto. 2008. Bitcoin: A peer-to-peer electronic cash system,âĂİ http://bitcoin.org/bitcoin.pdf. (2008).
[17] Benjamin Négrevergne, Alexandre Termier, Marie-Christine Rousset, and Jean-François Méhaut. 2014. Para Miner: a generic pattern mining algorithm for multi-core architectures. *Data Min. Knowl. Discov.* 28 (2014), 593–633.
[18] Thai Pham and Steven Lee. 2016. Anomaly Detection in Bitcoin Network Using Unsupervised Learning Methods. *CoRR* abs/1611.03941 (2016).
[19] Michele Spagnuolo, Federico Maggi, and Stefano Zanero. 2014. *BitIodine: Extracting Intelligence from the Bitcoin Network*. Springer Berlin Heidelberg, Berlin, Heidelberg, 457–468.
[20] Marie Vasek and Tyler Moore. 2015. *There's No Free Lunch, Even Using Bitcoin: Tracking the Popularity and Profits of Virtual Currency Scams*. 44–61.
[21] Marie Vasek, Micah Thornton, and Tyler Moore. 2014. *Empirical Analysis of Denial-of-Service Attacks in the Bitcoin Ecosystem*. 57–71.
[22] Xifeng Yan and Jiawei Han. 2002. gSpan: Graph-Based Substructure Pattern Mining. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM '02)*. IEEE Computer Society.
[23] Xifeng Yan and Jiawei Han. 2003. CloseGraph: Mining Closed Frequent Graph Patterns. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '03)*. 286–295.