PAKDD : Forecasting Bitcoin Price with Graph Chainlets

Cuneyt G. Akcora, Asim Kumer Dey, Yulia R. Gel, and Murat Kantarcioglu

University of Texas at Dallas, Richardson, USA
{cuneyt.akcora,adey,ygl,muratk}@utdallas.edu

Abstract. Over the last couple of years, Bitcoin cryptocurrency and the Blockchain technology that forms the basis of Bitcoin have witnessed a flood of attention. In contrast to fiat currencies used worldwide, the Bitcoin distributed ledger is publicly available by design. This facilitates observing all financial interactions on the network, and analyzing how the network evolves in time. We introduce a novel concept of chainlets, or Bitcoin subgraphs, which allows us to evaluate the local topological structure of the Bitcoin graph over time. Furthermore, we assess the role of chainlets on Bitcoin price formation and dynamics. We investigate the predictive Granger causality of chainlets and identify certain types of chainlets that exhibit the highest predictive influence on Bitcoin price and investment risk.

1 Introduction

Bitcoin cryptocurrency [17] has seen tremendous interest and has achieved skyrocketing adoption over the last couple of years. The bitcoin phenomenon is due not only to revolutionizing online payments but also to a big number of applications the underlying blockchain technology has witnessed in various domains [21].

One interesting aspect of Bitcoin is that a distributed ledger (i.e., blockchain) is maintained by all the participants to verify the authenticity of each Bitcoin transaction. The existence of such a distributed ledger creates unique opportunities with respect to graph analysis. Already, different applications have used the distributed ledger and the Bitcoin graph information to track sex trafficking [19] and money laundering activity [16].

We believe that the Bitcoin graph can be used for interesting off-the-beaten track applications. For instance, in most stock analysis platforms, the market trend is usually predicted by using historical prices and other financial and economic indicators only, without accounting for financial network structure effects. Since we can observe the complete Bitcoin graph, a natural question to ask is whether the *local graph structure* impacts the price of an asset (e.g., Bitcoin). In other domains, local higher-order structures of complex networks, or multiple-node subgraphs, are found to be an indispensable tool for analysis of network organization beyond the trivial scale of individual vertices and edges. The core

idea is that if a particular subgraph occurs more or less frequently than the expected baseline occurrence, then such a subgraph is likely to play an important role in network functionality.

Furthermore, structural properties of multiple complex networks can be compared in terms of their (dis)similarities in subgraph patterns. The role of small subgraphs, or network *motifs* and *graphlets*, in organization of complex systems has been first discussed in conjunction with the assessment of stability and robustness of biological networks [15], and later have been studied in a variety of contexts, from social networks to power grids (for overviews see [1] and references therein). Most recently, network motifs are shown to provide an invaluable insight into analysis of functionality and early warning stability indicators in financial networks [9]. However, compared to biological networks, motif-induced inference in financial systems is still an emerging field, and there yet exist no studies on the role of motifs in the analysis of blockchain.

To our knowledge, we are the first to address the impact of local topological structures/motifs on Bitcoin price. We can summarize our contributions as follows:

- We introduce and formalize the notion of *chainlet* motifs to understand the impact of local topological structures on Bitcoin price dynamics.
- We develop techniques to understand which local topological structures (i.e., chainlets) have a higher impact on the price dynamics and use those "important" chainlets for price prediction.
- We compare our techniques to the state of art time series analysis approaches and show that employing chainlets leads to more competitive price prediction mechanisms.

The remainder of this paper is organized as follows: In Section 2, we discuss the related work. In Section 3, we formally define chainlets using a generalized heterogeneous graph model. In Section 4 we compare the price prediction models that use chainlets to other existing models to see the impact of chainlets on price. Finally, in Section 5, we conclude with the summary of our results.

2 Related Work

Since the seminal Bitcoin paper [17] in 2008, digital coins [21] have been the most prominent Blockchain applications. Among these, Bitcoin has been the main focus of Blockchain analysis (see [2] for a review).

The earliest studies focused on the transaction graph to locate the coins used in illegal activities, such as money laundering and blackmailing [3, 18], which is known as the taint analysis [5]. Moser et al. [16] analyzed the opportunities and limitations of anti-money laundering on Bitcoin by looking at how successive transactions are used to transfer money.

The Bitcoin network itself has also been studied from multiple aspects. For instance, [4] analyzed centralities, and [13] found that since 2010 the Bitcoin network can be considered a scale-free network. Furthermore, [12] tracked the

evolution of the Bitcoin transaction network, and modeled degree distributions with power-laws. Although these studies analyzed the Bitcoin graphs, the primary focus was on global graph characteristics. In turn, our *chainlet analysis* sheds light onto local topological structures of Bitcoin and their role on price formation.

A number of recent studies show the utility of global graph features to predict the price [11, 7, 14]. For instance, [20] analyzed the predictive effects of average balance, clustering coefficient, and number of new edges on the Bitcoin price. Two network flow measures were recently proposed by [23] to quantify the dynamics of the Bitcoin transaction network and to assess the relationship between flow complexity and Bitcoin market variables. Furthermore, [14] identified 16 features for 30, 60 or 120 minute intervals and used Random Forest models to predict the price. The core idea behind all these approaches is to extract certain global network features and to employ them for predictions. On the other hand, chainlets provide a finer grained insight at the network transactions. In practice, chainlets can be used to refine the above-mentioned models, so that features are computed on selected subgraphs only. Furthermore, network flows can be detailed in terms of successive chainlets.

3 Methodology

The Bitcoin graph has three main components: *addresses*, *transactions* and *blocks*. A transaction is a transfer of bitcoins from input addresses to output addresses. Figure 1 shows such a network for 4 transactions and 13 addresses.

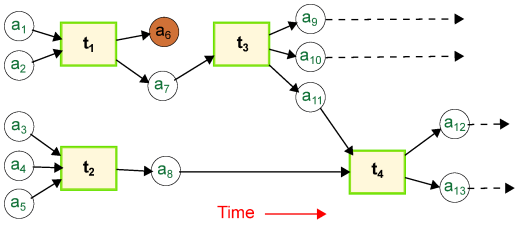


Fig. 1: A transaction-address graph representation of the Bitcoin network. Addresses and transactions are shown with circles and rectangles, respectively. An edge indicates a transfer of coins. The coins at address a_6 are unspent.

Our Bitcoin data come from the official Bitcoin software; we installed the Bitcoin core wallet [1] and had the wallet download the entire Bitcoin history from 2009 to 2018. Afterwards, we parsed the Bitcoin blockchain files, and extracted blocks, transactions and addresses. The source code of our Spark project is available on our Github repository. [2]

We model the Bitcoin graph as the following heterogeneous network with two node types: addresses and transactions.

[1] https://bitcoin.org/en/download
[2] https://github.com/cakcora/coinworks

The Bitcoin Graph Model. The Bitcoin network is a directed graph $\mathcal{G} = (V, E, B)$ where V is a set of vertices, and $E \subseteq V \times V$ is a set of edges. $B = \{\textbf{Address}, \textbf{Transaction}\}$ represents the set of vertex types. For any vertex $u \in V$, it has a vertex type $\phi(u) \in B$. For each edge $e_{u,v} \in E$ between adjacent nodes u and v, we have $\phi(u) \neq \phi(v)$, and either $\phi(u) = \{\textbf{Transaction}\}$ or $\phi(v) = \{\textbf{Transaction}\}$. That is, an edge $e \in E$ represents a coin transfer between an address node and a transaction node. This heterogeneous graph model subsumes the homogeneous case (i.e., $|B| = 1$), where only transaction or address nodes are used, and edges link vertices of the same type. In this paper, we focus on the case where each address node is linked (i.e., input or output address of a transaction) via a transaction node to another address node.

We emphasize three graph rules that shape the actual Bitcoin graph. First, input coins from multiple transactions can be merged and spent in a single transaction (as in transaction t_4 in Fig. 1). Second, in a Bitcoin transaction the input-output address mappings are not explicitly recorded. For instance, consider the transaction t_1 in Fig. 1. The output to address a_6 may come from either a_1 or a_2. Third, coins from multiple input transactions can be spent separately, but those received from one transaction must all be spent in a single transaction. Any amount that is not transferred is considered to be the transaction fee, and gets collected by the miner who creates the block. For this reason, unless it specifies itself as output address again, an address cannot transfer some bitcoins from a previous transaction and keep the change. As a community practice, this address reuse is discouraged, hence most nodes appear in the graph two times; once when they receive coins and once when they spend it. See [2] for a detailed graph representation of Blockchain.

Blocks order transactions in time, whereas each transaction with its input and output nodes represents an immutable decision that is encoded as a subgraph on the Bitcoin network. Rather than using individual edges or nodes, we chose to use this subgraph as the building block in our Bitcoin analysis. We use the term **chainlet** to refer to such subgraphs.

Our choice is due to two reasons. First, the subgraph can be taken as a single data unit because inclusion of nodes and edges in it is based on a single decision. As a transaction is immutable, joint inclusion of input/output nodes in its subgraph cannot be changed afterwards. This is unlike the case on a social network where nodes can become closer on the graph because of actions of their neighbors. Second, we argue and prove that subgraphs have distinct shapes that reflect their role in the network, and we can aggregate these roles to analyze network dynamics.

3.1 Graph Chainlets

We introduce the concept of k-chainlets to assess local higher order topological structure of the Bitcoin graph.

The k-Chainlet Model A Bitcoin subgraph $\mathcal{G}' = (V', E', B)$ is a *subgraph* of $\mathcal{G}$, if $V' \subseteq V$ and $E' \subseteq E$. If $\mathcal{G}' = (V', E', B)$ is a subgraph of $\mathcal{G}$ and E' contains

all edges $e_{u,v} \in E$ such that $(u, v) \in V'$, then G' is called an *induced* subgraph of G. Two graphs $\mathcal{G}' = (V', E', B)$ and $\mathcal{G}'' = (V'', E'', B)$ are called *isomorphic* if there exists a bijection $h : V' \to V''$ such that all node pairs u, v of G' are adjacent in G' if and only if u and v are adjacent in G''.

Let k-chainlet $\mathcal{G}_k = (V_k, E_k, B)$ be a subgraph of $\mathcal{G}$ with k nodes of type {**Transaction**}. If there exists an isomorphism between $\mathcal{G}_k$ and $\mathcal{G}'$, $\mathcal{G}' \in \mathcal{G}$, we say that there exists an *occurrence*, or *embedding* of $\mathcal{G}_k$ in $\mathcal{G}$. If a $\mathcal{G}_k$ occurs more/less frequently than expected by chance, it is called a blockchain *k-chainlet*. A *k-chainlet signature* $f_{\mathcal{G}}(\mathcal{G}_k)$ is a number of occurrences of $\mathcal{G}_k$ in $\mathcal{G}$.

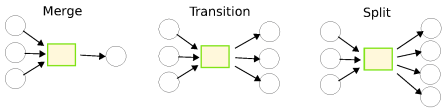


Fig. 2: Merge ($\mathbb{C}_{3\to1}$), Transition ($\mathbb{C}_{3\to3}$) and Split ($\mathbb{C}_{3\to4}$) chainlets for 3 inputs.

We start by focusing on the 1-chainlet signatures and their properties. For simplicity, we refer to *1-chainlets as chainlets*. A natural classification of chainlets can be made in terms of the number of inputs x and outputs y since there is only one transaction involved.

For a chainlet, we denote $\mathbb{C}_{x\to y}$ if it has x inputs and y outputs. If the branch is merging with other branches, the corresponding chainlet will have a higher number of inputs, compared to outputs. We call these **merge** chainlets, i.e., $\mathbb{C}_{x\to y}$ such that $x > y$, which show an aggregation of coins into fewer addresses. Two other classes of chainlets are **transition** and **split** chainlets with $x = y$ and $x < y$, respectively, as shown in Fig. 2. In what follows, we refer to these three chainlet types as the **aggregate chainlets**.

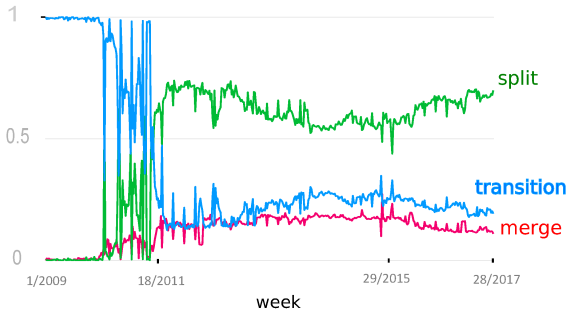


Fig. 3: Percentage of aggregate chainlets in weeks. Splits constitute around 60% of all transactions.

Fig. 3 visualizes the percentage of aggregate chainlets in time. For example, the transition chainlets are those $\mathbb{C}_{x\to x}$ for $x \geq 1$. Fig. 3 shows that starting as an unknown project, the Bitcoin network stabilized only after summer 2011. From 2014 and onwards, the split chainlets continued to steadily rise, compared to merge and transition chainlets.

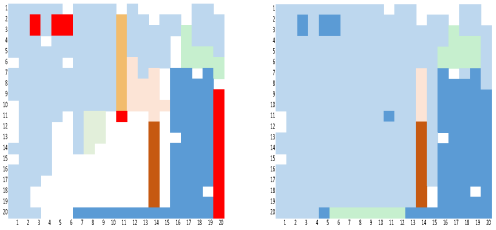
3.2 Clustering Chainlets

The Bitcoin protocol restricts numbers of input and output addresses in a transaction by putting a limit on the block size (1MB), but the number of inputs and outputs can still reach thousands. As a result, we can have millions of distinct chainlets (e.g., $\mathbb{C}_{1900\to200}$, $\mathbb{C}_{1901\to200}$ or $\mathbb{C}_{1900\to201}$).

We use a matrix representation to model the Bitcoin graph in time with chainlets. For a given time granularity, such as one day, we take snapshots of the Bitcoin network and construct a Bitcoin graph. Chainlet counts obtained from this graph are stored as an $n \times n$-matrix $\mathcal{O}$ such that for $i \leq n, j \leq n$

$$\mathcal{O}[i,j] = \begin{cases} \#\mathbb{C}_{i\to j} & \text{if } i < n \text{ and } j < n, \\ \sum_{z=n}^{\infty} \#\mathbb{C}_{i\to z} & \text{if } i < n \text{ and } j = n, \\ \sum_{y=n}^{\infty} \#\mathbb{C}_{y\to j} & \text{if } i = n \text{ and } j < n, \\ \sum_{y=n}^{\infty}\sum_{z=n}^{\infty} \#\mathbb{C}_{y\to z} & \text{if } i = n \text{ and } j = n. \end{cases}$$

In this matrix notation, choosing an n value, e.g., $n = 5$, means that a chainlet with more than 5 inputs/outputs (i.e., $\mathbb{C}_{x\to y}$ s.t., $x \geq 5$ or $y \geq 5$) is recorded in the n-th row or column. That is, we aggregate chainlets with large dimensions that would otherwise fall outside matrix dimensions. In what follows we use the term **extreme chainlets** to refer to these aggregated chainlets on the n-th row and column.



(a) Daily clusters. (b) Weekly clusters.

Fig. 4: [Color online]. Chainlet clusters with day and week granularities. A chainlet $\mathbb{C}_{x\to y}$ is the intersection cell of the x-th row and y-th column.

To select a suitable value for the matrix dimension n, we analyzed the entire Bitcoin history. We found that % 90.50 of the chainlets have n of 5 (i.e., $\mathbb{C}_{x\to y}$ s.t., $x < 5$ and $y < 5$) in average for daily snapshots. This value reaches % 97.57 for n of 20. We chose to take n of 20, because it can distinguish a sufficiently large number (i.e., 400) of chainlets, and still offers a dense matrix.

With daily and weekly snapshots of the Bitcoin network, we constructed 3.284 and 443 daily and weekly matrices, respectively (with data from 2009 to 2018). Each of the 400 chainlets is represented as a vector of its count in time.

We hierarchically clustered chainlets by using Cosine Similarity [8] over chainlet vectors, and used a similarity cut threshold of 0.7 to create clusters from the

hierarchical dendogram. Fig. 4 shows the resulting clusters. Cluster memberships are shown with the same color. A white cell denotes a chainlet that constitutes a cluster of its own. In both Fig. 4a and 4b, higher n values in the right low corner are clustered together, and in the daily clusters extreme chainlets ($\mathbb{C}_{\{x|x>8\}\to 20}$) have their own cluster. An interesting result is that in both matrices extreme chainlets belong to the same clusters with some considerably smaller chainlets such as $\mathbb{C}_{2\to 3}$, $\mathbb{C}_{3\to 3}$ and $\mathbb{C}_{2\to 6}$. In Section 4.2 we show that their similarity extends to their impact on price predictions.

4 Experiments

Our experiments first prove the predictive power of chainlets with Granger Causality. We then show how chainlets can be used to predict Bitcoin price.

4.1 Granger Causality

To assess a potential predictive role of chainlets in Bitcoin price formation, we employ a widely adopted econometric concept of Granger causality [6]. The causality test assesses whether one time series is useful in predicting another (see an overview by White et al. [22]). In particular, assume $\mathbf{Y}_t$, $t \in Z^+$ is a $p \times 1$-random vector (e.g., Bitcoin price) and let $\mathcal{F}^t_{(\mathbf{Y})} = \sigma\{\mathbf{Y}_s : s = 0, 1, \ldots, t\}$ denote a σ-algebra generated from all observations of $\mathbf{Y}$ in the market up to time t. Consider a sequence of $(k+2)$-tuples of random vectors $\{\mathbf{Y}_t, \mathbf{X}_t, \mathbf{Z}^1_t, \ldots, \mathbf{Z}^k_t\}$. For example, in the context of this paper $\mathbf{X}$ can be chainlets and $\mathbf{Z}^1, \ldots, \mathbf{Z}^k_t$ can be number of transactions. Suppose that for all $h \in Z^+$

$$F_{t+h}\left(\cdot|\mathcal{F}^{t-1}_{(\mathbf{Y},\mathbf{X},\mathbf{Z}^1,\ldots,\mathbf{Z}^k)}\right) = F_{t+h}\left(\cdot|\mathcal{F}^{t-1}_{(\mathbf{Y},\mathbf{Z}^1,\ldots,\mathbf{Z}^k)}\right), \tag{1}$$

where $F_{t+h}\left(\cdot|\mathcal{F}^{t-1}_{(\mathbf{Y},\mathbf{X},\mathbf{Z}^1,\ldots,\mathbf{Z}^k)}\right)$ and $F_{t+h}\left(\cdot|\mathcal{F}^{t-1}_{(\mathbf{Y},\mathbf{Z}^1,\ldots,\mathbf{Z}^k)}\right)$ are conditional distributions of $\mathbf{Y}_{t+h}$, given $\mathbf{Y}_{t-1}, \mathbf{X}_{t-1}, \mathbf{Z}^1_{t-1}, \ldots, \mathbf{Z}^k_{t-1}$ and $\mathbf{Y}_{t-1}, \mathbf{Z}^1_{t-1}, \ldots, \mathbf{Z}^k_{t-1}$, respectively. Then, $\mathbf{X}_{t-1}$ is said *not* to Granger cause (G-cause) $\mathbf{Y}_{t+h}$ with respect to $\mathcal{F}^{t-1}_{(\mathbf{Y},\mathbf{Z}^1,\ldots,\mathbf{Z}^k)}$. Otherwise, $\mathbf{X}$ is said to G-cause $\mathbf{Y}$, which can be denoted by $G_{\mathbf{X}\rightarrowtail\mathbf{Y}}$, where $\rightarrowtail$ represents the direction of causality. Hence, G-causality means that given information on the past of $\mathbf{Y}$ and $\mathbf{Z}^1, \ldots, \mathbf{Z}^k$, the past of $\mathbf{X}$ does not deliver any new information that can be used for predicting $\mathbf{Y}_{t+h}$.

In practice G-causality is typically performed by fitting two linear vector autoregressive (VAR) models of finite order d to $\mathbf{Y}$, with and without $\mathbf{X}$, respectively, and then testing for statistical significance of model coefficients associated with $\mathbf{X}$. Alternatively, we can compare predictive performance of two models (i.e., with and without $\mathbf{X}$), using an F-test, under the null hypothesis of no explanatory power in $\mathbf{X}$. For instance, consider a case of univariate time series y_t, x_t and z_t. To test G-causality of x_t, we compare the fit of the full model $y_t = \alpha_0 + \sum_{k=1}^d \alpha_k y_{t-k} + \sum_{k=1}^d \beta_k x_{t-k} + \sum_{k=1}^d \gamma_k z_{t-k} + e_t$, versus the

fit of the reduced model $y_t = \alpha_0 + \sum_{k=1}^{d} \alpha_k y_{t-k} + \sum_{k=1}^{d} \beta_k x_{t-k} + \tilde{e}_t$. That is, under the null hypothesis of no predictive effect in x onto y (i.e., x does not G-cause y), $Var(e_t) = Var(\tilde{e}_t)$. If $Var(e_t)$ is (statistically) significantly lower than $Var(\tilde{e}_t)$, then we conclude that x contains additional information that can improve forecasting of y, i.e., $G_{x \rightarrowtail y}$.

Armed with the time series of chainlets, we are now interested in evaluating the potential impact of local graph structures on *future* bitcoin price formation and investment risk. We are primarily interested in two interlinked questions:

1. Do changes in chainlet characteristics exhibit any causal effect on future Bitcoin price and Bitcoin returns?
2. Do chainlets convey some unique information about future Bitcoin prices, given more conventional economic variables and non-network blockchain characteristics?

Table 1 provides summary results of the Granger causality tests for predictive utility of individual/aggregate chainlets, and chainlet clusters [3] in analysis of the Bitcoin price and its log returns (see Fig. 4a for the clusters). Log returns of Bitcoin prices measure the relative change in prices and are defined as $LR_t = \log y_t - \log y_{t-1}$. As a more conventional predictor, we also include the total number of transactions (# of Trans.) into the baseline models. Direction of causality is denoted by $\rightarrowtail$. Table 1 indicates that individual chainlets, e.g., $\mathbb{C}_{6\rightarrow 1}$, $\mathbb{C}_{1\rightarrow 7}$, $\mathbb{C}_{20\rightarrow 12}$, as well as aggregate chainlets, e.g., split chainlets, have a predictive impact on price formation, and in some cases also exhibit causal linkage with future log returns. Some chainlet clusters have predictive relationship only with Bitcoin price, whereas Cluster 35 G-causes both price and log returns. As expected, total number of transactions also has causality effects on both Bitcoin price and log returns. The G-causality relationships of different chainlets and Bitcoin price indicate that they are likely to contain important predictive information on Bitcoin price formation and volatility.

4.2 Price prediction

In Section 4.1 we show that chainlets G-cause the Bitcoin price and hence, exhibit predictive impact on prices. We are now interested in quantifying the forecasting utility of chainlets. To evaluate the chainlets' predictive power, *we can use any forecasting model and compare predictive performances with and without chainlets*. Typically such a comparative analysis is performed based on the Box-Jenkins (BJ) class of parametric linear models. However, as indicated by [10], more flexible Random Forest (RF) models often tend to outperform the BJ models in their predictive capabilities. In particular, we find that the optimal baseline autoregressive integrated moving average (ARIMA(p, d, q)) models selected by minimizing the Akaike Information criterion (AIC), yield from 0.2% to 40% higher prediction root mean squared error (RMSE) than the RF baseline

[3] Some representative chainlets from daily clusters 7, 8, 16 and 35 are $\mathbb{C}_{9\rightarrow 11}$, $\mathbb{C}_{3\rightarrow 17}$, $\mathbb{C}_{8\rightarrow 14}$ and $\mathbb{C}_{1\rightarrow 1}$, respectively.

Table 1: In G-causality, P and LR denote significance in price & log returns, respectively; blank space implies no significance. Confidence level is 95%.

Covariate Types	**Causality**	**Outcome with lag effects** 1	2	3	4	5
# of Trans.	Total # Trans. $\rightarrowtail$ Outcome	LR	LR	P/LR	P/LR	
Aggregate Chainlets	Merge Chainlets $\rightarrowtail$ Outcome	-	-	-	-	-
	Split Chainlets $\rightarrowtail$ Outcome	-	LR	P/LR	P	-
	Trans. Chainlets $\rightarrowtail$ Outcome	-	-	-	-	-
Individual Chainlets	$\mathbb{C}_{1\rightarrow 7} \rightarrowtail$ Outcome	P	P	P	P	P
	$\mathbb{C}_{6\rightarrow 1} \rightarrowtail$ Outcome	-	P	P	P	-
	$\mathbb{C}_{3\rightarrow 3} \rightarrowtail$ Outcome	-	P	P	P	-
Extreme Chainlets	$\mathbb{C}_{20\rightarrow 2} \rightarrowtail$ Outcome	LR	P/LR	P/LR	P/LR	P
	$\mathbb{C}_{20\rightarrow 3} \rightarrowtail$ Outcome	P	P	P	P	P
	$\mathbb{C}_{20\rightarrow 12} \rightarrowtail$ Outcome	P	P	P	P	P
	$\mathbb{C}_{20\rightarrow 17} \rightarrowtail$ Outcome	-	-	P	P	P
Chainlet Clusters	Cluster 35 $\rightarrowtail$ Outcome	LR	LR	P/LR	P/LR	-
	Cluster 16 $\rightarrowtail$ Outcome	-	LR	-	-	-
	Cluster 8 $\rightarrowtail$ Outcome	-	P	P	P	P
	Cluster 7 $\rightarrowtail$ Outcome	-	P	P	P	P

models. Here RMSE $= \sqrt{(1/n)\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$, where y_t is the test set of Bitcoin price and $\hat{y}_t$ is the corresponding predicted value. ARIMA and RF models deliver comparable results, therefore, due to space limitations, we present the comparison study based only on the RF type of models.

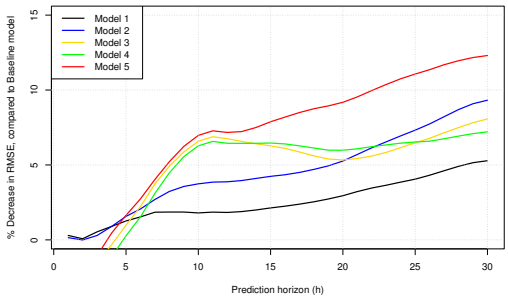


Fig. 5: % Change (decrease) in RMSE compared to the baseline model.

We performed extensive experiments with various chainlets and selected to showcase six of these RF models. Table 2 provides an overview of the constructed models. The baseline model includes only the lagged (past period) values of the Bitcoin price. Other models comprise of lagged prices with different covariates, mainly chainlets or some functions of chainlets such as the mean of all aggregate/split type chainlets and mean of all chainlets in a specific cluster.

In our study each RF model used 500 trees, and sampling all rows of the data set is done with replacement. Number of variables used at each split are, for example, 2, 3 and 4, for Models 1, 2 and 5, respectively.

Table 2: Model description for Bitcoin price (response) and varying predictors.

Model	**Predictors**
Baseline M_0	Price lag 1, Price lag 2, Price lag 3
Model 1	Price lag 1, Price lag 2, Price lag 3, # Trans lag 1 , # Trans lag 2, # Trans lag 3
Model 2	Price lag 1, Price lag 2, Price lag 3, Split Pattern lag 1, Split Pattern lag 2, Split Pattern lag 3 Cluster 8 lag 1, Cluster 8 lag 2, Cluster 8 lag 3
Model 3	Price lag 1, Price lag 2, Price lag 3, $\mathbb{C}_{1\to7}$ lag 1, $\mathbb{C}_{1\to7}$ lag 2, $\mathbb{C}_{1\to7}$ lag 3
Model 4	Price lag 1, Price lag 2, Price lag 3, $\mathbb{C}_{1\to7}$ lag 1, $\mathbb{C}_{1\to7}$ lag 2, $\mathbb{C}_{1\to7}$ lag 2, $\mathbb{C}_{6\to1}$ lag 1, $\mathbb{C}_{6\to1}$ lag 2, $\mathbb{C}_{6\to1}$ lag 3
Model 5	Price lag 1, Price lag 2, Price lag 3, $\mathbb{C}_{1\to7}$ lag 1, $\mathbb{C}_{1\to7}$ lag 2, $\mathbb{C}_{1\to7}$ lag 2, $\mathbb{C}_{6\to1}$ lag 1, $\mathbb{C}_{6\to1}$ lag 2, $\mathbb{C}_{6\to1}$ lag 3, $\mathbb{C}_{3\to3}$ lag 1, $\mathbb{C}_{3\to3}$ lag 2, $\mathbb{C}_{3\to3}$ lag 3

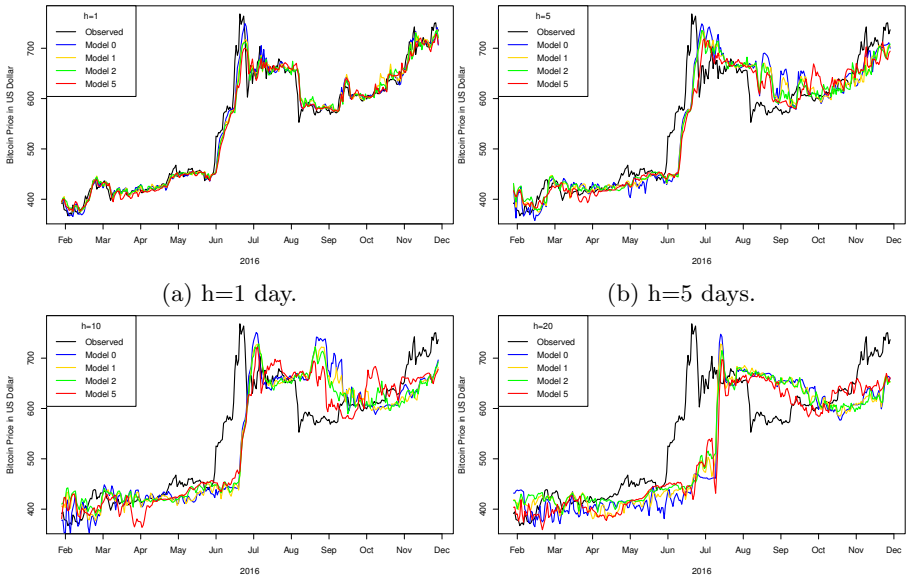
We continuously change the training data using a sliding window technique, where we choose the window size of 200. That is, at each time step we train our model based on the past 200 values, and armed with this estimated model, we then construct a h step ahead forecast.

Predictive utilities of models in Table 2 over the baseline model can be measured as $\Psi_{(X\rightarrowtail Y)} = \psi(M)/\psi(M_0)$, where ψ is a measure of prediction error, e.g., root mean squared error (RMSE). Here $\psi(M_0)$ is the prediction error of baseline model, where lagged prices are the only predictor; and $\psi(M)$ is the prediction error of a given model, where predictors are lagged prices and other exogenous covariates ($\mathbf{X}$). If $\Psi_{(X\to Y)} < 1$, the covariate ($\mathbf{X}$) is said to improve prediction of Y. We also calculate the percentage change in ψ for a specific model w.r.t. M_0 as $\Delta = \big(1 - \Psi_{(X\rightarrowtail y)}\big)100\%$.

Fig. 5 compares the percent decrease in RMSE for different models, calculated for varying prediction horizons $h = 1, \ldots, 30$. For 1-step ahead forecast, chainlets and other covariates do not contribute useful predictive information over history of Bitcoin price. However, for 3 or more steps ahead forecasts, chainlets play an increasingly significant predictive role in Bitcoin price formation, even when other more conventional factors, such as historical price and number of transactions, are already in the model.

Furthermore, some chainlets has a higher utility for price prediction. For example, in Model 5, we observe the highest decrease in RMSE, compared to the baseline model. Models 3 and 4 yield the second highest decrease in RMSE until the forecast horizon h of 20. After h of 20, Model 2 delivers the second highest reduction in RMSE over the baseline model.

Fig. 6 compares the observed data with fitted values from baseline model and three other models, i.e., Model 1, 2, and 5. For h of 1, all models deliver similar prediction accuracy and capture the variability of the data very well. Although, as expected, the prediction performance of all models deteriorates as forecasting



(a) h=1 day. (b) h=5 days.

(c) h=10 days. (d) h=20 days.

Fig. 6: [Color Online]. Price prediction for 2016 with 1, 5, 10 for 20 day horizons.

horizon $h \to \infty$, Models 1, 2, and 5 still yield a noticeably higher predictive accuracy, compared to the baseline model without chainlets.

5 Conclusion

We introduce a novel concept of k-chainlets on Bitcoin that expands the ideas of motifs and graphlets to Blockchain graphs. Chainlet analysis provides a deeper insight into local topological properties of the Blockchain and the role of those local higher-order topologies in the Bitcoin price formation. We find that certain types of chainlets have a high predictive utility for Bitcoin prices. Furthermore, extreme chainlets exhibit an important role in the Bitcoin price prediction.

Acknowledgments

This research was supported in part by NIH 1R01HG006844, NSF CNS-1111529, CICI-1547324, IIS-1633331, DMS-1736368 and ARO W911NF-17-1-0356.

References

1. Ahmed, N.K., Neville, J., Rossi, R.A., Duffield, N., Willke, T.L.: Graphlet decomposition: Framework, algorithms, and applications. KAIS **50**, 1–32 (2016)

2. Akcora, C.G., Gel, Y.R., Kantarcioglu, M.: Blockchain: A graph primer. arXiv preprint arXiv:1708.08749 (2017)
3. Androulaki, E., Karame, G.O., Roeschlin, M., Scherer, T., Capkun, S.: Evaluating user privacy in bitcoin. In: IFCA. pp. 34–51. Springer (2013)
4. Baumann, A., Fabian, B., Lischke, M.: Exploring the Bitcoin network. In: WEBIST (1). pp. 369–374 (2014)
5. Di Battista, G., Di Donato, V.and Patrignani, M., Pizzonia, M., Roselli, V., Tamassia, R.: Bitconeview: visualization of flows in the bitcoin transaction graph. In: IEEE VizSec. pp. 1–8 (2015)
6. Granger, C.W.J.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica **37**(3), 424–438 (1969)
7. Greaves, A., Au, B.: Using the bitcoin transaction graph to predict the price of bitcoin. No Data (2015)
8. Huang, A.: Similarity measures for text document clustering. In: NZCSRSC. pp. 49–56 (2008)
9. Jiang, X. F., C.T.T., Zheng, B.: Structure of local interactions in complex financial dynamics. Scientific Reports **4**(5321) (2014)
10. Kane, M.J., Price, N., Scotch, M., Rabinowitz, P.: Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. BMC Bioinformatics **15**(1), 276 (2014)
11. Kondor, D., Csabai, I., Szüle, J., Pósfai, M.and Vattay, G.: Inferring the interplay between network structure and market effects in Bitcoin. New J. of Phys. **16**(12), 125003 (2014)
12. Kondor, D., Pósfai, M., Csabai, I., Vattay, G.: Do the rich get richer? An empirical analysis of the Bitcoin transaction network. PLOS One **9**(2), e86197 (2014)
13. Lischke, M., Fabian, B.: Analyzing the bitcoin network: The first four years. Future Internet **8**(1), 7 (2016)
14. Madan, I.and Saluja, S., Zhao, A.: Automated bitcoin trading via machine learning algorithms (2015)
15. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: Simple building blocks of complex networks. Science **298**(5594), 824–827 (2002)
16. Moser, M.and Bohme, R., Breuker, D.: An inquiry into money laundering tools in the bitcoin ecosystem. In: eCRS. pp. 1–14. IEEE (2013)
17. Nakamoto, S.: Bitcoin: A peer-to-peer electronic cash system (2008)
18. Ober, M., Katzenbeisser, S., Hamacher, K.: Structure and anonymity of the bitcoin transaction graph. Future Internet **5**(2), 237–250 (2013)
19. Portnoff, R.S., Huang, D.Y., Doerfler, P., Afroz, S., McCoy, D.: Backpage and bitcoin: Uncovering human traffickers. In: SIGKDD. pp. 1595–1604. ACM (2017)
20. Sorgente, M., Cibils, C.: The reaction of a network: Exploring the relationship between the Bitcoin network structure and the Bitcoin price. No Data (2014)
21. Tschorsch, F., Scheuermann, B.: Bitcoin and beyond: A technical survey on decentralized digital currencies. IEEE COMMUN SURV/TUT **18**(3), 2084–2123 (2016)
22. White, H., Chalak, K., X., L.: Linking Granger causality and the Pearl causal model with settable systems. In: JMLR. vol. 12, pp. 1–29 (2011)
23. Yang, S.Y., Kim, J.: Bitcoin market return and volatility forecasting using transaction network flow properties. In: IEEE SSCI. pp. 1778–1785 (2015)