NUS RMI Working Paper Series – No. 2019-03

Discover Regional and Size Effects in Global Bitcoin Blockchain via Sparse-Group Network AutoRegressive Modeling

Ying CHEN, Simon TRIMBORN and Jiejie ZHANG

25 August 2019

NUS Risk Management Institute
21 HENG MUI KENG TERRACE, #04-03 I^3 BUILDING, SINGAPORE 119613

www.rmi.nus.edu.sg/research/rmi-working-paper-series

Discover Regional and Size Effects in Global Bitcoin Blockchain via Sparse-Group Network AutoRegressive Modeling

Ying Chen[1,2,3,4], Simon Trimborn *[1], and Jiejie Zhang[2]

[1]Department of Mathematics, National University of Singapore, Singapore

[2]Department of Statistics & Applied Probability, National University of Singapore, Singapore

[3]Department of Economics, National University of Singapore, Singapore

[4]Risk Management Institute, National University of Singapore, Singapore

August 25, 2019

Abstract

Bitcoin blockchain has grown into an active global virtual money network with millions of accounts. We propose a Sparse-Group Network AutoRegressive

*Corresponding author, phone: +65 6516-1245, E-Mail: simon.trimborn@nus.edu.sg

(SGNAR) model to understand the dynamics of its cross-border transactions. It describes the money flows of virtual funds, with a focus on the regional and size effects in the Bitcoin network at a global level. In particular, we develop a regularized estimator with two-layer sparsity, which enables discovering 1) the active regions with influential impact on the global network and 2) the size of the groups which lead the dynamic evolution of the Bitcoin transaction network. Our study considers the up-to-date Bitcoin blockchain, from February 2012 to July 2017, with all the transactions being classified into 60 groups according to region and size. We found that mostly the users with the smallest and largest sizes of transactions from North America, Europe, South America, Africa and Asia were driving the Bitcoin transactions, while the other groups and all the groups in Oceania were either followers or isolated. The global connectivity remained low in the period from 2013 to 2015, although it was high in 2012 and enhanced in the recent years of 2016 and 2017.

Keywords: Bitcoin Blockchain, Network Dynamics, Two-Layer sparsity

JEL classification: C55, C58, C60, G17

1 Introduction

Powered by the blockchain technology, Bitcoin (BTC) brought an innovative financial asset class into the market. On the blockchain, BTC is transacted as a borderless decentralized digital currency and has grown into an active global virtual money network with millions of accounts. Just as banking customers in the United States send USD denominated transactions to settle their financial obligations, so BTC blockchain users send BTC denominated transactions to each other. The number of BTC transactions has increased incredibly. According to blockchain.com, a total of 368 million transactions were sent by 1 January 2019. The average daily number of transactions was 91 in 2009, further rose to 69,084 in 2014 and peaked in 2017 with 283,281 average daily transactions, an increase of 310% between the years 2014 and 2017. BTC has even started playing a key role in the unsteady financial systems of some developing countries, such as Venezuela or Zimbabwe.

Despite its impressive growth, the public's attention is on the potential massive risks of BTC from, e.g. sudden price drops and liquidation risk. In December 2013 and January 2018, the price dropped by 50% and 63% over two and four weeks respectively. An extreme case would happen if users stopped interacting on the BTC blockchain, which would cause every user to suffer the loss of their invested capital. This has already happened to 1661 cryptocurrencies (CCs), according to deadcoins.com. The situation in BTC seems better but definitely not optimistic. Several studies have reported high volatility and tail risk in BTC (Elendner et al., 2017; Feng et al., 2018), frequent jumps (Scaillet et al., 2018), informed trading (Feng et al., 2017), bubbles, and sudden drops in market value (Hafner, 2018). Despite the huge market risks, one observes a demand on the exchanges to enter the market. After buying BTC on a CC exchange, one either owns them for the purpose of speculative investments or uses them for transactions with others via the blockchain. While investors' behaviour can be studied via Limit Order Books and the evolution of its price, one knows little about the users' behaviour in the BTC blockchain. The anonymity of the BTC blockchain masks the purposes of BTC transactions and often also their frequency. However the dynamic evolution of the virtual money flows via blockchain can provide a number of insightful implications about the users of BTC. Since the BTC blockchain is a payment network, transactions need to be carried out first to enable new transactions by other users. Taking into consideration the geographical information paves the way to study the influence of certain regions and certain types of users on the growth of the network.

For the borderless BTC, naturally this question has to be addressed at the global level. Europe and North America have for many years been considered as leaders in the financial markets. However the recent frantic enthusiasm for crypto mining in certain areas, in particular China, Japan and Korea (as reported in the media), calls into question the composition in the BTC blockchain network. The first question of our study is:

Q1: Where are the active users coming from or which regions are leading the transactions and how much influence do they have on the BTC blockchain network?

The active users are defined as those who have an influential impact on the dynamic evolution of transactions. For exchanges where BTC are traded, Kristoufek (2015) considered the US and China and found a high correlation between the evolution of the price in the two markets and a strong positive correlation between the Chinese and US trading volumes. Darlington III (2014) studied the geographical usage of BTC at a global level by metrics defined on the BTC related search queries on Google and the number of mining nodes. For the BTC blockchain, where the virtual currency is sent from user to user directly, Ron and Shamir (2013) analysed the transaction behaviours of the accounts from the emergence of Bitcoin until 13 May 2012. While studies like Lischke and Fabian (2016) have looked into the geographical distribution and the official vendor BTC transactions, little is known about the regional effect of users.

To investigate the regional interactions of the entire global network, we group the BTC transactions by continent. Inside each of these groups, the data are further split into 10 groups according to the transaction size, resulting in all in 60 groups. These serve as a proxy for the wealth of a BTC investor, which is used to classify the users. It follows the heuristic that only big Bitcoiners are able to execute large transactions: small Bitcoiners contribute to the small transactions. Detecting the essential dynamic interactions between the regional and size groups can help to answer the second question:

Q2: What are the features of the active users in terms of transaction size?

Our analysis of the dynamic network activity finds the presence of serial cross-correlation. The existence of serial correlations motivates the adoption of Vector AutoRegressive (VAR) models for analysing the transactions of the BTC network. Already since Ord (1975), VAR has been used to investigate spatial interactions in networks. Pesaran et al. (2004) investigates the exposure of economies to each other, Chudik and Pesaran (2011) study Infinite-dimensional VARs under the assumption that each node is related to a small number of neighbouring nodes and a large number of non-neighbouring ones. Dees et al. (2007) study the network between the European Union and 26 countries. Creal et al. (2013) propose Generalized AutoRegressive models and study the relation between exchange rates and credit risk ratings. Zhu et al. (2017) develop the Network vector AutoRegressive (NAR) model, where the connectivity of the network is represented by an adjacency matrix that is a given or pre-determined binary matrix, see also Zhou et al. (2017). Both papers assume that the dynamic network connectivity is controlled by one network parameter, which, in combination with the given adjacency matrix, circumvents the dimensionality problem with large-scale networks. Though simple, modelling with one single network parameter and, more importantly, a known adjacency matrix, is a strong and unrealistic constraint for studying the BTC blockchain. While the geographical origin of a transaction can be identified, the geographical destination of a transaction is unknown, requiring an estimation of the adjacency matrix of the network.[1] This motivates using a flexible VAR model with unknown adjacency matrix, which encounters the overfitting problem for high dimensional networks. The estimation is often inefficient or even infeasible, unless one imposes some lower-dimensional structural assumptions, e.g. sparsity in the parameter space, see Basu and Michailidis (2015).

Regularization approaches were originally designed for the univariate case in regressions, but have recently been brought to a vector time series context including the high-dimensional VAR models. In an investigation of large Vector AutoRegressive models with exogenous variables (VARX), Nicholson et al. (2017) propose five kinds of penalties. Song and Bickel (2011) assume a sparse structure for the lags and apply group sparsity to the columns of the parameter matrix. These studies build on the l_1/l_2-norm penalties proposed by Hoerl and Kennard (1988), Tibshirani (1996) and Zou and Hastie (2005), also known as ridge regression, the lasso, and naïve elastic net. Yuan and Lin (2006) develop the group sparsity method for regression models. The spline-lasso (Guo et al., 2016) allows for smoothly changing coefficients, which is motivated by the fused lasso (Tibshirani et al., 2005), encouraging locally constant coefficients within groups. Adopting both the l_2-norm and the l_1-norm in the regression context, Simon et al. (2013) develop an algorithm to search for the solution with group lasso penalization while allowing for individual penalizations inside of the

[1] The BTC blockchain uses relaying nodes to distribute the transactions to each participant. The IP address of the relaying node can be observed and provides an approximation of the origin of a transaction. Since the ownership of the funds is recorded in a public database distributed to each user and not in the accounts only, the destination of the transaction is not observable. For more details on the procedure used to observe this information, refer to Section 2.

groups.

We propose a Sparse-Group Network AutoRegressive (SGNAR) model to study the dynamics in the BTC blockchain. The entries of the adjacency matrix are considered unknown and not necessarily binary, introducing a flexibility of the existence and level of connectivity in the network. By doing this, we essentially assume that only a few nodes are active. Moreover, diverse magnitudes of the parameters within the groups, with some being zero, implies the existence of individual sparsity. The sparsity assumption is necessary due to the huge dimensionality in combination with the limited data availability. The SGNAR model adopts two kinds of sparsity. Group sparsity, Yuan and Lin (2006), is applied to the columns (nodes) to identify the influential groups in certain continents, referred to as active nodes. Individual sparsity, Tibshirani (1996), is imposed on the individual parameters in an active node, indicating that the active node does not have an effect on every other group. The proposed SGNAR estimator with this two-layer sparsity enables discovering 1) the active regions that dominate the global transactions and 2) the size groups who lead the dynamic evolution of the network. For the optimization of the SGNAR, we develop an algorithm for the two-layer sparsity for high-dimensional networks.

This research is related to previous studies, yet there are several differences. SGNAR provides the adjacency matrix estimator with sparse-group sparsity. Song and Bickel (2011) use a lasso type sparsity as a pre-selector for un-regularized time series modelling. In the SGNAR framework, we derive the least square estimator under two penalties and introduce a two-step algorithm into the SGNAR model to obtain the estimator numerically. Simon et al. (2013) derive a related algorithm for univariate regression models, whereas we develop the algorithm for high-dimensional VAR type models. The interconnections strongly challenge the algorithm compared to the univariate case, making it computationally intensive. Lastly, the SGNAR model contributes to the literature on econometric modelling for BTC networks. To the best of our knowledge, the BTC blockchain has not been studied from an economic viewpoint that considers the time dependent impact of regional transaction activity on the same continent and other regions. Our study provides an analysis of the real data of the BTC transactions from 25 February 2012 to 17 July 2017. Our results demonstrate the spatial connections and dynamic changes in the BTC blockchain. In particular, it is shown that:

- In 2012, connectivity in the network was present before it vanished in 2013–2015 completely. Yet from 2016 onward, the connectivity improved in the BTC network.
- Certain user groups from Europe, North America, Asia and South America led the BTC blockchain network, with a dynamic influence on the rest of the world in 2012 and 2016, and in most cases very large and very small Bitcoiners of the continents play a key role.
- Spatial differences are apparent in the global network. While Europe, North America, South America, Africa and Asia are influential, Oceania is a follower. In Europe, North America and Asia, only small Bitcoiners play a role, while in South America and Africa the big ones are important.
- Taking into account that most Bitcoin mining farms are in Asia, it is surprising to some extent that Asia is not the sole driver but operates Bitcoin for Europe, North America, Africa and South America, fostering the importance of these regions in the blockchain.

This paper is organized as follows. Section 2 describes the BTC transaction data. Section 3 presents the Sparse-Group Network AutoRegression (SGNAR) model; 3.1 gives details on the estimation of the adjacency matrix, then 3.2 presents the algorithm to optimize the SGNAR. Section 4 applies the SGNAR model to real

Bitcoin transaction data and presents an interpretation and discussion. Section 5 presents some conclusions. The codes to carry out the numerical calculations are available on the corresponding author's GitHub account.

2 Data description

We consider the BTC blockchain from 25 February 2012 to 17 July 2017 (1969 days with 1843 observed days). The raw data are published on the blockchain at 10-minute frequency[2] with attributes of transaction size, account ID, accounts participating in the transactions, the timestamp of the transaction, source Blockchain.info. Blockchain.info provides additional information on the IP address from the relaying party of the origin of the transaction which is used to label the region. However the IP address is not available after July 2017, which unfortunately makes a dynamic regional analysis of the year 2018 and onwards impossible. We group the data into 6 continents: Africa (AF), Asia (AS), Europe (EU), North America (NA), Oceania (OC) and South America (SA). The continent is identified depending on the IP address compared with a dataset of IP address from MaxMind Inc. We follow Reid and Harrigan (2013) in tracking the approximate location of the origin of the transaction [3]. By assuming that the node that informs first about a transaction is the location where the transaction takes place, one can approximately identify the location where the transaction originates. This approach only works as long as the running node does not use an anonymizing technology.

Each continental group is further categorized according to the transaction size. Due to anonymity, characterizing BTC users is not easy. We thus group the users

[2] Note that on the Bitcoin blockchain, the transactions are not published at the moment they occur. The miners collect the records and publish them as a block at a 10-minute frequency.

[3] The location of the relaying node gets observed, which is geographically close to the origin of the transaction. Since the information is saved in the blockchain and each user has a copy of it, no information on receiving node gets recorded. Consequently the final destination is not traceable.

according to the size of the transactions associated with the accounts. The heuristic behind this is that only big Bitcoiners are able to execute large transactions, while small Bitcoiners contribute to small transactions. One should note that this is a stringent assumption, yet we deem it valid in the unique transaction network of BTC. Inside each continental grouping, the data are separated into 10 size groups, depending on the deciles of the sizes of the transactions. The first group, indicated by a 1 placed after the abbreviation for the continent, has the smallest transactions, corresponding to the 0%–10% percentile, while the tenth group, with the largest transactions, is indicated by a 10, and corresponds to the 91%–100% percentile. Later, for robustness analysis, we also consider a 3 group per continent setting, where users are clustered into three size groups corresponding to 0%–30%, 31%–70%, and 71%–100% percentiles for small, medium, and big investors, respectively.

The identification of the originating continent and building of the groups we conducted based on the raw data with 10-min frequency. Except for Europe and North America, there are 1% and 25% zeros, meaning no transactions. A lack of liquidity can be challenging for the model estimation. We overcome the liquidity problem by accumulating the raw data to a daily frequency.

For the further analysis, we consider the log transactions. To avoid $-\infty$ in the data for cases without any transactions in a continental grouping within a day, we add 1 Satoshi [4] to each transaction. Given the large numbers under consideration, the bias effect of the correction is negligible.

Figure 1 displays the evolution of the daily log accumulated transaction sizes over all groups in each continent. Europe and North America on average have the largest transactions and the dynamic pattern is quite steady. Asia and Oceania contain a few days (8 and 19) without transactions, even after accumulating to daily frequency.

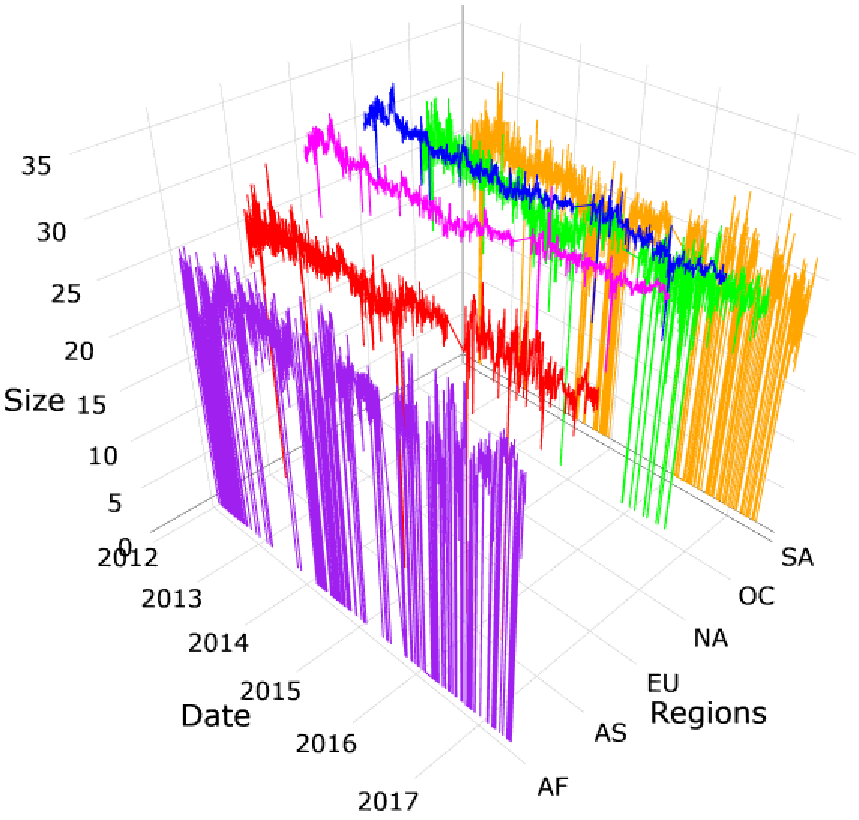
[4] The BTC transactions are reported in Satoshi values, the smallest fraction of a BTC, where 1 BTC $= 100,000,000$ Satoshi.

They are also more volatile than Europe and North America. Africa and South America are the most volatile and have a relatively larger number of days, 364 and 241, respectively, without transactions. The interpretation of Figure 1 is further supported by the descriptive statistics, presented in Table 1. Inferring from the mean and standard deviation, Europe and North America, Asia and Oceania, Africa and South America indeed show a related behaviour. Also the minimum values indicate the existence of zero transactions, a lack of liquidity, in some areas, supporting the previous analysis.

Table 2 provides the average daily transaction volumes in USD in each decile of each continent. For the conversion to USD we consider the daily closing price of BTC reported on YahooFinance. One sees that the daily transaction volumes can be very low, especially in Africa and South America, see the lowest decile. At the same time, the transactions in the top decile of Europe have a mean transaction volume of over 144 million USD per day. In Africa and South America for the same decile, it still ranges to over 200,000 and 425,000 USD on daily average. Apparently there are quite high transaction volumes, especially when considering that BTC is still an emerging asset.

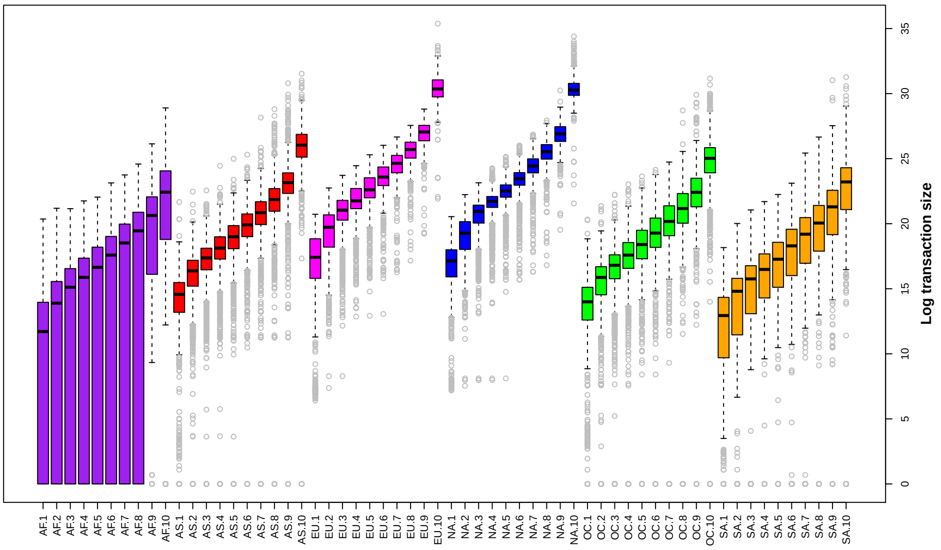
For deeper insights into the features of the data of the groups in each continent, the empirical distribution of the log of the sizes of the transactions is displayed as a boxplot in Figure 2. For each continent, the left plot corresponds to the first group, namely group 1 with the smallest transactions, and the right one to group 10 with the largest transactions, leading to an increasing pattern within each continent. The narrow box width of Europe and North America suggests a smooth evolution of the transaction sizes with few spikes. There are hardly any occurrences of zero transactions, indicating a healthy liquidity in these regions. This indicates a more mature market in Europe and North America, hence a clearer structure within an estimated model is to be expected. Asia and Oceania are relatively more dispersely distributed. The daily transaction sizes are more volatile, inferred from the size of the center box and the length of the whiskers. South America becomes again extreme in the sense of showing longer whiskers, translating to a larger variation of the sizes of the transactions within each group. Even in group 10 with the highest transaction sizes, there are days without any transactions. Africa follows a very different pattern from the other continents. The respective boxplots indicate high volatilities with frequent drops to zero transaction volume. The divergences between the groups eventually suggests, for the modelling, an adjacency matrix with a flexible choice of parameters.

Figure 1: Time series of daily log accumulated transactions. The time period is 25 February 2012 until 17 July 2017 in the 6 continents Africa, Asia, Europe, North America, Oceania, South America.



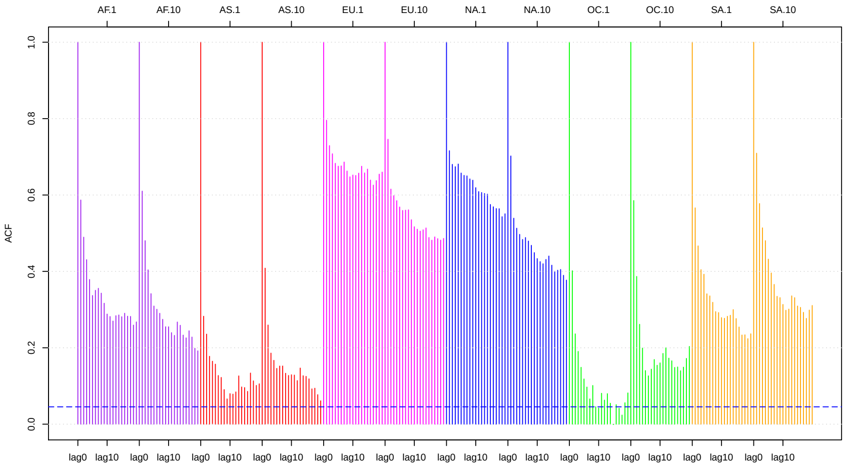
There remains the question if there is any dynamic dependence between the BTC transactions. As an illustration, Figure 3 displays an autocorrelation effect within the highest and lowest groups on each continent. The two extreme groups in each

Figure 2: Daily log transactions of the 10 groups displayed as boxplots, where the left boxplot represents group 1 and the right one group 10 of the respective continent. The time period is 25 February 2012 until 17 July 2017 in the 6 continents Africa, Asia, Europe, North America, Oceania, South America. The first 8 boxplots for Africa range to 0 due to the little number of transactions in this continent in several days.



continent exhibit serial autocorrelation and persistence, with a slow decay of the autocorrelations. Not surprisingly, the effect is the strongest in Europe and North America. In each of these two continents, the lowest group exhibits a stronger effect than the largest. Oceania and Asia, on the other hand, have weaker serial dependences. The effect in the highest group 10, is stronger than in their corresponding lowest group 1. The remaining continents, Africa and South America, share similar serial dependence. Moreover, there are network effects within the BTC blockchain, as reflected by the lag 1 cross-correlations between the groups and the regions, see Figure 4. The diagonal block of the heat map shows the lead–lag dependence among the groups within the same continent, while the off-diagonal shows the intra-continental cross-dependence. Europe and North America exhibit a stronger cross dependence, both inter-continent and intra-continent, in terms of their influence on the others (lead) and being affected by the others (lag). The network effect is much less

Figure 3: Autocorrelation Functions of the total value of each day's transactions in the 6 continents Africa, Asia, Europe, North America, Oceania, South America.



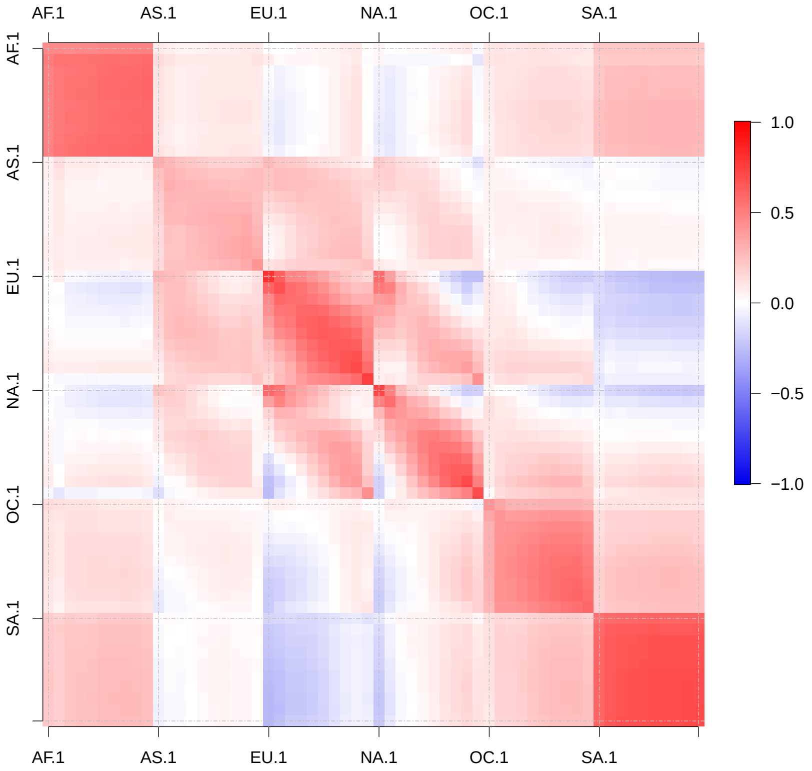
between the other continents. South America and Africa exhibit a connection within themselves and simultaneously sparse cross-dependence with the other continents. This indicates that these continents are self-dependent.

The magnitude of the cross dependence differs, suggesting flexibility in the dynamic modeling parameters. In the heat map, there are a number of zeros, displayed as blank fields, and values close to zero, which implies sparsity in the dynamic structure. All the intra-continental dependences are positive and with similar values, although in Asia the parameter values are a bit lower than in the other continents. This motivates considering one common parameter to represent the autocorrelation.

Table 1: Descriptive statistics of the log accumulated transactions of the 6 regions Africa (AF), Asia (AS), Europe (EU), North America (NA), Oceania (OC), South America (SA).

	AF	AS	EU	NA	OC	SA
mean	18.65	26.01	30.43	30.41	24.76	20.48
sd	9.47	2.19	0.95	0.85	3.08	8.15
skewness	-1.34	-7.17	-1.03	-0.74	-5.34	-1.96
kurtosis	3.06	85.75	11.73	15.79	43.06	5.24
min	0.00	0.00	22.04	21.78	0.00	0.00
max	28.91	31.65	35.38	34.40	31.21	31.85

Figure 4: Lag 1 cross-correlations between the size of the transactions – ordered in 10 groups – in the 6 regions: Africa (AF), Asia (AS), Europe (EU), North America (NA), Oceania (OC), and South America (SA). Each block on the diagonal represents the lag 1 dependence within a continent, while the off-diagonal blocks represent the inter-continental effects.



3 Sparse-Group Network AutoRegression

We propose a Sparse-Group Network AutoRegression (SGNAR) model to describe the dynamic dependence in a network with an unknown and sparse adjacency matrix. The adjacency matrix reflects both the connectivity with non-zero values and their strengths, shown as the magnitudes among the nodes. The serial dependence on its own lagged value is controlled by a common parameter. To detect the essential dynamic dependence, a two-layer sparsity is imposed on both group and individual effects. We develop a regularized least squares estimator and a gradient descent algorithm for modelling the high dimensional network.

Table 2: Mean daily transaction value (in USD) in the deciles of the transactions of the 6 regions Africa (AF), Asia (AS), Europe (EU), North America (NA), Oceania (OC), South America (SA).

	AF	AS	EU	NA	OC	SA
.1	7.18	52.52	920.34	333.46	11.71	4.06
.2	34.53	205.43	5931.73	2470.26	56.53	19.53
.3	89.94	448.50	18743.82	7341.04	147.04	57.45
.4	197.07	1039.13	39903.12	15438.44	342.60	131.56
.5	431.86	2074.87	83607.85	34735.14	771.13	302.91
.6	947.88	4386.15	177291.31	80343.77	1683.52	713.64
.7	2179.41	17316.84	390226.22	193029.66	3853.24	1970.66
.8	5834.54	66690.19	1011731.37	531628.26	10173.54	4955.48
.9	20647.78	235919.06	3677787.40	2040912.93	40651.43	17731.31
.10	425972.68	2789979.66	144060061.43	71684792.44	573119.69	211864.35

Let N denote the size of the network and $Y_{i,t}$ denote the transaction size of Node i, $1 \leq i \leq N$ at time t, $1 \leq t \leq T$, where T is the length of the time period. The SGNAR model is defined as follows

$$Y_{i,t} = \beta_0 + \beta_1 Y_{i,(t-1)} + \sum_{j=1}^{N} a_{ij} Y_{j,(t-1)} + \mathbf{Z}_i^\top \gamma + \varepsilon_{i,t} \tag{1}$$

where the parameter β_1 controls the autoregressive dependence. The adjacency matrix $A = (a_{ij})_{1\leq i,\ j\leq N}$ represents the connectivity. The elements of A reflect both the connectivity between Node i and the lagged value of Node j, if nonzero, but also the strength of the dynamic influence of Node j's lag on Node i. The adjacency matrix is assumed to be sparse, with few non-zero entries, highlighting active groups and nodes. If $a_{ij} \neq 0$, Node j is active and has influence on Node i. For $a_{ij} = 0$, Node j has no influence on Node i. If $a_{ij} = 0$ for all i, then Node j is inactive. It is unknown which elements are zeros and which are not. Since the autoregressive dependence is parametrized by β_1, the diagonal elements of A are forced to be zeros (i.e. $a_{ii} = 0$, $1 \leq i \leq N$). SGNAR also allows the measurement of the impact of exogenous variables, if applicable. Suppose there is a q-dimensional random vector $\mathbf{Z}_i = (Z_{i1}, \cdots, Z_{iq})^\top \in \mathbb{R}^q$ observed for node i. The SGNAR model allows the estimation of $\gamma = (\gamma_1, \cdots, \gamma_p)^\top$. In addition, $\varepsilon_{i,t}$ is white noise s.t. $\mathrm{E}(\varepsilon_{i,t}) = 0$, $\mathrm{E}(\varepsilon_{i,s}\varepsilon_{i,\tau}) = 0$, $\mathrm{Var}(\varepsilon_{i,t}) = \sigma_i^2$, $1 \leq i \leq N$ and $1 \leq t,\ s,\ \tau \leq T$.

Define $\mathbf{Y}_t = (Y_{1t}, \cdots, Y_{Nt})^\top \in \mathbb{R}^N$, $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_N)^\top \in R^{N\times q}$ $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \cdots, \varepsilon_{Nt})^\top$ and write I_N for the N dimensional identity matrix. The SGNAR model (1) can be represented in compact matrix form:

$$\mathbf{Y}_t = \mathbf{1}\beta_0 + (I_N\beta_1 + A)\mathbf{Y}_{t-1} + \mathbf{Z}^\top\gamma + \boldsymbol{\varepsilon}_t. \tag{2}$$

Our interest is to detect 1) the active groups and 2) the active elements within the active groups, namely to estimate the adjacency matrix A under sparsity. The large size of the network challenges the estimation of the $N \times N$ adjacency matrix, due to the limited data availability with T as the number of observations. This translates to the famous small-T-large-N issue with $T << N^2$.

Under the two-layer sparsity assumption, also referred to as sparse-group, the estimation is achieved by carrying out a nonlinear regularized optimization:

$$\begin{aligned} \min_\theta \frac{1}{2N}\sum_{t=2}^{T} ||\mathbf{Y}_t - \mathbf{1}\beta_0 - (I_N\beta_1 + A)\mathbf{Y}_{t-1} - \mathbf{Z}^\top\gamma||_F^2 \\ + \sum_{i=1}^{N}(1-\alpha)\lambda||A_{\cdot i}||_F + \sum_{i=1}^{N}\sum_{j\neq i}^{N}\alpha\lambda|a_{ij}| \end{aligned} \tag{3}$$

where $\theta = (\beta_0, \beta_1, A, \gamma)^\top$. Group sparsity is applied to the columns of the adjacency matrix. The matrix A is partitioned to $A_{\cdot i}$ with all the elements being 0 except for the ith column, i.e. $A_{\cdot i} = \{A|a_{kj} = 0\,\forall\, k \wedge (j \neq i)\}$, and $A_{\cdot -i} = \{A|a_{kj} = 0\,\forall\, k \wedge (j = i)\}$ with the ith column being 0. Individual sparsity is further applied only to the nonzero columns, namely, the active groups. If a group is inactive, the entire corresponding columns of the adjacency matrix will be shrunk to zero. Here α and λ are the tuning parameters and $||A||_F = \sqrt{\sum_{i,j} a_{ij}^2}$ refers to the Frobenius norm. The term $(1-\alpha)\lambda$

controls the group sparsity and $\alpha\lambda$ the individual sparsity.

3.1 Gradient Descent

To solve the optimization problem, we develop a gradient descent algorithm and iteratively apply it to each column of A. In every iteration, the parameters of a particular group are optimized, while the remaining parameters are fixed.

Specifically, let $A_{\cdot i}$ be the ith column/group to be optimized in an iteration step. The remaining parameters in $A_{\cdot -i}$ are held fixed when optimizing the ith column. We construct the partial residuals of $\mathbf{Y}_t$, which contain the dependence unexplained by the already optimized parameters:

$$\begin{aligned} r_{t,-\beta_0} &= \mathbf{Y}_t - (I_N\beta_1 + A)\mathbf{Y}_{t-1} - \mathbf{Z}^\top\gamma, \\ r_{t,-\beta_1} &= \mathbf{Y}_t - \mathbf{1}\beta_0 - A\mathbf{Y}_{t-1} - \mathbf{Z}^\top\gamma, \\ r_{t,-A_{\cdot i}} &= \mathbf{Y}_t - \mathbf{1}\beta_0 - (I_N\beta_1 + A_{\cdot -i})\mathbf{Y}_{t-1} - \mathbf{Z}^\top\gamma, \\ r_{t,-\gamma} &= \mathbf{Y}_t - \mathbf{1}\beta_0 - (I_N\beta_1 + A)\mathbf{Y}_{t-1}. \end{aligned}$$

The following are the loss functions:

$$\begin{aligned} L(r_{-\beta_0};\beta_0) &= \frac{1}{2N}\sum_{t=2}^{T}||r_{t,-\beta_0} - \mathbf{1}\beta_0||_F^2. \\ L(r_{-\beta_1};\beta_1) &= \frac{1}{2N}\sum_{t=2}^{T}||r_{t,-\beta_1} - I_N\beta_1\mathbf{Y}_{t-1}||_F^2, \\ L(r_{-A_{\cdot i}};A_{\cdot i}) &= \frac{1}{2N}\sum_{t=2}^{T}||r_{t,-A_{\cdot i}} - A_{\cdot i}\mathbf{Y}_{t-1}||_F^2, \\ L(r_{-\gamma};\gamma) &= \frac{1}{2N}\sum_{t=2}^{T}||r_{t,-\gamma} - \mathbf{Z}^\top\gamma||_F^2. \end{aligned}$$

To simplify the notation, let $\theta_1 = \beta_0$, $\theta_2 = \beta_1$, $\theta_3 = \gamma$ and $\theta_k = A_{\cdot i}$, $k =$

$4, \cdots, N+3$. We rewrite the optimization in this particular iterative step as

$$\hat{\theta}_k = \underset{\theta_k}{\operatorname{argmin}} \, L(r_{-\theta_k}; \theta_k) + (1-\alpha)\lambda ||\theta_k||_F + \sum_{i=1}^{N} \alpha\lambda |\theta_{k,i}| \tag{4}$$

where for $k = \{1, 2, 3\}$ the penalty term is set to $\lambda = 0$, namely no sparsity penalization is applied for β_0, β_1 and γ.

There is no closed form solution for the non-convex optimization problem in (3). We introduce a two-step gradient descent algorithm to numerically estimate β_0, β_1, γ and A in the SGNAR framework. We derive the updating function for each iteration step l. Using a Taylor expansion, we formulate an upper bound for $L(r_{-\theta_k^{(l)}}; \theta_k^{(l)})$ depending on the $\theta_k^{(l-1)}$ that has been optimized in the previous iteration step $l-1$. The minimization problem can be equivalently solved by minimizing

$$\begin{aligned} M(\theta_k^{(l)}) =& L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)}) + (\theta_k^{(l)} - \theta_k^{(l-1)})^\top \nabla L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)}) + \frac{1}{2\xi} ||\theta_k^{(l)} - \theta_k^{(l-1)}||_F^2 \\ &+ (1-\alpha)\lambda ||\theta_k^{(l)}||_F + \sum_{i=1}^{N} \alpha\lambda |\theta_{k,i}^{(l)}|, \end{aligned} \tag{5}$$

where ξ is small enough so that the quadratic term dominates the Hessian of the loss function. This approach is also known as the majorize-minimization approach, Wu and Lange (2008). Note in the case of $k = \{1, 2, 3\}$, λ is set to 0.

The first term of Equation (5) does not depend on $\theta_k^{(l)}$, thus it can be further simplified to

$$\begin{aligned} M(\theta_k^{(l)}) \propto& \frac{1}{2\xi} ||\theta_k^{(l)} - \{\theta_k^{(l-1)} - \xi \nabla L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)})\}||_F^2 \\ &+ (1-\alpha)\lambda ||\theta_k^{(l)}||_F + \sum_{i=1}^{N} \alpha\lambda |\theta_{k,i}^{(l)}|. \end{aligned} \tag{6}$$

The loss function is embedded into the thresholding function of the Lasso as follows:

$$S(z, \alpha\lambda) = \text{sign}(z) \circ (|z| - \alpha\lambda)_+,$$

where $\circ$ denotes the Hadamard product. This leads to $\hat{\theta}_k = 0$ if

$$||S\left\{\theta_k^{(l-1)} - \xi \nabla L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)}), \xi\alpha\lambda\right\}||_F \leq \xi(1-\alpha)\lambda$$

and otherwise

$$\left\{1 + \xi(1-\alpha)\lambda / ||\theta_k^{(l)}||_F\right\} \theta_k^{(l)} = S\left\{\theta_k^{(l-1)} - \xi \nabla L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)}), \xi\alpha\lambda\right\}$$

The solution to (6) satisfies

$$\theta_k^{(l)} = \left(1 - \frac{\xi(1-\alpha)\lambda}{||S(\theta_k^{(l-1)} - \xi \nabla L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)}), \xi\alpha\lambda)||_F}\right)_+ S(\theta_k^{(l-1)} - \xi \nabla L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)}), \xi\alpha\lambda). \tag{7}$$

3.2 Algorithm

The two-layer sparsity has both group and individual terms which are inseparably connected. Friedman et al. (2010) outline an idea for an algorithm that would be applicable in such situations. Yet the idea was designed for the univariate case and the groups are defined on the rows. In our multivariate case, we define the group on the columns, as we are looking for leading groups which influence the future values of other groups. This makes the groups dependent on each other.

Simon et al. (2013) formulated the algorithm for univariate regression models, which translates to a regression on a vector of length $(T-1)$. Sparsity of the rows instead of the columns would result in an optimization problem which requires less computation time to find a solution since each group can be optimized independently from the others. However this does not allow a network interpretation. Because of the interdependency of the groups, the optimization problem cannot be written in a vectorized form. The complexity of the model challenges the algorithm, resulting in a longer runtime. We propose a new algorithm customized for the multivariate case with groups defined on the columns. The algorithm initializes with all parameters set to be 0. It iterates through each group of parameters by starting with the parameters β_0, β_1, γ to control for the effects of the intercept, autoregressive dependence, and external influences, before optimizing on the groups in the adjacency matrix A. The algorithm optimizes at the update step width ξ, before the current group θ_k gets updated. The update of θ_k is performed until an a priori chosen vectorized threshold value ϵ_2 is reached. When θ_k has been updated, the next group gets optimized until a full walk through all the groups of parameters has been performed. This procedure repeats until it converges. In detail, the algorithm works as described in Algorithm 1.

The parameter ϵ_1 can be set to any value in $(0, 1)$. Its value controls the density of the grid in which the search for the updating value of parameters takes place. The smaller it is, the faster the algorithm, so one can use it to speed up the computationally intensive method. The entire algorithm works under a chosen mixing parameter α and a penalty parameter λ. The algorithm converges when a vectorized threshold parameter ϵ_3 is satisfied.

The algorithm depends on the hyperparameter λ. It controls the level of penalization, which balances the sparseness of the model against the fit. We derive first which level of λ sets all groups to 0 by following the approach of Simon et al. (2013). The path is started with λ_{max} and from there on a halving sequence is created. In the spirit of Simon et al. (2013), the mixing parameter α is set to be $\alpha = 1/N$, which gives equal importance to group and individual sparsity.

Algorithm 1 SGNAR optimization algorithm

```
Input: Data $\mathbf{Y}_t$ for all $t = 1, \dots, N$
Output: Adjacency matrix $A$
Initialization $\beta_0 = 0,\ \beta_1 = 0,\ \gamma = 0,\ A = 0,\ m = 1$
Set $\theta_1 = \beta_0, \theta_2 = \beta_1, \theta_3 = \gamma, \theta_k = A_{.i}, i = 1, \dots, N, k = i + 3$
while $\text{vec}\{A^{(m)} - A^{(m-1)}\} < \epsilon_3,\ \beta_0^{(m)} - \beta_0^{(m-1)} < \epsilon_3,\ \beta_1^{(m)} - \beta_1^{(m-1)} < \epsilon_3$ or $\gamma^{(m)} - \gamma^{(m-1)} < \epsilon_3$ do
    for $k = 1, \dots, N + 3$ do
        $l = 2$
        while $\theta_k^{(l)} - \theta_k^{(l-1)} < \epsilon_2$ do
            $\xi = 1$
            while $\xi$ small enough such that it holds $L(r_{-U}; U) \leq L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)}) + (U - \theta_k^{(l-1)})^\top \nabla L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)}) + \frac{1}{2\xi}||U - \theta_k^{(l-1)}||^2$ do
                $z = \theta_k^{(l-1)} - \xi * \nabla L(r_{-\theta_k^{(l-1)}}; \theta_k^{(l-1)})$
                $S = \text{sign}(z) \circ (|z| - \alpha\lambda)_+$
                $U = \{1 - \xi(1 - \alpha)\lambda / ||S||\}_+ S$
                $\xi = \epsilon_1 * \xi$
            end while
            $\theta_k^{(l)} = U_{l-1} + \frac{(l-1)}{(l+2)}(U_l - U_{l-1})$
            $l = l + 1$
        end while
    end for
    $m = m + 1$
end while
```

The settings for running Algorithm 1 are as described in Algorithm 2.

Algorithm 2 SGNAR algorithmic procedure

```
1: Run **Algorithm 1** for each $\lambda$ with $J$ iterations
2: Fix identified groups from step 1.
3: To obtain warm starting values, run **Algorithm 1** without penalization for
   identified groups
4: Utilize results from 3. as starting values
5: Run **Algorithm 1** with $\lambda$ sequence
```

4 Real Data: Bitcoin Transaction Analysis

In this section, we analyse the BTC blockchain and implement the SGNAR model to detect the regional and size effects in the global virtual currency transactions in the BTC blockchain network.

4.1 Estimation procedure

We use the BTC transaction data described in Section 2 from February 2012 to July 2017. To provide a better interpretation, the data is demeaned and scaled with the GARCH volatility. As such the magnitudes of the parameters become comparable. The intercept β_0 is not required in the estimation. Since we concentrate our analysis on the transaction streams in the BTC blockchain, we omit γ too. We are modelling the transactions on a daily basis as follows:

$$Y_{i,t} = \beta_1 Y_{i,(t-1)} + \sum_{j=1}^{N} a_{ij} Y_{j,(t-1)} + \varepsilon_{i,t} \tag{8}$$

where the parameters are defined as in (1). We focus on the estimation of the unknown adjacency matrix $A = \{a_{ij}\}$ for network connectivity and the parameter β_1 of serial dependence. To understand the time related dependence in the network, we split the activity into years and perform the estimation independently for each year. In total, there are 6 samples; each contains the daily transactions of the 60 groups within the particular year.

The estimation relies on the choice of α and λ as defined in the regularized optimization (3). The mixing parameter α is set to be $1/N$, where N is the number of groups: in this case $N = 60$. The choice of λ on the other hand is data-driven. Although cross-validation is a standard technique, it ignores the serial dependence in time series, see Nicholson et al. (2017). Hence, we use a forward-looking criterion by selecting λ such that the out-of-sample forecast errors, measured by the root mean squared error (RMSE), are minimized on the next year's data. This approach was also used in Bańbura et al. (2010), Song and Bickel (2011) and Nicholson et al. (2017). As an example, for the period of 2015, the SGNAR estimation is conducted on the sample period from 1 January 2015 to 31 December 2015. The hyperparameter λ is selected such that the forecasts for the next 196 days from 1 January 2016, computed with the adjacency matrix and β_1 estimated in 2015, has the minimal out-of-sample RMSE among all the alternatives. We chose 196 days to have comparable results given that the 2017 data are only available until mid-July. By means of this procedure we use the maximum available amount of data over all years, which ensures comparability between years and a maximum of used information. Since SGNAR is a method dependent on the evaluation period, a consistent choice over all periods is crucial for a meaningful comparison. Note that the observation for 17 July was omitted, since it changes the result in an unreasonable manner and thus is considered as an outlier. To select λ, we carried out this estimation exercise for each period from 2012 to 2016 until we reached the end of the sample, i.e. 2017.

4.2 Evaluation

We evaluate the estimation performance using metrics including the in-sample R^2 for the year estimated, out-of-sample R^2 for the next year's observations, in-sample RMSE and out-of-sample RMSE. The first two metrics measure the predictability, and the last two measure the variation around the realized value.

We illustrate the serial cross-dependence with chord diagrams. They demonstrate the essential dynamic connectivity in the global BTC blockchain network. A chord diagram displays the direction and magnitude of the influence of each node by showing the magnitude by means of the circle and the destination of the signal by the chord. The wider the space on the circle, the larger the magnitude and hence the higher the dynamic impact in the network. A chord diagram does not differentiate between positive and negative influences. The sum of the absolute values of the parameters (magnitude) is displayed on the circle. Moreover, the colour of the chord corresponds to the colour of the continent to which the effect is directed. For an example, consider Figure 5e, where EU.1 is outstanding with a magnitude of 12.5 and about one-third of the magnitude directly influencing the other European groups. The remaining magnitude mostly reflects an influence on North American, South American, Oceanian and Asian groups.

4.3 Results and interpretation: 10-groups

SGNAR is applied to the grouped transactions (10 groups per continent) and we tackle the problem of answering if user groups defined by transactions are related in a time dependent manner between years. In the year-to-year analysis, zero entries in the adjacency matrix indicate that the past transactions of the corresponding group have no influence on the future transactions of another group. If there are only zero entries in one column, this indicates the lack of network connectivity of the particular regional size group with all the other groups. On the other hand, a group with a non-zero entry in the adjacency matrix is considered as an active group as it is able to influence the dynamic evolution of the virtual money flows.

We focus on the years where the adjacency matrices are not zero, in other words, where network effects appeared. Figure 5 illustrates the active network connectivity based on the 10 groups per continent over the whole sample period. It shows there are network effects for 2012 and 2016. 2012 was the year when BTC received increasing attention. Its price doubled before it skyrocketed in 2013 with its price reaching over 1000 USD for the first time in November 2013. Simultaneously with a decreasing price evolution and a period of the creation of a plethora of alternative CCs, the BTC Blockchain showed no network effects. In 2016, akin to 2012, the price doubled on the exchanges before it skyrocketed at the end of 2017. In particular, we observe network effects for 2012 and 196 days of the evaluation period coming from a small group in North America, NA.1, the smallest group from Asia, AS.1, as too the top group in South America, a medium and top group in Africa. In terms of the magnitude[5] of the groups, SA.10 has about the same magnitude as the North American and medium African groups combined, hence stronger network effects come from users who move larger transactions from South America. One should note that the transaction amounts in SA.10 are about the size of a medium group in North America. This picture changes strongly in 2013. From 2013 to 2015 not a single group is active, hence we observe a decreasing network connectivity. In contrast, in 2016 effects again become visible. The smallest groups from Europe and Africa become active. For Africa one observes that the network effects are frequently directed towards South America and Africa itself. The frequent drops to 0 transactions in 2017 of these two continents are an explanation for the active state of AF.1, compare Figure 1. The

[5] Recall that the data are standardized, hence the magnitudes of the parameters are comparable in their values.

network effects of AF.1 reflect this dynamics within South America and Africa. The group EU.1 has a substantially larger magnitude and sends network effects to all continents, particularly strong ones to other European groups and South America, followed by Oceania and Asia. Over the years, one observes strong changes in the network. After all, Bitcoin is an emerging asset and its usage has been changing over time. One can only speculate about the actual usage, yet major known activities include gambling and trading.

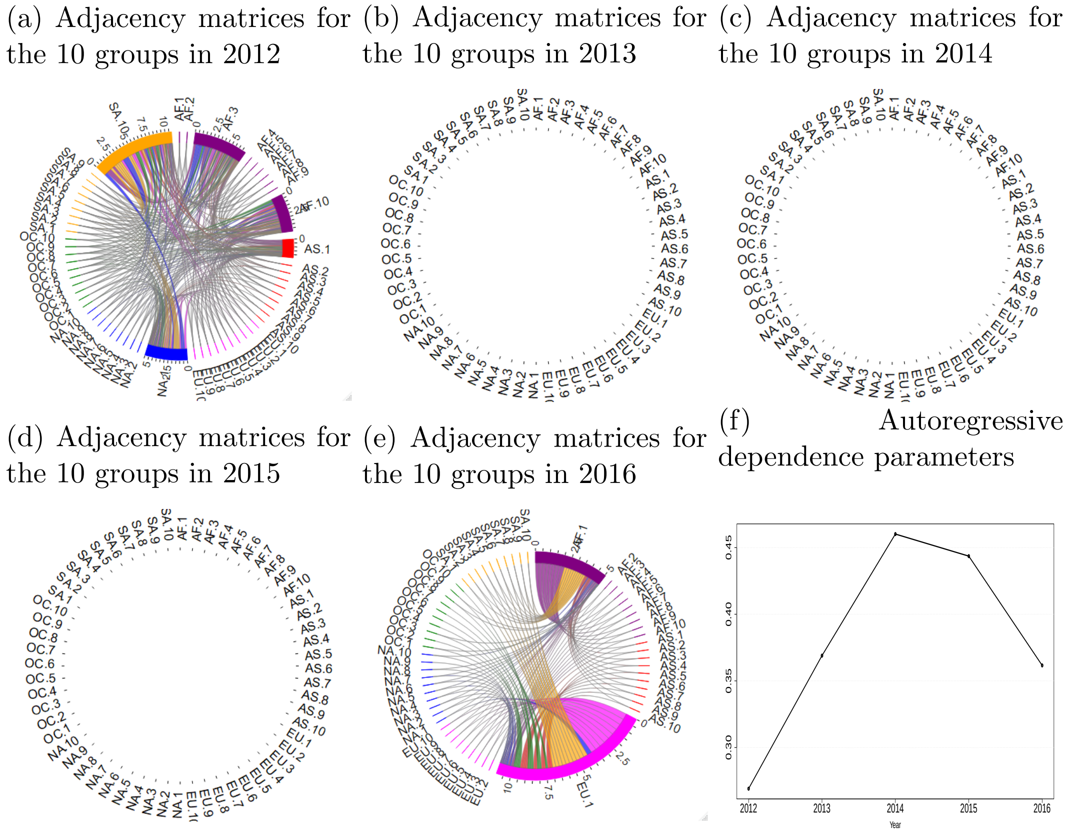
The transactions before 2016 showed decreasing network effects from daily data on the next days observations, but this picture seems to change in 2016. The groups and continents involved seem surprising to some extent, because media reports often focus on the roles of CCs in Asia rather than in Europe, especially in terms of mining. But comparing the time series plots, Figure 2, it is obvious the volumes of transactions in Europe and North America are higher than in other regions. This gives a good rationale for the effects coming from these two regions, even from smaller groups like EU.1. Further support for this finding comes from the surprising number of null values in Africa, South America and Oceania. Explanations for these values may be that users from these regions switched to other CCs, since in this period a bunch of altcoins (CCs other than BTC) became important. Secondly, the number of transactions in the BTC blockchain increased strongly (150%) in this time, see Figure 6, and simultaneously the maximum block size of 1 Megabyte was reached. Since each block of transactions has a limit on the possible number of included transactions, it is likely that certain users from Europe completely dominated the transaction chain in this period. Via the willingness to pay transaction fees, the miners' decisions about privileging one transaction over another by including it into the next block can be influenced. The miners have an incentive to include small transactions which pay high transaction fees, since by this action they can maximize their personal profit. Hence the respective transactions would be prioritized, which leads to the conclusion that high value transactions originated from the continents detected in the analysis. This provides good evidence for the economic reasons described for our finding. The limit on the possible number of transactions included in each block led the developers to introduce a BTC without this restriction, called Bitcoin Cash (BCH) on 1 August 2017. This event was a fork of the BTC source code in which the code was amended so that it would fit the features wished for. As a result, BTC and BCH exist as individual CCs. Finally, it can be inferred that the blockchain transactions developed a more dense network pattern in the last two years, with network effects from Europe and Africa, which was fostered by market reasoning.

Table 3: R^2, RMSE and λ penalties for in- and out-of-sample performance of the models found in the respective years for 10 groups per continent with 196 days evaluation period length.

	R^2 in	R^2 out	RMSE in	RMSE out	$\alpha * \lambda$	$(1-\alpha) * \lambda$
2012	0.17	0.20	0.87	0.99	0.01	0.54
2013	0.24	0.24	0.93	0.92	0.04	2.45
2014	0.27	0.35	0.90	0.85	0.02	1.41
2015	0.28	0.26	0.92	0.88	0.04	2.12
2016	0.21	0.27	0.85	0.70	0.01	0.73

Having a closer look at the actual values assigned to the diagonal of the adjacency matrix, we observe they are quite low or of medium strength. In the first 4 years, the autoregressive part takes on the estimates 0.27, 0.37, 0.46 and 0.44. Apparently the autoregressive effect is not strong in these years. In 2016 the autoregressive parameter is again quite low, taking the value 0.36. It is a strong sign that the effects got assigned to the columns instead of the diagonal. In case the network effects could be represented by the autoregressive dependence (diagonal), the effects would be assigned accordingly, which is a feature of SGNAR. Yet since this is not the case, it is a strong hint the network effects are of sufficient importance to be considered in the SGNAR network. Considering the R^2 and RMSE for the in- and out-of-sample analysis, presented in Table 3, one observes a high overall out-of-sample prediction

Figure 5: Adjacency matrices and serial dependence parameter in analysis with 10 groups and evaluation period length of 196 days.



(a) Adjacency matrices for the 10 groups in 2012

(b) Adjacency matrices for the 10 groups in 2013

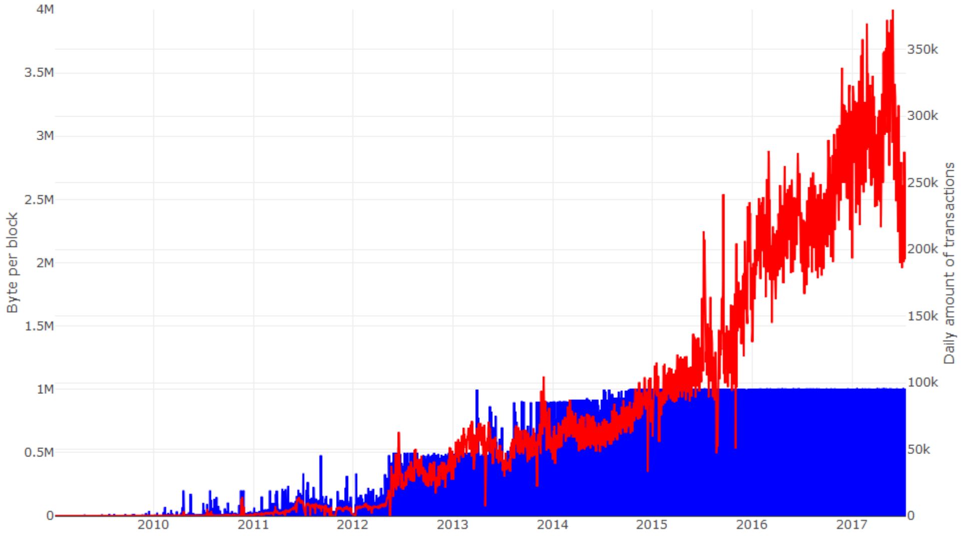
(c) Adjacency matrices for the 10 groups in 2014

(d) Adjacency matrices for the 10 groups in 2015

(e) Adjacency matrices for the 10 groups in 2016

(f) Autoregressive dependence parameters

Figure 6: Daily Bitcoin Transactions (line) and the Block Size of Bitcoin (filled area) in the time period 4 January 2009 to 17 July 2017.



accuracy since all R^2 are above 20%. 2016 is a notable year since it is when the model shows strong network effects after no effects in 2013–2015. The RMSE shrinks a lot compared to earlier years' models. It is notable that the out-of-sample RMSE is much lower than the in-sample one for 2016, hinting that the year-to-year structure becomes more related. Considering the regularization parameters in 2013–2015, the high λ penalties indicate the network effects were spurious, hence not containing information, and therefore shrunk to 0. The much lower valued of λ in 2012 and 2016 show the importance of the identified network effects and that hardly any other effects exist since little penalization is necessary.

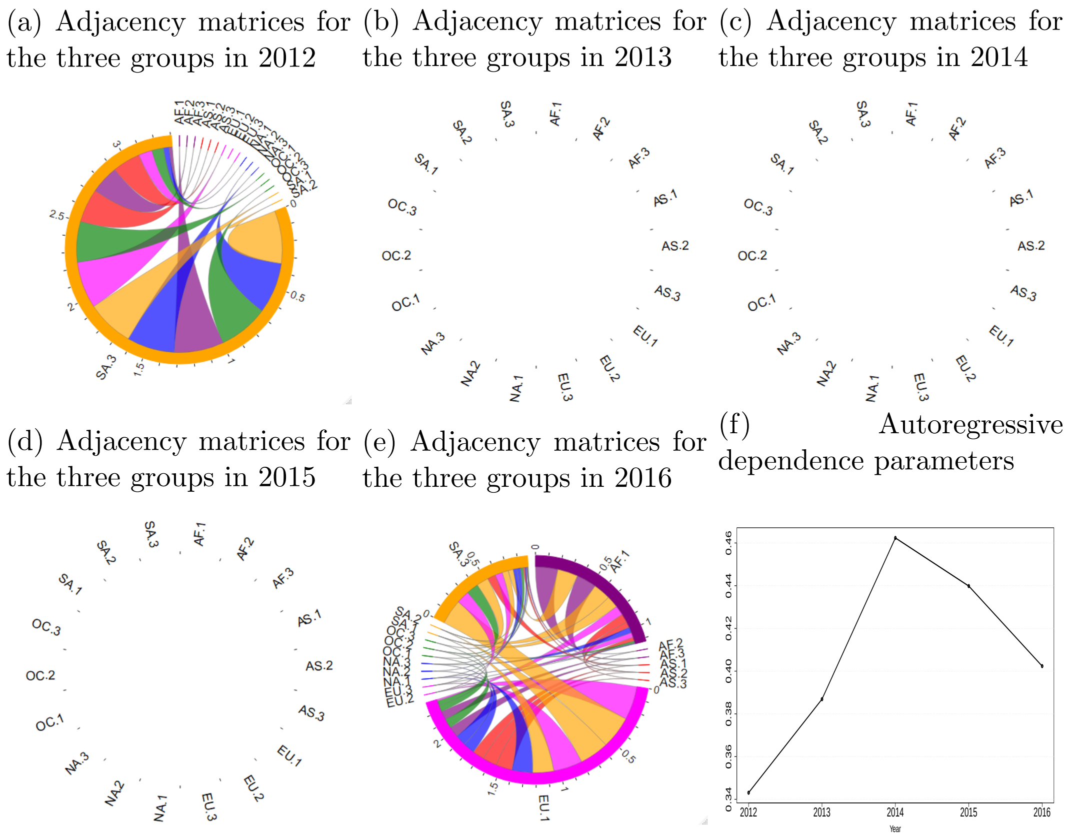
4.4 Robustness check with alternative grouping: 3-groups

In order to understand whether the regional and size effects are robust to the grouping, we carried out robustness checks with alternative three groups. For the three groups, users in each continent are further split into small, medium, and big groups, according to the sizes of their transactions. The other model settings remain the same as previous. The resulting adjacency matrices for the three groups per continent are illustrated in Figure 7, which again were obtained by minimizing the out-of-sample RMSE on 196 days forecast.

Concerning the analysis with three groups per continent, the Bitcoin Blockchain showed network effects from big users in South America and small users in Europe and Africa, SA.3, EU.1 and AF.1. Autoregressive dependence effects are illustrated in Figures 7f and indicate a medium sized connectivity within the groups. Interestingly, the autoregressive dependence parameter shrinks markedly in 2016, when strong network effects on 2017 become apparent. No network effects were uncovered for 2013–2015. Therefore the observation of no network activity is robust to the 10-grouping. Again in 2016, strong network effects were uncovered, in terms of connectivity as well as in terms of their magnitude. The previous result from the 10-grouping is therefore robust.

The network effects in 2016 originate from EU.1 and AF.1, which is consistent with the 10-group case. Also South America is now active and its influence is frequently directed towards itself and Europe. In contrast, Europe mostly influences other continents, rather than itself. Table 4 gives the in- and out-of-sample R^2 as well as the RMSE. The out-of-sample R^2 is in all the years above 20%, indicating a high prediction accuracy. The RMSE stays more or less in the same range, just in 2016 it was much lower out-of-sample. Interestingly, the out-of-sample RMSE stayed lower than for 2012, when several network effects were detected, hinting at a more difficult data structure to be modelled in 2012 compared to 2016.

Figure 7: Adjacency matrices and serial dependence parameter in analysis with three groups and evaluation period length of 196 days.



(a) Adjacency matrices for the three groups in 2012

(b) Adjacency matrices for the three groups in 2013

(c) Adjacency matrices for the three groups in 2014

(d) Adjacency matrices for the three groups in 2015

(e) Adjacency matrices for the three groups in 2016

(f) Autoregressive dependence parameters

Table 4: R^2, RMSE and λ penalties for in- and out-of-sample performance of the models found in the respective years for 3 groups per continent with 196 days evaluation period length.

	R^2 in	R^2 out	RMSE in	RMSE out	$\alpha * \lambda$	$(1-\alpha) * \lambda$
2012	0.18	0.23	0.88	0.99	0.03	0.51
2013	0.25	0.25	0.91	0.91	0.07	1.24
2014	0.27	0.37	0.90	0.81	0.04	0.67
2015	0.27	0.25	0.92	0.91	0.06	1.02
2016	0.24	0.30	0.88	0.72	0.02	0.33

5 Conclusion

Cryptocurrencies have become interesting asset classes. BTC, being the elephant in the room, is traded all over the globe and virtually uncorrelated to any other asset class, which in principle is good for purposes of diversification. Besides the trading data on the exchanges, the blockchain displays a second layer of transactions, which are the actual shifts of funds directly between users without a middleman. The anonymity of the blockchain challenges analysis, even though understanding the state of the network is important to understand a cryptocurrency. For the analysis of the blockchain, the huge dimensionality of the blockchain is challenging. We have proposed a Sparse-Group Network AutoRegressive (SGNAR) model to analyse the time dependent network relations between the users of the BTC blockchain. We have provided an algorithm to derive the adjacency matrix of SGNAR and find spatial connections in the BTC blockchain. We found the global connectivity remained low in the period from 2013 to 2015, although it was high in 2012 and enhanced in the recent years of 2016 and 2017, driven by specific user groups from all over the globe. Taking into account that most Bitcoin mining farms are in Asia, it is surprising to some extent that Asia is not the sole driver but operates Bitcoin for Europe, North America, Africa and South America, fostering the importance of these regions in the blockchain. In particular we found that mostly the users with the smallest and largest sizes of transactions from North America, Europe, South America, Africa and Asia were driving the Bitcoin transactions, while the other groups and all the groups in Oceania were either followers or isolated.

References

Bańbura, M., D. Giannone, and L. Reichlin (2010). “Large Bayesian vector auto regressions”. *Journal of Applied Econometrics* 25.1, pp. 71–92.

Basu, S. and G. Michailidis (2015). “Regularized estimation in sparse high-dimensional time series models”. *The Annals of Statistics* 43.4, pp. 1535–1567.

Chudik, A. and M. H. Pesaran (2011). “Infinite-dimensional VARs and factor models”. *Journal of Econometrics* 163.1, pp. 4–22.

Creal, D., S. J. Koopman, and A. Lucas (2013). “Generalized Autoregressive Score Models with Applications”. *Journal of Applied Econometrics* 28.5, pp. 777–795.

Darlington III, J. K. (2014). “The Future of Bitcoin: Mapping the Global Adoption of World’s Largest Cryptocurrency Through Benefit Analysis”. *University of Tennessee Honors Thesis.*

Dees, S., F. d. Mauro, M. H. Pesaran, and L. V. Smith (2007). “Exploring the international linkages of the euro area: a global VAR analysis”. *Journal of Applied Econometrics* 22.1, pp. 1–38.

Elendner, H., S. Trimborn, B. Ong, and T. M. Lee (2017). “The Cross-Section of Crypto-Currencies as Financial Assets: Investing in Crypto-currencies beyond Bitcoin”. *Handbook of Blockchain, Digital Finance and Inclusion: Cryptocurrency, FinTech, InsurTech, and Regulation.* Ed. by D. Lee Kuo Chuen and R. Deng. 1st ed. Vol. 1. Elsevier, pp. 145–173.

Feng, W., Y. Wang, and Z. Zhang (2017). “Informed trading in the Bitcoin market”. *Finance Research Letters.*

Feng, W., Y. Wang, and Z. Zhang (2018). “Can cryptocurrencies be a safe haven: a tail risk perspective analysis”. *Applied Economics* 50.44, pp. 4745–4762.

Friedman, J., T. Hastie, and R. Tibshirani (2010). “A note on the group lasso and a sparse group lasso”.

Guo, J., J. Hu, B.-Y. Jing, and Z. Zhang (2016). “Spline-Lasso in High-Dimensional Linear Regression”. *Journal of the American Statistical Association* 111.513, pp. 288–297.

Hafner, C. (2018). “Testing for Bubbles in Cryptocurrencies with Time-Varying Volatility”. *Journal of Financial Econometrics.*

Hoerl, A. and R. Kennard (1988). “Ridge regression”. *in Encyclopedia of Statistical Sciences* 8, pp. 129–136.

Kristoufek, L. (2015). “What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis”. *PLOS ONE* 10, pp. 1–15.

Lischke, M. and B. Fabian (2016). “Analyzing the Bitcoin Network: The First Four Years”. *Future Internet* 8.1.

Nicholson, W., D. Matteson, and J. Bien (2017). “VARX-L: Structured Regularization for Large Vector Autoregressions with Exogenous Variables”. *International Journal of Forecasting* 33.3, pp. 627–651.

Ord, K. (1975). “Estimation Methods for Models of Spatial Interaction”. *Journal of the American Statistical Association* 70.349, pp. 120–126.

Pesaran, M. H., T. Schuermann, and S. M. Weiner (2004). “Modeling Regional Interdependencies Using a Global Error-Correcting Macroeconometric Model”. *Journal of Business & Economic Statistics* 22.2, pp. 129–162.

Reid, F. and M. Harrigan (2013). “An Analysis of Anonymity in the Bitcoin System”. *Security and Privacy in Social Networks.* Ed. by Y. Altshuler, Y. Elovici, A. B. Cremers, N. Aharony, and A. Pentland. Springer New York, pp. 197–223.

Ron, D. and A. Shamir (2013). “Quantitative Analysis of the Full Bitcoin Transaction Graph”. *Financial Cryptography and Data Security.* Ed. by A.-R. Sadeghi. Lecture Notes in Computer Science 7859. Springer Berlin Heidelberg, pp. 6–24.

Scaillet, O., A. Treccani, and C. Trevisan (2018). “High-frequency jump analysis of the bitcoin market”. *Journal of Financial Econometrics.*

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). “A Sparse-Group Lasso”. *Journal of Computational and Graphical Statistics* 22.2, pp. 231–245.

Song, S. and P. J. Bickel (2011). “Large vector auto regressions”. *arXiv preprint arXiv:1106.3915.*

Tibshirani, R. (1996). “Regression Shrinkage and Selection via the Lasso”. *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, pp. 267–288.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). “Sparsity and smoothness via the fused lasso”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1, pp. 91–108.

Wu, T. T. and K. Lange (2008). “Coordinate descent algorithms for lasso penalized regression”. *The Annals of Applied Statistics* 2.1, pp. 224–244.

Yuan, M. and Y. Lin (2006). “Model selection and estimation in regression with grouped variables”. *Journal of the Royal Statistical Society, Series B* 68, pp. 49–67.

Zhou, J., Y. Tu, Y. Chen, and H. Wang (2017). “Estimating Spatial Autocorrelation With Sampled Network Data”. *Journal of Business & Economic Statistics* 35.1, pp. 130–138.

Zhu, X., R. Pan, G. Li, Y. Liu, and H. Wang (2017). “Network vector autoregression”. *The Annals of Statistics* 45.3, pp. 1096–1123.

Zou, H. and T. Hastie (2005). “Regularization and variable selection via the elastic net”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.