Bitcoin’s Crypto Flow Network

Yoshi Fujiwara¹ and Rubaiyat Islam¹

¹Graduate School of Information Science, University of Hyogo, Kobe 650-0047, Japan
E-mail: yoshi.fujiwara@gmail.com

(Received May 31, 2021)

How crypto flows among Bitcoin users is an important question for understanding the structure and dynamics of the cryptoasset at a global scale. We compiled all the blockchain data of Bitcoin from its genesis to the year 2020, identified users from anonymous addresses of wallets, and constructed monthly snapshots of networks by focusing on regular users as big players. We apply the methods of bow-tie structure and Hodge decomposition in order to locate the users in the upstream, downstream, and core of the entire crypto flow. Additionally, we reveal principal components hidden in the flow by using non-negative matrix factorization, which we interpret as a probabilistic model. We show that the model is equivalent to a probabilistic latent semantic analysis in natural language processing, enabling us to estimate the number of such hidden components. Moreover, we find that the bow-tie structure and the principal components are quite stable among those big players. This study can be a solid basis on which one can further investigate the temporal change of crypto flow, entry and exit of big players, and so forth.

Presented at the conference “Blockchain in Kyoto 2021”, and forthcoming in the JPS Conference Proceedings.

KEYWORDS: Bitcoin, cryptoasset, bow-tie structure, Hodge decomposition, non-negative matrix factorization, latent Dirichlet allocation, complex network

1. Introduction

Cryptoasset or cryptocurrency is essentially a digital ledger to record transactions between creditors and debtors, just like money. The digital system is based on a collection of non-centralized ledgers, called blockchain, which contains all the historical record of transactions among anonymous users. Today there are many cryptoassets being exchanged in markets with fiat currencies and also with each other. The market capitalization is so huge in total ranging from one to a few trillion USD, and highly volatile potentially having a big impact even on asset markets and prices of non-crypto at a global scale.

In this paper, we study Bitcoin, the largest one dominating nearly half of the market capitalization at the time of writing. We attempt to understand the flow of crypto as a complex network comprising of the users as nodes and the crypto flow as links. There are a number of studies from such a viewpoint of complex network on cryptoassets. See [1–16] for example, and references therein.

Specifically, in this paper, we focus on “big players” who are defined as persistently appearing users, likely to be involved in transactions of high frequencies and large amounts, and address the following questions. First, it is important to identify the users in the upstream, downstream, or core in the entire crypto flow. We shall examine the so-called “bow-tie” structure of the network of those big players to classify the location of the crypto flow based on the binary relationship of links. Second, we measure the location in a more quantitative way by using the information of flow along the links in the combinatorial method of Hodge decomposition. Third question is to extract “principal components” hidden in the entire crypto flow so as to uncover a certain number of latent factors or components.
Fourth, because the network is changing in time, what can one say about the stability of the crypto flow?

In Section 2, we describe our dataset of Bitcoin to identify the anonymous addresses representing wallets into users. Then we define regular users as big players and construct networks. In Section 3.1, we perform the bow-tie analysis to locate users in the stream of crypto flow. In Section 3.2, we use the method of Hodge decomposition to quantify the location of users. In Section 3.3, we introduce non-negative matrix factorization as a method of matrix decomposition to reveal principal components hidden in the flow. We shall show in Section 3.4 that the method can be interpreted as a probabilistic model, from which one can estimate the number of components. In Section 3.5, we find about a dozen of principal components among several hundred big players, and show that the temporal change of the network is quite stable. In Section 4, we discuss about several aspects, being worth further investigation, and conclude in Section 5. We add Appendix A for the identities (actual names, business, and so forth) of selected users. Appendix B illustrates the networks in adjacency matrices. Appendix C briefly summarizes the above mentioned probabilistic model in relation to latent Dirichlet allocation, from which the estimation on the number of components is done in Appendix D.

2. Data

We employ the dataset of all transactions recorded in the Bitcoin blockchain from the genesis block (first block issued on January 9, 2009) until the block of height 63,299 (inclusive; issued on June 4, 2020). Each transaction is a transfer of a certain amount of BTC (monetary unit of Bitcoin) from one or more addresses to others as we will see shortly. We call such a transfer of BTC as crypto flow. An address is something like a wallet possessed by a user who can be an individual or, more frequently today, an agent in the business of exchanges, services, gambling, and so forth.

In the dataset, total number of transactions was 1.38 billion, while the number of different addresses was about 657 million. To study crypto flow of Bitcoin, one needs to know users, rather than the addresses. However, it is not straightforward to identify users from addresses because of the very nature of anonymity inherent in the core technology of blockchain. See [17] for technical details.

Let us employ a simple but useful method to identify users from addresses to construct a giant graph comprising of nodes as users and edges as crypto flow. We shall see that more than 60% of the addresses can be identified with users. Additionally we will see that a number of users can be revealed with their actual names, types of business, and sometimes geographical location at a global scale. Then we define regular users as big players in order to focus on a subgraph comprising of frequently appearing users who are involved in crypto flow with huge amounts of BTC. The subgraph will be studied in the subsequent sections.

2.1 Identification of Users from Addresses

Consider an example of such a transaction (TX) that one day Alice transferred 1 BTC to Bob:

$$\text{TX}_1 : \{a_1, a_2\} \rightarrow \{a_{123}, a_1\}, \quad (1)$$

where the addresses $a_1$ and $a_2$ belong to Alice, while $a_{123}$ belongs to Bob. Alice needs more than one address as input of TX$_1$, because a single one was not sufficient to fulfill the amount of 1 BTC. Output of TX$_1$ includes $a_1$ representing the change. Another day, Alice did another transaction:

$$\text{TX}_2 : \{a_1, a_3\} \rightarrow \{a_{45}, a_3\}, \quad (2)$$

where the address $a_3$ also belongs to Alice. Obviously, multiple addresses, if and only if they appear in an input of a transaction, belong to the same user, namely her wallets. As a consequence from both of (1) and (2), it follows that $a_1$, $a_2$, $a_3$ can be identified to belong to the same user. Note that $a_2$ and
Fig. 1. Rank-size plot for the number of addresses identified with users. Size (horizontal) is the number of addresses identified with users of type A (a group of addresses appearing in multiple inputs of transactions, so identified as user; see main text). Rank (vertical) is the descending order of the size. Maximum size is nearly 20 million, while minimum is 2 and average is 6.7.

$\alpha_3$ did not appear in any record of transactions, TX$_1$ and TX$_2$. By looking at all the transaction in the history, one can identify many addresses with users.

This simple but useful method to identify users from addresses was proposed by [1] and has been extensively used in the literature (see [4, 5] and the data of [18, 19], for example). We implemented an efficient algorithm of this method to process the above mentioned 657 million addresses in the 1.375 billion transactions. We found that 402 million addresses (more than 60%) can be identified to obtain 60 million different users (denoted as type A). The rest of 255 million addresses, which never appeared as multiple inputs in any transactions, are regarded as different users (type B). This procedure results in 315 million users all together.

Fig. 1 shows a rank-size plot, in which the size is the number of addresses identified with users of type A, and the rank is its descending order. The minimum size is 2 by construction, while the average is 6.7. It is interesting to observe that the rank-size plot is highly skewed with the maximum size being nearly 20 million! We labeled all the users of type A in a sequential user ID corresponding to the rank (e.g. 0000012345), while users of type B is labeled with its address (e.g. 3CjqmbuRA1LEWmLHiWoSWHcWuTEVPfu24P).

A bunch of transactions is packed into a block in the blockchain. Each block has a timestamp of its birth, which is UTC (Coordinated Universal Time) when the block was mined or born as an empty ledger to be filled with transactions. We associate the time with the transactions contained in the block. Each block is mined in a mean time of 20 minutes or longer, as is well known, so it should be understood that by time we mean a rough estimate on when the transaction was made within an accuracy of an hour or so. One can use this information of the timestamp to obtain the intra-day activities of users. Fig. 2 is a schematic diagram for typical three users, presumably located in Asia, Europe including Africa, and USA.

Moreover, with the help of laborious investigation by [19], it is possible to unravel the identity of users of type A with actual names, types of business, and sometimes geographical location for many users. Appendix A is the summary of the unraveled identity. Table A-1 gives the classification of business into exchanges, gambling, pools, services and others. Nearly one third in the list are exchanges. Table A-2 is the list of countries of those exchanges, where China, UK, and USA are
2.2 Crypto Flow Network

Now all the transactions among addresses are converted into transfers from users to users, each of which has the following information:

- user of source \( s \),
- user of destination \( d \),
- amount of Bitcoin transferred from \( s \) to \( d \), i.e. \( s \to d \),
- UTC time of transfer (from the block containing the transaction).

During a certain period of time \( T \), for a pair of users, there can be more than one transfer as depicted in Fig. 3 (see the left-hand side). In this example, there are three transfer of crypto for \( i \to j \), one for \( j \to i \), and also two for \( i \to i \). The last case of self-loop is possible, because one can receive a change in a transaction, and also because different addresses are possibly identified with a user, like

\[ \text{Cryptoasset of XRP has a similar weekly pattern according to [20]. In the early history of Bitcoin, the weekly pattern was quite the opposite: more active in weekends, presumably because individuals were big players at that time.} \]
Fig. 3. Aggregation of transactions to construct crypto flow network. Left: During a given period of time, for a pair of users $i$ and $j$, three transfers for $i \rightarrow j$ with 0.1, 0.2, 0.4 in BTC, one of $j \rightarrow i$ with 0.1, and two for $i \rightarrow i$ with 1 and 2. Right: After aggregation, one has $i \rightarrow j$ with frequency $f_{ij} = 3$ and amount of flow $g_{ij} = 0.7$. Similarly, $j \rightarrow i$ with $f_{ji} = 1$ and amount of flow $g_{ji} = 0.1$, and $i \rightarrow i$ with $f_{ii} = 2$ and $g_{ii} = 3$.

an exchange. Given a time-scale $T$, it would be reasonable to aggregate these transactions as shown in the right-hand side of Fig. 3.

After the aggregation, one has a network comprising of nodes as users and edges with direction given by the transfer of crypto, frequency and amount of transfer occurred during the period of time. Let us denote the following variables which represent the strength or weight of each edge by

- **frequency** $f_{ij}$ ≡ frequency of transfers for $i \rightarrow j$.
- **amount of flow** $g_{ij}$ ≡ amount of total transfers for $i \rightarrow j$.

Regarding the time-scale $T$ for aggregating the transactions and the epoch to select in the historical data, we choose one month and the calendar year of 2019. By examining the time-series for the daily number and amount of transactions, we assumed that the period of one month is adequate to study the stability and temporal change of the crypto flow. Shorter period may lead to a trivial result for the stability and could be insufficient to detect the temporal change, if any change is present. Also longer period would be misleading due to non-equilibrium nature of the system. The year 2019 was chosen as the epoch, in which one does not see violent bubble or crush in the price.

The number of all the users is huge, 315 million users in total. Fortunately, however, it is not necessary to include all of them, because most of them do not appear frequently. In the next section, we shall extract only a tiny part of the network by focusing on the “regular users” who appeared everyday during the specified period.

### 2.3 Regular Users as Big Players

For our purpose in this paper, it is sufficient to focus on the crypto flow with high frequency and big amount of Bitcoin, because infrequent and/or small amount of flows is obviously unimportant for the understanding the entire flow. In other words, it suffices to focus on “big players” who are playing some dominant role in the game of crypto flow. It would be possible to define such a big player in different ways. In this study, we define it by looking at how persistently the user appears in transactions during our specified period of time.

Fig. 4 depicts examples of users who appear in different numbers of transactions on daily basis. The user of the top case is persistent in committing transactions with other users, which can be labeled as a “regular user”. The middle case changes the persistency from being inactive with no transaction to being active in an abrupt way. The bottom case has little activity, just a few transactions on particular days having a strong intermittency.

We define **regular users** as those appearing everyday during the period of one year in 2019, and use them as big players. The number of regular users was 479. Then we construct a subgraph in each month, which is comprised of the regular users as nodes and the crypto flow as links, the latter of which are aggregated as described in Fig. 3. Thus we have 12 snapshots of such subgraphs, each
Fig. 4. Illustrative examples for the activities of users. Each plot depicts the daily number of transactions, in each of which the user is either source or destination of the transaction for the period of year 2019. Self-loops (case of the same source and destination) are excluded. Top is a “regular” user appearing every day. The middle user became active from being inactive, while the bottom one has intermittency in its activity. We focus on regular users in this paper.

Fig. 5. Activities of regular users, defined by each user’s peak of transactions, in UTC. Shaded region corresponds to the daytime of Asia, Europe, and USA (left to right).

corresponding to each month in the year, from January to December. Summarizing the processing of the whole dataset, we constructed the snapshots of networks denoted by $G_t = (V_t, E_t)$, where $t$ is the month, $V_t$ is the set of regular users, $E_t$ is the set of links among them, each having the frequency and amount as depicted in Fig. 3.

We add Fig. 5 showing the histogram for the UTC time of highest activities of all the regular
users in the year of 2019. This information gives geographical locations of those regular users.

3. Analysis of Crypto Flow Network

3.1 Basic Properties of Network

For each month $t$, we constructed a network denoted by $G_t = (V_t, E_t)$ as described in the preceding section. Table I is the summary of basic properties of the networks in the year 2019. $V_t$ corresponds to the regular users in each month, the number of which is shown by the column $|V_t|$. The column $|E_t|$ is the number of edges, namely the number of different crypto flow from one user to another or to itself (self-loops as shown in parentheses). Most of the users have self-loops. Temporal change of the network causes the changes of $V_t$ and $E_t$. The column of $|V_t \cap V_{t+1}|$ is the number of users that are common to successive months in their appearance. One can see that most of the users are appearing successively. The same is true for the edges as shown in the column of $|E_t \cap E_{t+1}|$. In other words, the network is not changing drastically in terms of the entry and exit of nodes and edges during the time-scale of months.

| $t$ | $|V_t|$ | $|E_t|$ | $|V_t \cap V_{t+1}|$ | $|E_t \cap E_{t+1}|$ | GWCC | GSCC/IN/OUT/TE |
|-----|-------|-------|----------------|----------------|------|----------------|
| 01  | 470   | 17,215(408) | 468 | 13,313 | 468(3) | 327/24/113/4 |
| 02  | 470   | 16,658(407) | 468 | 13,357 | 468(3) | 322/20/120/6 |
| 03  | 473   | 17,618(408) | 471 | 13,942 | 470(4) | 327/25/113/5 |
| 04  | 473   | 17,691(415) | 470 | 13,960 | 469(5) | 336/16/115/2 |
| 05  | 472   | 17,903(416) | 471 | 14,012 | 468(5) | 330/23/112/3 |
| 06  | 471   | 17,787(412) | 469 | 13,705 | 467(5) | 318/23/124/2 |
| 07  | 472   | 17,292(410) | 468 | 13,108 | 470(3) | 333/21/110/6 |
| 08  | 468   | 16,534(402) | 467 | 12,628 | 466(3) | 321/18/123/4 |
| 09  | 470   | 16,288(407) | 468 | 12,652 | 466(5) | 325/23/113/5 |
| 10  | 469   | 16,317(402) | 468 | 12,476 | 467(3) | 322/21/119/5 |
| 11  | 470   | 15,859(409) | 466 | 12,080 | 468(3) | 324/23/116/5 |
| 12  | 466   | 15,584(390) | —    | —     | 466(1) | 323/18/120/5 |

Each column represents the following.

- $t =$ month of the year 2019
- $|V_t|$ = number of nodes
- $|E_t|$ = number of edges (number of self-loops in parentheses)
- $|V_t \cap V_{t+1}|$ = number of nodes common to successive months
- $|E_t \cap E_{t+1}|$ = number of edges common to successive months
- GWCC = number of nodes in giant weakly connected component (number of components in parentheses)
- GSCC/IN/OUT/TE = number of nodes in giant strongly connected component/IN/OUT/tendrils

We found that most of the users have self-loops, as shown in the parentheses of column $|E_t|$, with the frequencies $f_{ii}$ and the amounts $g_{ii}$ being highly correlated with the number of addresses identified in the preceding Section 2.1 as naturally expected. Because our main interest in this paper is the crypto flow from one user to another, we remove all the self-loops in what follows.

Adjacent matrices with the strength of links given by the frequencies $f_{ij}$ of $G_t$ for all $t$’s are illustrated in Appendix B. One can see that the overall picture does not change in time, but the illustration does not help to uncover the nature of connectivity and flow.

To see the connectivity of network, namely how those regular users are linked among them and also how they are located in the stream of crypto flow, let us examine the property of connected com-
components. First, decompose $G_t$ into weakly connected components (WCC), i.e. connected components when regarded as an undirected graph. We found that there exists a giant WCC (GWCC) containing most of the users. See the column GWCC of Table I. There was only a small number of disconnected components as shown in the same column.

Then, in order to identify the location of users contained in the GWCC, we employed the well-known analysis of “bow-tie” structure [23]. In general, GWCC can be decomposed into the following parts:

**GSCC** Giant strongly connected component: the largest connected component when viewed as a directed graph. One or more directed paths exist for an arbitrary pair of firms in the component.

**IN** The nodes from which the GSCC is reached via at least one directed path.

**OUT** The nodes that are reachable from the GSCC via at least one directed path.

**TE** “Tendrils”; the rest of the GWCC.

It follows that

$$GWCC = GSCC + IN + OUT + TE \quad (3)$$

GSCC is the core of the crypto flow’s circulation. The IN and OUT parts are upstream and downstream of the flow respectively. The users in the part of IN are playing a role of suppliers of crypto, while the OUT users are considered to be consumers of crypto.

Table I shows the bow-tie structure in the column of GSCC/IN/OUT/TE. For example, in September, 470 users are located into GSCC (325 users), IN (23), OUT (113), and TE (5). One can observe that a large fraction of the users in the GWCC is located in the GSCC, as one can easily interpret this fact in the way that those regular users are circulating crypto globally. There are a less fraction of the users in IN and OUT with asymmetry in the numbers.

It would be interesting to see how the individual users are located in the temporal change of the network. Fig. 6 depicts such a diagram of temporal change from one month to its successive one in the whole year. One can see that the groups of GSCC, IN, and OUT are very stable in each membership of users. This fact means that those users appearing in successive months are playing stable roles in the crypto flow’s circulation and the location of upstream and downstream. We remark that analysis...
of bow-tie structure is based on the binary links, namely either presence or absence among the nodes, but not on the strength of links such as frequency and amount of crypto flow. In the next section, we shall see how to quantify the location of users by using the so-called Hodge decomposition.

3.2 Hodge Decomposition

Helmholtz-Hodge-Kodaira decomposition, or simply Hodge decomposition, is a combinatorial method to decompose flow on a network into circulation and gradient flow. Original idea dates back to the Helmholtz theorem in vector analysis, which states that under appropriate conditions any vector field can be uniquely represented by the sum of an irrotational or rotation-free (curl-free) vector field and a divergence-free (solenoidal) vector field. The theorem can be generalized from Euclidean space to graph and other entity as shown by Hodge, Kodaira and others. See [24–26] for readable exposition.

We recapitulate the method briefly for the present manuscript to be self-contained. Let $A_{ij}$ denote the adjacency matrix:

$$A_{ij} = \begin{cases} 
1 & \text{if there is a link of transfer from user } i \text{ to } j, \\
0 & \text{otherwise.}
\end{cases} \quad (4)$$

We excluded all the self-loops, implying that $A_{ii} = 0$. Each link has a flow, denoted by $\tilde{F}_{ij}$, either of the frequency, $f_{ij}$, or the amount, $g_{ij}$, of the transfer from $i$ to $j$ (see Fig. 3). Define

$$\tilde{F}_{ij} = \begin{cases} 
 f_{ij} \text{ or } g_{ij} & \text{if } A_{ij} = 1, \\
0 & \text{otherwise.}
\end{cases} \quad (5)$$

Note that there can be a pair of users such that $A_{ij} = A_{ji} = 1$ and $\tilde{F}_{ij}, \tilde{F}_{ji} > 0$.

Let us define a “net flow” $F_{ij}$ by

$$F_{ij} = \tilde{F}_{ij} - \tilde{F}_{ji} \quad (6)$$

and a “net weight” $w_{ij}$ by

$$w_{ij} = A_{ij} + A_{ji} \quad (7)$$

Note that $w_{ij}$ is symmetric, i.e., $w_{ij} = w_{ji}$, and non-negative, i.e., $w_{ij} \geq 0$ for any pair of $i$ and $j$.

Hodge decomposition is given by

$$F_{ij} = F_{ij}^{(c)} + F_{ij}^{(g)} \quad (8)$$

where the circular flow $F_{ij}^{(c)}$ satisfies

$$\sum_j F_{ij}^{(c)} = 0 \quad (9)$$

which implies that the circular flow is divergence-free. The gradient flow $F_{ij}^{(g)}$ can be expressed as

$$F_{ij}^{(g)} = w_{ij} (\phi_i - \phi_j) \quad (10)$$

Thus the weight $w_{ij}$ serves to make the gradient flow possible only where a link exists. We refer to the quantity $\phi_i$ as the Hodge potential. Large value of $\phi_i$ implies that the user $i$ is in the upstream of the entire network, while small values implies $i$ is in the downstream.

---

2It is remarked that (7) is simply a convention to consider the effect of mutual links between $i$ and $j$. One could multiply (7) by 0.5 or an arbitrary positive number, which does not change the result significantly for a large network.
Combine (8), (9), and (10), one can derive the following equation to determine $\phi_i$.

$$
\sum_j L_{ij} \phi_j = \sum_j F_{ij},
$$

for $i = 1, \ldots, N$. Here, $L_{ij}$ is the so-called graph Laplacian and defined by

$$
L_{ij} = \delta_{ij} \sum_k w_{ik} - w_{ij},
$$

where $\delta_{ij}$ is the Kronecker delta.

It is easy to show that the matrix $L = (L_{ij})$ has only one zero mode (eigenvector with zero eigenvalue). The presence of this zero mode simply corresponds to the arbitrariness in the origin of $\phi$. All the other eigenvalues are positive (see, e.g., [30]). Therefore, (11) can be solved for the potentials by fixing the potentials’ origin. We assume that the average value of $\phi$ is zero.

Fig. 7 depicts the distributions for Hodge potentials of the users in GSCC, IN, and OUT. One can see that the entire set of distributions is bimodal having two peaks at positive and negative values, while there are a number of values around zero. Obviously, they correspond to IN, OUT, and GSCC, each being located in the upstream, downstream, and core of the entire crypto flow. Moreover, there exists a correlation between the value of the Hodge potential and the net amount of demand or supply of crypto by each user. See [30] for details, where we studied a daily snapshot of the network including all the users, not only big players. We claim that the same property holds also for the monthly data restricted to big players of regular users.

### 3.3 Non-negative Matrix Factorization

It would be a natural question whether there are distinctive ingredients of flows in the crypto flow or not. The analysis of bow-tie structure is based merely on the binary relationship of links, so does not give such information, because the crypto flows from upstream to downstream with circulation in the giant strongly connected component that occupies a large fraction of the entire network. In other
words, are there any “principal components” that constitute the entire flow in a decomposition? In order to find such principal components or latent factors in the transfer of crypto among big players, we shall apply non-negative matrix factorization (NMF) to the strength of links, namely the matrix of the frequencies and amounts of transfer. We recapitulate the method here. See [31–33] and references therein for introduction.

Let $X$ be an $N \times M$ non-negative matrix, in general, to start with; that is, its elements are all non-negative, denoted as $X \geq 0$. NMF gives an approximation of $X$ by a product of two matrices:

$$X \approx SD,$$ (13)

where $S$, $D$ are $N \times K$ and $K \times M$ non-negative matrices, $S, D \geq 0$, respectively\(^3\). In practice, one expects that $K$ is much smaller than $N$ and $M$ so that the factorization gives a compact representation of $X$. We shall assume that $N = M$ for our application of crypto flow among $N$ users in what follows.

Explicitly in components, (13) reads

$$X_{sd} \approx \sum_{k=1}^{K} S_{sk} D_{kd},$$ (14)

where the indices $s$ and $d$ represent source and destination ($s, d = 1, \ldots, N$) respectively, and $X_{sd}$ is the strength of crypto flow, quantified by frequency $f_{sd}$, amount $g_{sd}$, or similar variables, from $s$ to $d$ in a certain period of time. We choose $X_{sd} = f_{sd}$, (15)

in this paper. See Fig. B·1 in Appendix B for the illustration of $X_{sd}$. We would expect that $K \ll N$, because of the sparsity of $X$. How to determine $K$ is discussed later.

The approximation in (13) is actually given by the following optimization:

$$\min_{S, D \geq 0} F(X, SD),$$ (16)

where the function $F(\cdot, \cdot)$ is the so-called Kullback-Leibler (KL) divergence defined by

$$F(A, B) = KL(A \parallel B) \equiv \sum_{i,j} \left( A_{ij} \log \frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij} \right).$$ (17)

Note that $F(A, B) = 0$ if and only if $A = B$. The reason why we choose the particular function of (17) will be clarified later\(^4\). Technically, one can solve (16) iteratively with the initialization of $S, D$ using non-negative double singular value decomposition (see the review [32] and references therein). Although the iterative algorithm yields local minima, our numerical solutions under different random seeds gave essentially the same decomposition.

To understand the meaning of the decomposition, let us consider how a source distributes flow to different destinations. For an arbitrary source $s$, (14) can be written as

$$X_s \approx \sum_{k=1}^{K} S_{sk} D_k,$$ (18)

where $X_s$ is the vector of $s$-th row of $X$, and $D_k$ is the vector of $k$-th row of $D$. Equation (18) means that the flow from the source $s$ can be expanded in terms of “basis” vectors, $D_k (k = 1, \ldots, K)$. The

---

\(^3\)The decomposition is not unique due to trivial degrees of freedom. One is permutation, $SD = S \pi D$, where $\pi$ is a permutation matrix simply exchanging indices. Another is scale transformation, $SD = \sigma S \pi D$, where $\sigma$ is a diagonal matrix with all elements positive. We shall see that these degrees are fixed after appropriate normalization and ordering.

\(^4\)See Section 3.4 for a probabilistic interpretation of choosing the KL divergence. Another functional form, often used, is the so-called Frobenius norm: $F(A, B) = (1/2) \sum_{i,j} (A_{ij} - B_{ij})^2$, which leads to a different probabilistic model.
components \((D_k)_d = D_{kd}\) represent how destinations are distributed among users in the \(k\)-th NMF component. It is convenient to normalize \(D_k\) by L1-norm, that is, by defining

\[
\tilde{D}_{kd} = \frac{D_{kd}}{D_k} \quad \text{where} \quad D_k \equiv \sum_d D_{kd} ,
\]

so that one has

\[
\sum_d \tilde{D}_{kd} = 1 ,
\]

for all \(k\). With respect to this normalized basis vectors, the expansion in (18) is rewritten as

\[
X_s \approx K \sum_{k=1}^{K} (S_{sk} D_k) \tilde{D}_{k} \quad \text{where} \quad (\tilde{D}_k)_d = \tilde{D}_{kd} ,
\]

Thus the outgoing flow from the source \(s\) is approximately expressed by a linear combination of \(K\) normalized basis vectors \(\tilde{D}_k\) with coefficients given by \(S_{sk} D_k\).

Similarly, consider how a destination \(d\) collects flow from different sources. For an arbitrary destination \(d\), (14) reads

\[
X_d \approx \sum_{k=1}^{K} D_{kd} S_k ,
\]

where \(X_d\) is the vector of \(d\)-th column of \(X\), and \(S_k\) is the vector of \(k\)-th column of \(S\). The components of \((S_k)_s = S_{sk}\) represent how sources are distributed among users in this \(k\)-th NMF component. Define

\[
\bar{S}_{sk} \equiv \frac{S_{sk}}{S_k} \quad \text{where} \quad S_k \equiv \sum_s S_{sk} ,
\]

and one has

\[
\sum_s \bar{S}_{sk} = 1 ,
\]

for all \(k\). Then (22) is rewritten as

\[
X_d \approx \sum_{k=1}^{K} (D_{kd} S_k) \bar{S}_k \quad \text{where} \quad (\bar{S}_k)_s = \bar{S}_{sk} .
\]

Thus the incoming flow to the destination \(d\) is approximately expressed by a linear combination of \(K\) normalized basis vectors \(\bar{S}_k\) with coefficients given by \(D_{kd} S_k\).

How can one determine \(K\)? Obviously, the larger \(K\) is, the better the approximation (13) is, but with less parsimonious representation of the data. In the next section, let us make a detour to examine this issue from a different perspective.

### 3.4 NMF as a probabilistic model

We can interpret the NMF as a probabilistic model. Denote the right-hand of (14) by

\[
\xi_{sd} \equiv \sum_k S_{sk} D_{kd} ,
\]

which are regarded as parameters to be estimated from the data \(X\) such that \(X_{sd}\) is assumed to be a random number chosen from a Poisson distribution with the parameter \(\xi_{sd}\) as

\[
P(x | \xi) = e^{-\xi} \frac{\xi^x}{x!} .
\]
It is easy to see that the log likelihood function \( L(\xi_{sd}) \equiv \log P(X_{sd} | \xi_{sd}) \) takes the maximum value at \( \xi_{sd} = X_{sd} \). Then one can introduce a quantity to measure how much the estimation of the parameters is good, that is
\[
\sum_{s,d} (L(X_{sd}) - L(\xi_{sd})) = \sum_{s,d} \left( X_{sd} \log \frac{X_{sd}}{\xi_{sd}} - X_{sd} + \xi_{sd} \right),
\]
(28)
to be minimized. One can see that this quantity is equivalent to the KL divergence in (17)\(^5\).

To express the entire framework in probabilistic terms more explicitly, let us normalize the data \( X \) in (15) by
\[
\tilde{X}_{sd} = \frac{X_{sd}}{\sum_{s',d'} X_{s'd'}}.
\]
(29)
Then let us rewrite (14) as
\[
\tilde{X}_{sd} \approx \sum_k r_k \tilde{S}_{sk} \tilde{D}_{kd},
\]
(30)
where \( \tilde{D}_{kd} \) and \( \tilde{S}_{sk} \) were given by (19) and (23) respectively, and
\[
 r_k \equiv \frac{S_k D_k}{\sum_{k'} S_{k'} D_{k'}},
\]
(31)
which satisfies that \( \sum_k r_k = 1 \). Let us denote the right-hand side of (30) by
\[
p_{sd} \equiv \sum_k r_k \tilde{S}_{sk} \tilde{D}_{kd},
\]
(32)
which satisfies that \( \sum_{s,d} p_{sd} = 1 \). We remark that the normalized weight \( r_k \) defined by (31) gives the information of relative importance of the \( k \)-th NMF component in the expansion with normalized basis vectors in (32). One can determine the ordering of NMF components uniquely according to the magnitudes of \( r_k \).

Suppose that there are \( N_f \) transfers in total during a period of time. For each pair of source and destination, \( s \) and \( d \), generate a transfer \( s \rightarrow d \) with the probability given by \( p_{sd} \), being independently of other pairs. Under the assumption of a small probability of \( p_{sd} \) and a large number of \( N_f \), \( X_{sd} \) follows a Poisson distribution with the parameter, \( \xi_{sd} = N_f p_{sd} \).

It turns out that the decomposition in (26), or equivalently (32), has an interesting connection with machine learning. In natural language processing, it is often necessary to extract topics among documents comprising of words or terms. In a situation of unsupervised learning, the task is to infer topics as hidden or latent variables, which can explain a collection of documents, each being an unordered set of terms. Probabilistic latent semantic analysis (PLSA) is a probabilistic model for doing such a task [36]. Suppose that there are \( N \) documents and \( M \) terms. Then the occurrence of terms can be expressed by a document-term matrix \( X \) with size \( N \times M \), each element of which is the frequency of occurrence of a term in a document. Topics are latent variables to explain the data \( X \). A topic is actually a probability distribution for the occurrence of terms with different probabilities. A document can have a mixture of topics. An example is a document on “influence of hosting Olympics to economy” with a mixture of topics on sports and economy.

One of the widely used model of PLSA is latent Dirichlet allocation (LDA). See Appendix C and references therein. For our purpose, it suffices to understand how terms are generated at locations of documents in a probabilistic way. The probability that a term is chosen at a location in a document is given by the sum of \( K \) factors, each of which is the product of two probabilities; the probability that

\(^5\)We became aware that this argument is known in the literature. If one assumes Gaussian instead of Poisson, one would have Frobenius norm for KL divergence in (17). See [34]. One of the present authors (YF) learned a hint on the argument from Itsuki Noda in his application of NMF to transportation data [35].
Fig. 8. Determining $K$, the number of NMF components, by using the concept of coherence in LDA. (a) Different methods [38–40] are compared. Each measure of coherence is drawn in the vertical axis so that it is to be minimized to find the optimal number of components. Maximum and minimum values in the region of $K$ are scaled to 1 and 0 respectively to make the comparison easier. One can see that $K = 11 \sim 13$ are optimal. (b) Monte Carlo simulations by the method [39] with 20 runs for each $K$. Averages (points) and 99% level (gray band, narrow) calculated from standard errors are drawn. We conclude that $K = 13$ is optimal from this result (b).

a topic is selected in the document and the one that the term is chosen under the selected topic. See the equation of (C·11) in Appendix C. One can immediately see that (C·11) is essentially the same as (26), or equivalently (32).

Thus the matrix decomposition of NMF can be put in the framework of probabilistic model of PLSA and LDA. As a bonus, one can adopt the method of estimating the number of topics to our problem of determining the number of NMF components, denoted by $K$ in both cases. Interested readers are guided to look at the literature [37–40] and others given at the end of Appendix C. Let us take a look at our results in the next section finishing the detour of this section.

3.5 Result of NMF for Crypto Flow

We first show a few results for a snapshot of September in the year 2019 (denoted as 2019-09), in order to verify if the idea in the preceding section works to determine the number of NMF components. Fig. 8 shows the measures of coherence by employing three different methods of LDA [38–40]. The methods give mostly the same optimal values of $K$, namely $K = 11 \sim 13$, consistently as shown in Fig. 8 (a). We found that the measure given in [39] is relatively stable and potentially useful to determine a specific value of $K$. So we performed Monte Carlo simulations in Fig. 8 (b), and were able to determine the optimal value as $K = 13$. For this data, $X$ has the dimension of $N = 470$, so we conclude that one can have a small number of NMF components that can explain the entire flow among those regular users.

In Appendix D, we summarize the same result for the data in all the other months of the year 2019. We found that the optimal number $K$ is quite small in the range more than 10 and less than 20, much smaller than the number of users, $N \sim 500$ (see Table I). Additionally, $K$ is relatively stable irrespectively of the temporal change. See Table D·1 and Fig. D·1.

Let us examine each NMF components obtained with the optimal value of $K$. Fig. 9 and Fig. 10 show the NMF components in terms of the basis vectors, $\tilde{D}_k$ and $\tilde{S}_k$, respectively for $k = 1, \ldots, K$. In Fig. 9, each plot shows the vector components of $(\tilde{D}_k)_d = D_{kd}$ meaning how destinations are distributed among users $d$ in the $k$-th NMF component. Similarly in Fig. 10, each plot shows the
vector components of \( \tilde{S}_k \) means how sources are distributed among users \( s \) in the \( k \)-th NMF component. See (19) and (23), and also note the normalization therein. Note that in each of Fig. 9 and Fig. 10, the plots are ordered (from top to bottom) in the descending order of the probability \( r_k \) given in (31).

One can immediately notice from the figures that the components of these basis vectors are concentrated on a limited number of users, but are not distributed among many users. To quantify the effective number of the concentration, let us use the inverse Herfindahl-Hirschman index, abbreviated as IHH, which is defined as follows. Consider “shares” \( x_i \geq 0 \) among \( i = 1, \ldots, N \) things with the sum equal to 1, i.e. \( \sum_i x_i = 1 \). The IHH is defined by

\[
\text{IHH} \equiv \left( \sum_{i=1}^{N} x_i^2 \right)^{-1}.
\]

When the shares are equal, \( x_i = 1/N \) for all \( i \), then \( \text{IHH} = N \). On the other hand, when there is the strongest concentration, namely, \( x_i = 1 \) for a particular \( i \) and \( x_i = 0 \) otherwise, then \( \text{IHH} = 1 \). So IHH can give an estimate of the effective number of large shares\(^6\). The idea can be applied to the basis vectors, because the vectors are normalized in the same way as shares. In Fig. 9 and Fig. 10, we displayed all the calculated IHH’s. One can see that the IHH’s are quite small ranging from a few to a dozen or so, compared with the total number of users \( N = 470 \) for the data 2019-09 (see Table I).

How can we use the NMF components to understand the crypto flow? Choose a particular user \( s \) as a source. The flow from \( s \) was approximately expressed by a linear combination of \( K \) normalized basis vectors \( \tilde{D}_k \), each depicted in Fig. 9, with the coefficients given in (21), i.e. \( S_{sk} \tilde{D}_k \). The coefficients represent the strength of the decomposed flow from the source \( s \). A similar argument holds by choosing a particular user \( d \) as a destination. The flow to \( d \) was expressed by a linear combination in (25) with the coefficients, \( D_{kd} S_k \).

For example, consider the user with ID \( 0000000000 \) that is located in the GSCC, as a source \( s \). Fig. 11 shows the coefficients corresponding to \( K \) components; see (a). One can see that the coefficients are non-zero at only four components. Such a sparseness tells that the flow from this user can be expressed with a few components. And the corresponding components have non-zero values at a small number (recall the IHH’s) of the vector components, \( D_{kd} \), as shown in Fig. 9, implying that the users corresponding to these non-zero components constitute a cluster for the outgoing flow from the source \( s \).

The same user \( 0000000000 \) can be regarded as a destination \( d \) in the GSCC. Fig. 11 (b) shows the coefficients, again non-zero at only one or two components. Together with Fig. 10, one can find another cluster composed of a small number of users for the incoming flow to the \( d \). Similar arguments hold for the users \( 0000006178 \) and \( 0000000012 \), respectively located in the IN and OUT. See Fig. 11 (c) and (d). In this way, one can find clusters for either of or both of the outgoing and incoming flows of each user.

Each NMF component can be represented by a matrix, because the non-negative matrix of \( X_{sd} \) or its probabilistic counterpart \( P_{sd} \) can be expressed by (18) or (32). It is possible to depict each component \( k \) by the matrix of \( S_{sk} \tilde{D}_{kd} \) in the normalized way. Fig. 12 and Fig. 13 illustrate such matrices for the data of 2019-09. This should be compared with \( X_{sd} \) in Appendix B. One can see that the NMF components provide sparse matrices.

Finally, we find that the NMF components are relatively stable in the temporal change of network. Fig. 14 (a) shows the result for cosine similarities of the NMF basis vectors \( D_i \) for the two successive months of 2019-09 and 2019-10. Fig. 14 (b) shows the result for \( S_k \) for the same data. In either of these results, one can see that the NMF components are quite similar except only a few permutation of indices.

\(^6\)It is remarked that the paper [20] in this volume also applied but modifies Herfindahl-Hirschman index in an interesting way to characterize the frequencies of transactions in XRP.
Fig. 9. NMF components $\tilde{D}_k (k = 1, \ldots, 13)$ from top to bottom. Each plot $(\tilde{D}_k)_d = D_{kd}$ shows how destinations are distributed among users $d$ in the $k$-th NMF component. Note that $\tilde{D}_k$ is normalized, i.e. $\sum_d D_{kd} = 1$ for all $k$. See (19) and (20) in the main text. Also shown in each plot are the probability of the component, denoted by “Prob”, and the inverse Herfindahl-Hirschman index “IHH” representing the effective number of dominant users in the component. Data: 2019-09.
Fig. 10. NMF components $\tilde{S}_k$ ($k = 1, \ldots, 13$) from top to bottom. Each plot $(\tilde{S}_k)_s = S_{sk}$ shows how sources are distributed among users $s$ in the $k$-th NMF component. Note that $\tilde{S}_k$ is normalized, i.e. $\sum_s S_{sk} = 1$ for all $k$. See (23) and (24) in the main text. See the caption of Fig. 9 for “Prob” and “IHH” in each plot. Data: 2019-09.
Fig. 11. Coefficients, with which crypto flow from or to a selected user is expanded with respect to the NMF components. The selected users are (a,b) included in the GSCC of bow-tie structure, (c) in the IN, and (d) in the OUT. The user can be source (a) and destination (b). The user of (c) is a source, and the user of (d) is a destination. The expansion is given by (21) for the selected source $s$, and by (25) for the selected destination $d$. Data: 2019-09.
Fig. 12. NMF components $k = 1, \ldots, 8$ as matrices defined by $\tilde{S}_{ik} \tilde{D}_{kd}$ in (32). Data: 2019-09.
Fig. 13. (Continued) NMF components $k = 9, \ldots, 13$ as matrices defined by $\tilde{S}_{ak}\tilde{D}_{kd}$ in (32). Data: 2019-09.
Fig. 14. Temporal change of the NMF components from one month \( t \) to its successive month \( t+1 \). (a) Cosine similarities calculated for all the pairs of \( D_k \) \((k = 1, \ldots, 13)\) between \( t \) (vertical) and \( t+1 \). (b) The same for all the pairs of \( S_k \). For the vertical and horizontal axes in (a) and (b), the order of indices along vertical and horizontal axes corresponds to the descending order of the probability \( r_k \) in (31). Data: 2019-09 and 2019-10.
4. Discussions

Let us briefly discuss about several aspects that would be worth further investigation.

First, while we succeeded to extract the NMF components and found that the components have non-zero values only at a relatively small number of users, we still did not identify those users by exploiting the fact. It is quite likely the case that the extracted users must play important roles in each of the NMF components, either of key destinations or key sources. We attempted to identify a tiny fraction of such users by matching the list of such users with the identity given in Appendix A, but the identification was not sufficient in order to interpret the meaning of corresponding NMF components. Instead, such intra-day activities as shown in Fig. 2 of Section 2.1 could give us the geographical locations of those key users, possibly uncovering the crypto flow in each NMF component at a global scale. This issue remains to be investigated.

Second, even if the temporal change of the network in terms of the NMF components has such a stable structure as found in Fig. 14, we noticed that there exists interesting change of a few components in the same figure. A keen reader may have noticed that the components \( k = 3, 4, 5, 6 \) at time \( t \) are changed into among themselves at time \( t + 1 \), while the cosine similarities are close to 1. This means that the probabilities \( r_k \) for those components were changed from one month to the next. Also one can notice that the optimal number of NMF components showed a slow variation during the period (recall Table D-1 of Appendix D). These facts might give us a hint for how to treat the temporal change of network by paying attention to those slowing varying aspects. Additionally, while we focused only on regular users appearing everyday during the period under study, it would be necessary to include the process of entry and exit of big players.

Third, technically, we regarded the method of NMF as a probabilistic model that shares the same stochastic process as in the probabilistic latent semantic analysis (PLSA). As a bonus, we were able to employ the latent Dirichlet allocation (LDA) and its known methods to estimate the number of topics in the context of topic model, or the number of NMF components in our context. In principle, one could start with the full-fledged Bayesian framework in the LDA and its extension and variations. It would be worth pursuing in this direction, which is also related to the second point above, because there are several studies on how to treat temporally changing topics of documents in a long time-span.

Fourth, our methods in this paper can be easily applied to different cryptoassets including Ethereum and XRP. We are aware of the paper [20] in this volume, which is in a similar line of study. It would be interesting to apply our methods to the data of XRP.

Finally, it would be an extremely interesting problem how the crypto flow among big players is related to the prices in the exchange markets with fiat currencies and also with other cryptoassets. It is quite likely that the bubble/crash and their precursors might force the big players to react during such turmoils in a different way from tranquil periods. For example, exchanges need to reallocate cryptoassets in the necessity of making a reservoir or doing a release of cryptoassets under the risk.

5. Summary

Our purpose in this study on the cryptoasset of Bitcoin is to understand the structure and temporal change of crypto flow among big players. We compiled all the transactions contained in the blockchain of Bitcoin cryptoasset from its genesis to the year 2020, identified users from anonymous addresses, and constructed snapshots of networks comprising of users as nodes and links as crypto flow among the users. While the whole network is huge, we extracted sub-networks by focusing on regular users who appeared persistently during a certain period. Specifically, we extracted monthly snapshots during the year of 2019, and selected roughly 500 regular users.

We first analyzed the bow-tie structure from the binary relationship of flow, and then performed the Hodge decomposition based on the strength of flow defined by frequencies and amounts, in order
to locate users in the upstream, downstream, and core of the entire crypto flow. We found that the bow-tie structure is stable during the period, implying that those regular users have different roles in the crypto flow.

Then, to reveal important ingredients hidden in the flow, we employ the method of non-negative matrix factorization (NMF) to extract a set of principal components. We discussed that the NMF method can be regarded as a probabilistic model, which is equivalent to a probabilistic latent semantic analysis and its typical model of latent Dirichlet allocation. This observation brought us a method to estimate an optimal number of NMF components, which turned out to be a dozen or so. We found that the NMF components have non-zero values corresponding to a limited number of users, telling us their roles of destinations or sources of the crypto flow. Additionally, we found that the NMF components are quite stable in the temporal change for the time-scale of months.

There remain several points including the further investigation on the users contained in those NMF components, a treatment of temporally changing network, and technically interesting issues to be pursued in the future.

Acknowledgment

We would like to thank Hideaki Aoyama, Yuichi Ikeda, and Hiwon Yoon for discussions, Hiroshi Iyetomi for solving a technical issue of Hodge decomposition, Itsuki Noda for clarifying his application of NMF on transportation data, Takeaki Uno for an efficient algorithm to identify users, Shinya Kawata and Wajun Kawai for technical assistance. This work is supported by JSPS KAKENHI Grant Numbers, 19K22032 and 20H02391, by the Nomura Foundation (Grants for Social Science), and also by Kyoto University and Ripple’s collaboration scheme.

References

[1] F. Reid and M. Harrigan. An analysis of anonymity in the bitcoin system. In Y. Altshuler, Y. Elovici, A. Cremers, N. Aharony, and A. Pentland, editors, Security and Privacy in Social Networks, pages 197–223. Springer, New York, 2013.
[2] M. Ober, S. Katzenbeisser, and K. Hamacher. Structure and anonymity of the bitcoin transaction graph. Future Internet, 5(2):237–250, 2013.
[3] D. Ron and A. Shamir. Quantitative analysis of the full bitcoin transaction graph. In Financial Cryptography and Data Security, 2013.
[4] D. Kondor, M. Pósfai, I. Csabai, and G. Vattay. Do the rich get richer? an empirical analysis of the bitcoin transaction network. PLoS ONE, 9(2):e86197, 2014.
[5] D. Kondor, I. Csabai, S. J., M. Pósfai, and G. Vattay. Inferring the interplay between network structure and market effects in bitcoin. New Journal of Physics, 16(12):125003, 2014.
[6] B. Alvarez-Pereira, M. A. Ayres, A. M. G. López, S. Gorský, S. W. Hayes, Z. Qiao, and J. Santana. Network and conversation analyses of bitcoin. In 2014 Complex Systems Summer School Proceedings, 2014.
[7] A. Baumann, B. Fabian, and M. Lischke. Exploring the bitcoin network. In Proceedings of the 10th International Conference on Web Information Systems and Technologies – Volume 2: WEBIST, 2014.
[8] M. Fleder, M. S. Kester, and S. Pillai. Bitcoin transaction graph analysis. http://arxiv.org/abs/1502.01657, 2015.
[9] D. Maesa, A. Marino, and L. Ricci. Uncovering the bitcoin blockchain: An analysis of the full users graph. In Conference: 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2016.
[10] M. Lischke and B. Fabian. Analyzing the bitcoin network: The first four years. Future Internet, 8(1):7, 2016.
[11] C. G. Akcora, Y. R. Gel, and M. Kantarcioğlu. Blockchain: A graph primer. http://arxiv.org/abs/1708.08749, 2017.
[12] M. Bartoletti, A. Bracciali, S. Lande, and L. Pompeianu. A general framework for bitcoin analytics. http://arxiv.org/abs/1707.01021, 2017.
[13] C. Cachin, A. D. Caro, P. Moreno-Sanchez, B. Tackmann, and M. Vukolić. The transaction graph for modeling blockchain semantics. https://eprint.iacr.org/2017/1070, 2017.

[14] R. Cazabet, B. Rym, and M. Latapy. Tracking bitcoin users activity using community detection on a network of weak signals. http://arxiv.org/abs/1710.08158, 2017.

[15] D. Maesa, A. Marino, and L. Ricci. Data-driven analysis of bitcoin properties: exploiting the users graph. International Journal of Data Science and Analytics, 2017.

[16] S. Ranshous, C. Joslyn, S. Kreyling, K. Nowak, N. F. Samatova, C. L. West, and S. Winters. Exchange pattern mining in the bitcoin transaction directed hypergraph. In Financial Cryptography Workshops, 2017.

[17] A. M. Antonopoulos. Mastering Bitcoin: Programming the Open Blockchain. O'Reilly Media, 2 edition, 2017.

[18] Hungary research group: Elte bitcoin project. http://www.vo.elte.hu/bitcoin/; https://senseable2015-6.mit.edu/bitcoin/. Accessed: 2020-01-03.

[19] Bitcoin block explorer with address grouping and wallet labeling. https://www.walletexplorer.com. Accessed: 2021-03-18.

[20] H. Aoyama. Xrp network and proposal of flow index; see it in this volume., 2021.

[21] R. Islam, Y. Fujiwara, S. Kawata, and H. Yoon. Analyzing outliers activity from the time-series transaction pattern of bitcoin blockchain. Evolutionary and Institutional Economics Review, 16(1):239–257, 2019.

[22] R. Islam, Y. Fujiwara, S. Kawata, and H. Yoon. Unfolding identity of financial institutions in bitcoin blockchain by weekly pattern of network flows. in preparation, 2020.

[23] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. Computer Networks, 33(1-6):309–320, 2000.

[24] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Learning to rank with combinatorial hodge theory. available online, 2008. accessed in January, 2020.

[25] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial hodge theory. Mathematical Programming, 127(1):203–244, 2011.

[26] J. L. Johnson. Discrete hodge theory on graphs: a tutorial. Computing in Science & Engineering, 15(5):42–55, 2013.

[27] K. Miura and T. Aoki. Scaling of hodge-kodaira decomposition distinguishes learning rules of neural networks. IFAC-PapersOnLine, 48(18):175–180, 2015.

[28] Y. Kichikawa, H. Iyetomi, T. Iino, and H. Inoue. Hierarchical and circular flow structure of interfirm transaction networks in japan. https://ssrn.com/abstract=3173955, 2018. May 4, 2018.

[29] Y. Fujiwara, H. Inoue, T. Yamaguchi, H. Aoyama, T. Tanaka, and K. Kikuchi. Money flow network among firms’ accounts in a regional bank of japan. EPJ Data Science, 10:19, 2021.

[30] F. Y. and I. R. Advanced Studies of Financial Technologies and Cryptocurrency Markets, chapter Hodge Decomposition of Bitcoin Money Flow. Springer, Singapore, 2020.

[31] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791, 1999.

[32] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis, 52(1):155–173, 2007.

[33] R. Gaujoux and C. Seoighe. A flexible r package for nonnegative matrix factorization. BMC Bioinformatics, 11:367, 2010.

[34] H. Kameoka. Non-negative matrix factorization and its variants with applications to audio signal processing. Journal of the Japan Statistical Society, 44(2):383–407, 2015. Japanese.

[35] I. Noda and J. Ochiai. Analysis of od data of ondemand transportation by non-negative matrix factorization. 2019. Technical report in the workshop of SIG-DOCMAS, Japanese.

[36] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 42:177–196, 2001.

[37] T. L. Griffiths and M. Steyvers. Finding scientific topics. Proceedings of the National Academy of Sciences, 101:5228–5235, 2004.

[38] J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang. A density-based method for adaptive lda model selection. Neurocomputing, 72(7):1775–1781, 2009.
[39] R. Arun, V. Suresh, C. E. Veni Madhavan, and M. N. Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In M. J. Zaki, J. X. Yu, B. Ravindran, and V. Pudi, editors, Advances in Knowledge Discovery and Data Mining, pages 391–402. Springer Berlin Heidelberg, 2010.

[40] R. Deveaud, E. SanJuan, and P. Bellot. Accurate and effective latent concept modeling for ad hoc information retrieval. Document Numérique, 17(1):61–84, 2014.

[41] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[42] D. M. Blei and J. D. Lafferty. Text Mining: Classification, Clustering, and Applications, chapter Topic Models, pages 71–94. CRC Press, 2009.

[43] Grün, Bettina and Hornik, Kurt. topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software, 40(13):1–30, 2011.

[44] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. Journal of the American Statistical Association, 101(476):1–30, 2006.
Appendix A: Identity of Users of Type A

WalletExplorer.com [19] is a web site providing information about identity of addresses in Bitcoin blockchain. The site merges addresses together, if they are part of the same wallet, and also identifies wallets with actual names. According to the site, the method to merge addresses is:

*Just a basic algorithm is used to determine wallet addresses. Addresses are merged together, if they are co-spent in one transaction. So if addresses A and B are co-spent in transaction T1, and addresses B and C are co-spent in transaction T2, all addresses A, B and C will be part of one wallet.*

*Sometimes, an address belongs to some service but it was never co-spent with others. Then that address stays unnamed. It is typically more often at addresses with higher amount (as there is no need to co-spending).*

This method is precisely the same as [1], which is the one we employed in Section 2.1. In addition, the identification of actual names is done by WalletExplorer.com as follows:

*In most of the cases, I registered to service, made transaction(s) and saw which wallet bitcoins were merged with, or from which wallet it was withdrawn.*

*There is probably no easier way how to discover names other than this.*

*Please note that the name database is not updated, so it does not contain newer exchanges (or newer wallets of existing exchanges).*

We matched our data with the one in [19] to obtain the identity and additional attributes of users of type A (see Section 2.1 for the type). Table A.1 is the classification into exchanges, services, gambling, historic, and mining pools. Table A.2 shows the list of countries that exchanges belong to. Table A.3 is the complete list of this matching.

**Table A.1.** Classification of identified users (compiled from [19])

| Classification | #Users | Examples |
|----------------|--------|----------|
| Exchanges      | 84     | Bittrex.com, Huobi.com, Bit-x.com, HitBtc.com |
| Old/Historic   | 83     | AgoraMarket, EvolutionMarket, SilkRoadMarketplace |
| Services/Others| 45     | Xapo.com, ePay.info, Cubits.com |
| Gambling       | 41     | 999Dice.com, CoinGaming.io, SatoshiMines.com |
| Pools          | 11     | BTCCPool, SlushPool.com, BitMinter.com |
| **Total**      | 264    | —        |

**Table A.2.** Countries of identified exchanges (compiled from [19])

| Country   | #Users | Country   | #Users | Country       | #Users |
|-----------|--------|-----------|--------|--------------|--------|
| China     | 14     | Mexico    | 2      | Iran         | 1      |
| UK        | 13     | Netherlands| 2     | Korea        | 1      |
| USA       | 13     | Poland    | 2      | Luxembourg   | 1      |
| Canada    | 4      | Portugal  | 2      | Malta        | 1      |
| Australia | 3      | South Africa| 2   | Panama       | 1      |
| Brazil    | 3      | Austria   | 1      | Taiwan       | 1      |
| Singapore | 3      | Belize    | 1      | Thailand     | 1      |
| Russia    | 3      | Croatia   | 1      | Vanuatu      | 1      |
| Denmark   | 2      | Czech     | 1      | Vietnam      | 1      |
| Finland   | 2      | Germany   | 1      |              |        |
| **Total** | 84     |           |        |              |        |
Table A.3: Identity of Users (compiled from [19])

| No. | User ID | #Addr. (1) | #Addr. (2) | Name | Category | Country |
|-----|---------|------------|------------|------|----------|---------|
| 1   | 18,913,420 | 1         | Bit-x.com  |       | Exchanges | South Africa |
| 2   | 18,913,420 | 1         | Xapo.com   |       | Services/Others | — |
| 3   | 13,110,033 | 1,246,250 | ePay.info  |       | Services/Others | — |
| 4   | 5,302,867  | 1         | Cryptopay.me |     | Services/Others | — |
| 5   | 5,302,867  | 1         | Cubits.com |       | Services/Others | — |
| 6   | 5,302,867  | 1         | Luno.com   |       | Exchanges | South Africa |
| 7   | 5,302,867  | 1         | VirWoX.com |       | Exchanges | Austria |
| 8   | 5,302,867  | 1         | Xapo.com   |       | Services/Others | — |
| 9   | 5,302,867  | 487,776   | CoinPayments.net |     | Services/Others | — |
| 10  | 2,180,236  | 1,488,034 | Xapo.com   |       | Services/Others | — |
| 11  | 1,630,483  | 1         | Huobi.com  |       | Exchanges | China |
| 12  | 1,630,483  | 957,652   | Cubits.com |       | Services/Others | — |
| 13  | 1,538,476  | 1,377,461 | Bittrex.com |       | Exchanges | USA |
| 14  | 1,413,904  | 1         | Bitstamp.net |     | Exchanges | Luxembourg |
| 15  | 1,043,379  | 279,697   | Huobi.com  |       | Exchanges | China |
| 16  | 992,726    | 1         | VirWoX.com |       | Exchanges | Austria |
| 17  | 992,726    | 1         | Xapo.com   |       | Services/Others | — |
| 18  | 988,800    | 940,605   | Poloniex.com |     | Exchanges | USA |
| 19  | 912,950    | 133,020   | AnxPro.com |       | Exchanges | China |
| 20  | 912,950    | 770,486   | CoinTrader.net |     | Exchanges | Canada |
| 21  | 845,559    | 1         | Cubits.com |       | Services/Others | — |
| 22  | 811,809    | 2         | Cubits.com |       | Services/Others | — |
| 23  | 778,990    | 651,547   | 999Dice.com |     | Gambling | — |
| 24  | 682,176    | 1         | MoonBit.co.in |     | Services/Others | — |
| 25  | 659,820    | 656,699   | CoinGaming.io |     | Gambling | — |
| 26  | 631,985    | 1         | Luno.com   |       | Exchanges | South Africa |
| 27  | 581,525    | 1         | Cubits.com |       | Services/Others | — |
| 28  | 546,515    | 291,443   | Luno.com   |       | Exchanges | South Africa |
| 29  | 523,330    | 522,056   | LocalBitcoins.com |     | Exchanges | Finland |
| 30  | 498,001    | 498,001   | AgoraMarket |     | Old/Historic | — |
| 31  | 478,476    | 343,039   | Bitstamp.net |     | Exchanges | Luxembourg |
| 32  | 420,632    | 420,632   | EvolutionMarket |     | Old/Historic | — |
| 33  | 412,338    | 249,883   | Cryptonator.com |     | Services/Others | — |
| 34  | 398,349    | 1         | Xapo.com   |       | Services/Others | — |
| 35  | 377,140    | 305,518   | Cryptopay.me |     | Services/Others | — |
| 36  | 379,518    | 1         | SilkRoadMarketplace |     | Old/Historic | — |
| 37  | 370,536    | 350,036   | SilkRoad2Market |     | Old/Historic | — |
| 38  | 341,160    | 61,103    | MercadoBitcoin.com.br |     | Exchanges | Brazil |
| 39  | 325,365    | 1         | Xapo.com   |       | Services/Others | — |
| 40  | 307,489    | 307,451   | BTC-e.com |       | Exchanges | Russia |
| 41  | 294,238    | 1         | Xapo.com   |       | Services/Others | — |
| 42  | 278,973    | 2         | Cubits.com |       | Services/Others | — |
| 43  | 269,605    | 254,601   | SatoshiMiners.com |     | Gambling | — |
| 44  | 263,074    | 262,940   | YoBit.net |       | Exchanges | Russia |
| 45  | 250,920    | 191,689   | Bitcoin.de |       | Exchanges | Germany |
| 46  | 241,250    | 223,781   | BitcoinFog |       | Services/Others | — |
| 47  | 238,480    | 238,476   | Cex.io     |       | Exchanges | UK |
| 48  | 206,542    | 156,203   | Coinbase.com |     | Services/Others | — |
| 49  | 197,189    | 197,155   | NitrogenSports.eu |     | Gambling | — |
| 50  | 189,776    | 189,776   | AlphaBayMarket |     | Services/Others | — |
| 51  | 187,189    | 187,189   | HitBTC.com |       | Exchanges | UK |
| 52  | 186,000    | 186,000   | BitPay.com |       | Services/Others | — |
| 53  | 168,885    | 1         | Luno.com   |       | Exchanges | South Africa |
| 54  | 146,381    | 146,381   | NucleusMarket |     | Services/Others | — |
| 55  | 145,978    | 1         | Cryptonator.com |     | Services/Others | — |
| 56  | 145,978    | 1         | Cubits.com |       | Services/Others | — |
| 57  | 140,594    | 1         | Bitcoin.de |       | Exchanges | Germany |

Continue to next page
| No. | User ID | #Addr. (1) | #Addr. (2) | Name | Category | Country |
|-----|---------|------------|------------|------|----------|---------|
| 58  | 000000169 | 140,594    | 1          | Cubits.com | Services/Others |         |
| 59  | 000000169 | 140,594    | 1          | Poloniex.com | Exchanges | USA     |
| 60  | 000000176 | 134,559    | 134,559    | Cryptsy.com | Exchanges | USA     |
| 61  | 000000182 | 131,979    | 1          | Poloniex.com | Exchanges | USA     |
| 62  | 000000190 | 125,004    | 125,004    | PocketDice.io | Gambling |         |
| 63  | 000000195 | 122,249    | 1          | Bitstamp.net | Exchanges | Luxembourg |
| 64  | 000000201 | 120,548    | 120,491    | FortuneJack.com | Gambling |         |
| 65  | 000000208 | 119,119    | 119,065    | AbraxasMarket | Old/Historic |         |
| 66  | 000000210 | 115,775    | 115,775    | CoinKite.com | Services/Others |         |
| 67  | 000000216 | 114,458    | 114,458    | Kraken.com | Exchanges | USA     |
| 68  | 000000217 | 109,798    | 1          | Luno.com | Exchanges | South Africa |
| 69  | 000000219 | 107,479    | 85,122     | Bleutrade.com | Exchanges | Brazil |
| 70  | 000000234 | 96,890     | 96,890     | SecondsTrade.com | Gambling |         |
| 71  | 000000239 | 92,226     | 72,581     | HolyTransaction.com | Services/Others |         |
| 72  | 000000244 | 90,473     | 44,416     | CoinSpot.com.au | Exchanges | Australia |
| 73  | 000000252 | 85,637     | 85,626     | MintPal.com | Old/Historic |         |
| 74  | 000000253 | 85,566     | 84,679     | Hashnest.com | Exchanges | China |
| 75  | 000000259 | 83,723     | 83,517     | BtcTrade.com | Exchanges | China |
| 76  | 000000265 | 82,695     | 27,343     | BtcLan.com | Services/Others |         |
| 77  | 000000267 | 80,987     | 73,965     | OKCoin.com | Exchanges | China |
| 78  | 000000269 | 80,086     | 1          | Poloniex.com | Exchanges | USA     |
| 79  | 000000270 | 79,712     | 79,712     | Bter.com | Exchanges | China |
| 80  | 000000274 | 78,849     | 78,849     | BitZino.com | Gambling |         |
| 81  | 000000278 | 78,119     | 78,048     | OKCoin.com | Exchanges | China |
| 82  | 000000282 | 76,997     | 1          | Cubits.com | Services/Others |         |
| 83  | 000000284 | 76,997     | 52,925     | BitOEX.com | Services/Others |         |
| 84  | 000000286 | 74,602     | 74,602     | Rollin.io | Gambling |         |
| 85  | 000000286 | 74,602     | 74,602     | Rollin.io | Gambling |         |
| 86  | 000000288 | 68,956     | 67,795     | CloudBet.com | Gambling |         |
| 88  | 000000302 | 68,087     | 55,098     | VirWoX.com | Exchanges | Austria |
| 89  | 000000307 | 67,114     | 1          | Cubits.com | Services/Others |         |
| 90  | 000000312 | 66,748     | 1          | Luno.com | Exchanges | South Africa |
| 91  | 000000318 | 65,367     | 1          | VirWoX.com | Exchanges | Austria |
| 92  | 000000325 | 64,803     | 64,803     | BTCC.com | Exchanges | China |
| 93  | 000000347 | 57,770     | 57,753     | MailCoin.com | Exchanges | Taiwan |
| 94  | 000000350 | 56,952     | 56,952     | BTCCPool.com | Pools |         |
| 95  | 000000358 | 55,757     | 55,757     | PandoraOpenMarket | Old/Historic |         |
| 96  | 000000359 | 55,703     | 1          | MercadoBitcoin.com.br | Exchanges | Brazil |
| 97  | 000000362 | 55,167     | 55,167     | Paxful.com | Exchanges | USA     |
| 98  | 000000366 | 54,640     | 54,640     | PrimeDice.com | Gambling |         |
| 99  | 000000379 | 53,639     | 53,639     | SheppMarketplace | Old/Historic |         |
| 100 | 000000374 | 53,102     | 53,102     | Caviertex.com | Exchanges | Canada |
| 101 | 000000387 | 50,878     | 50,878     | BlackBankMarket | Old/Historic |         |
| 102 | 000000402 | 48,602     | 48,518     | BX.in.th | Exchanges | Thailand |
| 103 | 000000403 | 48,525     | 19,434     | MoonBit.co.in | Services/Others |         |
| 104 | 000000407 | 48,178     | 33,772     | HaoBtc.com | Services/Others |         |
| 105 | 000000412 | 47,295     | 47,281     | Mathea.com | Exchanges | UK |
| 106 | 000000444 | 42,750     | 41,150     | SatoshiDice.com | Gambling |         |
| 107 | 000000447 | 42,327     | 41,866     | BitcoinWallet.com | Services/Others |         |
| 108 | 000000468 | 40,270     | 1          | Cubits.com | Services/Others |         |
| 109 | 000000480 | 39,013     | 1          | VirWoX.com | Exchanges | Austria |
| 110 | 000000482 | 38,880     | 1          | Huobi.com | Exchanges | China |
| 111 | 000000501 | 36,999     | 36,999     | Justcoin.com | Old/Historic |         |
| 112 | 000000506 | 35,537     | 1          | Bitcoind.de | Exchanges | Germany |
| 113 | 000000508 | 35,537     | 28,567     | SafeDice.com | Gambling |         |
| 114 | 000000511 | 35,453     | 35,453     | MtxNOW.com | Old/Historic |         |
| 115 | 000000512 | 35,433     | 35,433     | C-Cex.com | Exchanges | UK |
| 116 | 000000527 | 34,149     | 34,149     | MiddleEarthMarketplace | Old/Historic |         |
| 117 | 000000533 | 33,436     | 33,389     | VirCurex.com | Exchanges | China |
| 118 | 000000537 | 32,823     | 32,823     | Purse.io | Services/Others |         |
| 119 | 000000539 | 32,701     | 32,701     | SatoshiBet.com | Gambling |         |
| 120 | 000000542 | 32,017     | 27,493     | SwcPoker.eu | Gambling |         |
| 121 | 000000545 | 31,940     | 30,755     | BitBargain.co.uk | Exchanges | UK |
| 122 | 000000556 | 30,965     | 30,965     | SealsWithClubs.eu | Old/Historic |         |
Table A-3 continued from previous page

| No. | User ID | #Addr (1) | #Addr (2) | Name | Category | Country |
|-----|---------|-----------|-----------|------|----------|---------|
| 122 | 0000000559 | 30,624 | 22,599 | BlockTrades.us | Exchanges | USA |
| 123 | 0000000562 | 30,251 | 16,052 | CoinMotion.com | Exchanges | Finland |
| 124 | 0000000563 | 30,187 | 30,183 | OkLink.com | Services/Others | — |
| 125 | 0000000571 | 29,256 | 29,256 | Huobi.com | Exchanges | China |
| 126 | 0000000591 | 27,653 | 23,515 | Bit-x.com | Exchanges | South Africa |
| 127 | 0000000665 | 26,622 | 26,622 | BtcDice.com | Old/Historic | — |
| 128 | 0000000614 | 26,013 | 24,204 | BitBay.net | Exchanges | Poland |
| 129 | 0000000615 | 25,960 | 25,960 | Betcoin.ag | Gambling | — |
| 130 | 0000000626 | 25,257 | 1 | Cubits.com | Services/Others | — |
| 131 | 0000000649 | 23,841 | 20,866 | Paymium.com | Services/Others | — |
| 132 | 0000000672 | 22,597 | 22,597 | Loanbase.com | Services/Others | — |
| 133 | 0000000870 | 20,866 | 20,866 | Betcoin.ag | Services/Others | — |
| 134 | 0000000687 | 17,809 | 17,809 | Coinroll.com | Gambling | — |
| 135 | 0000000678 | 16,490 | 16,490 | FaucetBOX.com | Services/Others | — |
| 136 | 0000000690 | 15,831 | 15,831 | CoinHako.com | Exchanges | Singapore |
| 137 | 0000000717 | 14,042 | 14,042 | FIBSG.com | Exchanges | Singapore |
| 138 | 0000000747 | 19,810 | 19,605 | TheRockTrading.com | Exchanges | Malta |
| 139 | 0000000762 | 18,997 | 18,997 | BlueSkyMarketplace | Old/Historic | — |
| 140 | 0000000774 | 18,489 | 18,489 | Crypto-Games.net | Gambling | — |
| 141 | 0000000794 | 17,705 | 17,705 | Coin-Swap.net | Old/Historic | — |
| 142 | 0000000894 | 17,531 | 17,531 | Luno.com | Exchanges | South Africa |
| 143 | 0000000889 | 15,965 | 15,965 | Anonibet.com | Gambling | — |
| 144 | 0000000887 | 15,757 | 15,757 | Bitmit.net | Old/Historic | — |
| 145 | 0000000891 | 15,495 | 15,495 | CoinApult.com | Services/Others | — |
| 146 | 0000000902 | 15,260 | 15,260 | BtcMarkets.net | Exchanges | Australia |
| 147 | 0000000926 | 14,566 | 14,566 | Inputs.io | Old/Historic | — |
| 148 | 0000000934 | 14,394 | 1 | Huobi.com | Exchanges | China |
| 149 | 0000000959 | 13,713 | 11,210 | Vaultoro.com | Exchanges | UK |
| 150 | 0000001066 | 12,486 | 12,486 | CryptoStocks.com | Services/Others | — |
| 151 | 0000001067 | 12,456 | 12,456 | BitAces.me | Old/Historic | — |
| 152 | 0000001071 | 11,221 | 11,221 | Coins-e.com | Exchanges | Canada |
| 153 | 0000001072 | 11,220 | 11,220 | Igot.com | Exchanges | Belize |
| 154 | 0000001093 | 10,901 | 10,900 | SatoshiRoulette.com | Gambling | — |
| 155 | 0000001117 | 9,667 | 1 | Bittrex.com | Exchanges | USA |
| 156 | 0000001199 | 9,512 | 9,512 | Crypto-Trade.com | Old/Historic | — |
| 157 | 0000001235 | 9,165 | 9,165 | Cryptorush.in | Old/Historic | — |
| 158 | 0000001240 | 9,122 | 9,122 | BTCOracle.com | Gambling | — |
| 159 | 0000001255 | 8,967 | 8,967 | Genesis-Mining.com | Services/Others | — |
| 160 | 0000001313 | 8,430 | 8,430 | Exmo.com | Exchanges | UK |
| 161 | 0000001325 | 8,371 | 4,343 | SlushPool.com | Pools | — |
| 162 | 0000001335 | 8,120 | 8,120 | VaultOfSatoshi.com | Old/Historic | — |
| 163 | 0000001368 | 8,032 | 8,032 | BitcoinVideoCasino.com | Gambling | — |
| 164 | 0000001393 | 7,865 | 7,865 | BTCGuild.com | Old/Historic | — |
| 165 | 0000001395 | 7,857 | 7,766 | Peerbet.org | Gambling | — |
| 166 | 0000001397 | 7,848 | 7,848 | 796.com | Exchanges | Vanuatu |
| 167 | 0000001421 | 7,585 | 7,585 | Btc38.com | Exchanges | UK |
| 168 | 0000001436 | 7,479 | 7,479 | Betcoins.net | Old/Historic | — |
| 169 | 0000001497 | 7,109 | 7,109 | LiteBit.eu | Exchanges | Netherlands |
| 170 | 0000001568 | 6,788 | 6,207 | Bitbond.com | Services/Others | — |
| 171 | 0000001588 | 6,669 | 5,369 | HappyCoins.com | Exchanges | Netherlands |
| 172 | 0000001631 | 6,477 | 6,477 | Bitcoin-Roulette.com | Old/Historic | — |
| 173 | 0000001666 | 6,309 | 6,309 | AllCoin.com | Old/Historic | — |
| 174 | 0000001685 | 6,242 | 6,242 | Coin.mx | Old/Historic | — |
| 175 | 0000001739 | 6,009 | 4,726 | LakeBTC.com | Exchanges | China |
| 176 | 0000001759 | 5,966 | 5,966 | 777Coin.com | Gambling | — |
| 177 | 0000001760 | 5,934 | 5,934 | GHash.io | Pools | — |
| 178 | 0000001796 | 5,762 | 5,762 | DoctorDMarket | Services/Others | — |
| 179 | 0000001862 | 5,481 | 5,481 | Coinomat.com | Exchanges | UK |
| 180 | 0000001916 | 5,297 | 5,295 | Coinmate.io | Exchanges | UK |
| 181 | 0000002012 | 5,024 | 5,024 | BitVC.com | Exchanges | China |
| 182 | 0000002024 | 4,996 | 4,875 | SatoshiCircle.com | Gambling | — |
| 183 | 0000002039 | 4,953 | 1 | Luno.com | Exchanges | South Africa |
| 184 | 0000002055 | 4,896 | 4,896 | MyBitcoin.com | Old/Historic | — |
| 185 | 0000002098 | 4,775 | 4,775 | AllCrypt.com | Old/Historic | — |

Continue to next page
| No. | User ID | #Addr (1) | #Addr (2) | Name | Category | Country |
|-----|---------|-----------|-----------|------|----------|---------|
| 186 | 0000002155 | 4,629 | 4,629 | GermanPlazaMarket | Services/Others | — |
| 187 | 0000002169 | 4,605 | 4,605 | MasterXchange.com | Old/Historic | — |
| 188 | 0000002181 | 4,551 | 4,272 | CoinCafe.com | Exchanges | USA |
| 189 | 0000002192 | 4,530 | 4,242 | BitKonan.com | Exchanges | Croatia |
| 190 | 0000002197 | 4,516 | 4,516 | QuadrigaCX.com | Exchanges | Canada |
| 191 | 0000002216 | 4,451 | 4,451 | BitElfin.com | Old/Historic | — |
| 192 | 0000002243 | 4,377 | 4,377 | OrderBook.net | Exchanges | USA |
| 193 | 0000002251 | 4,363 | 3,800 | SpectroCoin.com | Exchanges | UK |
| 194 | 0000002259 | 4,354 | 4,354 | Bitcurex.com | Exchanges | Poland |
| 195 | 0000002265 | 4,338 | 4,338 | Coinichiwa.com | Gambling | — |
| 196 | 0000002281 | 4,292 | 4,277 | Betcoin.tm | Gambling | — |
| 197 | 0000002298 | 4,232 | 4,227 | MEXBT.com | Exchanges | Mexico |
| 198 | 0000002303 | 3,999 | 3,999 | Bitfinex.com | Exchanges | China |
| 199 | 0000002424 | 3,947 | 1,571 | CoinVault | Old/Historic | — |
| 200 | 0000002464 | 3,861 | 3,861 | BetOfBitco.in | Gambling | — |
| 201 | 0000002471 | 3,840 | 3,840 | JetWin.com | Gambling | — |
| 202 | 0000002587 | 3,607 | 3,607 | BitZillions.com | Gambling | — |
| 203 | 0000002617 | 3,545 | 3,543 | Korbit.co.kr | Exchanges | Korea |
| 204 | 0000002661 | 3,485 | 3,485 | BTCpop.co | Services/Others | — |
| 205 | 0000002649 | 3,216 | 2,181 | YABTC.L.com | Gambling | — |
| 206 | 0000002922 | 3,121 | 3,121 | BitToomRa.com | Old/Historic | — |
| 207 | 0000002952 | 3,086 | 3,086 | BitYes.com | Old/Historic | — |
| 208 | 0000002965 | 3,075 | 2,640 | BetMoose.com | Gambling | — |
| 209 | 0000003031 | 2,979 | 2,978 | CoinURL.com | Services/Others | — |
| 210 | 0000003139 | 2,829 | 2,829 | CannabisRoadMarket | Old/Historic | — |
| 211 | 0000003195 | 2,760 | 2,760 | Ice-Dice.com | Old/Historic | — |
| 212 | 0000003211 | 2,744 | 2,744 | ChBtc.com | Exchanges | China |
| 213 | 0000003249 | 2,713 | 2,713 | CoinArch.com | Exchanges | Singapore |
| 214 | 0000003310 | 2,645 | 2,645 | Comkort.com | Old/Historic | — |
| 215 | 0000003340 | 2,618 | 2,618 | BitNZ.com | Services/Others | — |
| 216 | 0000003348 | 2,614 | 2,614 | CleverCoin.com | Exchanges | USA |
| 217 | 0000003388 | 2,575 | 2,575 | CoinMkt.com | Old/Historic | — |
| 218 | 0000003637 | 2,355 | 2,355 | DiceBitco.in | Old/Historic | — |
| 219 | 0000003756 | 2,276 | 2,276 | BitcoinVietnam.com.vn | Exchanges | Vietnam |
| 220 | 0000003840 | 2,221 | 2,221 | Indacoin.com | Exchanges | UK |
| 221 | 0000004134 | 2,023 | 2,023 | BitClix.com | Services/Others | — |
| 222 | 0000004187 | 1,992 | 1,992 | Coin-Sweeper.com | Old/Historic | — |
| 223 | 0000004271 | 1,948 | 1,948 | GoCellery.com | Services/Others | — |
| 224 | 0000004570 | 1,812 | 1,812 | Playt.in | Old/Historic | — |
| 225 | 0000004580 | 1,804 | 1,796 | Bitcash.cz | Old/Historic | — |
| 226 | 0000004586 | 1,802 | 1,802 | CampBX.com | Exchanges | USA |
| 227 | 0000004817 | 1,713 | 1,713 | BTCendl.org | Services/Others | — |
| 228 | 0000004840 | 1,704 | 1,704 | CoinChimp.com | Exchanges | Russia |
| 229 | 0000004863 | 1,699 | 1,699 | BtcExchange.ro | Old/Historic | — |
| 230 | 0000004882 | 1,690 | 1,690 | AdmiralCoin.com | Old/Historic | — |
| 231 | 0000005062 | 1,643 | 1,643 | Bitcoinica.com | Old/Historic | — |
| 232 | 0000005122 | 1,594 | 1,594 | Gatecoin.com | Exchanges | China |
| 233 | 0000005399 | 1,508 | 1,508 | BetChain.com-old | Gambling | — |
| 234 | 0000005547 | 1,471 | 1,471 | BabylonMarket | Old/Historic | — |
| 235 | 0000005637 | 1,443 | 1,443 | HelixMixer | Services/Others | — |
| 236 | 0000006104 | 1,314 | 1,314 | Bylls.com | Services/Others | — |
| 237 | 0000006381 | 1,246 | 1,246 | Btctst.com-pirateat40 | Old/Historic | — |
| 238 | 0000006688 | 1,189 | 1,189 | PocketRocketsCasino.eu | Old/Historic | — |
| 239 | 0000006697 | 1,188 | 1,188 | Bitso.com | Exchanges | Mexico |
| 240 | 0000006747 | 1,178 | 1,178 | BTC.com | Old/Historic | — |
| 241 | 0000006753 | 1,176 | 1,176 | DaDice.com | Old/Historic | — |
| 242 | 0000007201 | 1,095 | 1,095 | Cryptonit.net | Exchanges | UK |
| 243 | 0000007312 | 1,076 | 1,076 | BitStarz.com | Gambling | — |
| 244 | 0000007547 | 1,040 | 1,040 | Ccedk.com | Exchanges | Denmark |
| 245 | 0000007673 | 1,020 | 1,020 | Satoshi-Karoshi.com | Gambling | — |
| 246 | 0000007783 | 1,005 | 1 | VirWoX.com | Exchanges | Austria |
| 247 | 0000008133 | 978 | 978 | Just-Dice.com | Old/Historic | — |
| 248 | 0000008213 | 968 | 968 | CryptoLocker | Old/Historic | — |
| 249 | 0000008232 | 965 | 965 | GreenRoadMarket | Services/Others | — |

Continue to next page
| No.  | User ID     | #Addr (1) | #Addr (2)   | Name                  | Category          | Country |
|------|-------------|-----------|-------------|-----------------------|-------------------|---------|
| 250  | 0000008325  | 953       | 953         | CoinRoyale.com        | Gambling          | —       |
| 251  | 0000008352  | 950       | 950         | CryptoBounty.com      | Old Historic      | —       |
| 252  | 000009025   | 876       | 876         | iCoin.com             | Exchanges         | China   |
| 253  | 000009160   | 869       | 679         | Coingi.com            | Exchanges         | Panama  |
| 254  | 000010744   | 746       | 746         | BitcoinWebank.com     | Old Historic      | —       |
| 255  | 000010833   | 741       | 741         | EmpoEX.com            | Exchanges         | USA     |
| 256  | 000011035   | 729       | 729         | FairProof.com         | Gambling          | —       |
| 257  | 000011546   | 700       | 700         | UseCryptos.com        | Exchanges         | Portugal|
| 258  | 000011749   | 686       | 1           | Cubits.com            | Services/Others   | —       |
| 259  | 000011857   | 681       | 1           | VirWoX.com            | Exchanges         | Austria |
| 260  | 000012513   | 670       | 669         | AntPool.com           | Pools             | —       |
| 261  | 000012560   | 668       | 668         | CoinBrokers.io       | Exchanges         | USA     |
| 262  | 000012639   | 665       | 665         | UpDown.BT             | Old Historic      | —       |
| 263  | 000013022   | 648       | 648         | DiceNow.com           | Gambling          | —       |
| 264  | 000013139   | 643       | 643         | Dagensia.eu           | Old Historic      | —       |
| 265  | 000013870   | 614       | 614         | WatchMyBit.com        | Services/Others   | —       |
| 266  | 000014175   | 604       | 604         | MPEx.co               | Old Historic      | —       |
| 267  | 000014703   | 593       | 593         | Banx.io               | Exchanges         | USA     |
| 268  | 000015353   | 572       | 572         | CloudHashing.com      | Old Historic      | —       |
| 269  | 000015549   | 565       | 565         | Eligius.st            | Pools             | —       |
| 270  | 000016775   | 527       | 527         | EuropeX.eu            | Old Historic      | —       |
| 271  | 000017189   | 515       | 515         | EveryDice.com         | Old Historic      | —       |
| 272  | 000018179   | 499       | 499         | Brawker.com           | Old Historic      | —       |
| 273  | 000019489   | 473       | 473         | 10xBitco.in           | Old Historic      | —       |
| 274  | 000021056   | 440       | 1           | Cubits.com            | Services/Others   | —       |
| 275  | 000021084   | 440       | 439         | BitMinter.com         | Pools             | —       |
| 276  | 000021286   | 436       | 436         | ExchangeMyCoins.com   | Exchanges         | Denmark |
| 277  | 000021385   | 402       | 402         | BW.com                | Pools             | —       |
| 278  | 000023394   | 400       | 400         | ChainRoll.com         | Old Historic      | —       |
| 279  | 000024768   | 390       | 390         | DiceCoin.io           | Gambling          | —       |
| 280  | 000025249   | 382       | 382         | FoxBit.com.br         | Exchanges         | Brazil  |
| 281  | 000025389   | 380       | 380         | PonziCoin.co          | Old Historic      | —       |
| 282  | 000025554   | 377       | 377         | Birwo.com-old         | Old Historic      | —       |
| 283  | 000027509   | 351       | 338         | Zyado.com             | Exchanges         | Portugal|
| 284  | 000028366   | 341       | 341         | SuzukiDice.com        | Old Historic      | —       |
| 285  | 000028897   | 334       | 1           | Cubits.com            | Services/Others   | —       |
| 286  | 000032628   | 297       | 297         | KnCMiner.com          | Pools             | —       |
| 287  | 000032825   | 293       | 293         | BitcoinPokerTables.com| Gambling          | —       |
| 288  | 000034753   | 280       | 280         | Polmine.pl            | Old Historic      | —       |
| 289  | 000034912   | 279       | 279         | MineField.BitcoinLab.org| Gambling          | —       |
| 290  | 000045961   | 234       | 234         | SmenarnaBitcoin.cz    | Old Historic      | —       |
| 291  | 000062962   | 195       | 195         | Dgex.com              | Old Historic      | —       |
| 292  | 000064830   | 190       | 190         | BitLaunder.com        | Services/Others   | —       |
| 293  | 000072390   | 170       | 170         | BitMillions.com       | Old Historic      | —       |
| 294  | 000074148   | 166       | 1           | Cubits.com            | Services/Others   | —       |
| 295  | 000078381   | 157       | 1           | Cubits.com            | Services/Others   | —       |
| 296  | 000080904   | 125       | 125         | Vic-Socks.to          | Services/Others   | —       |
| 297  | 0000808438  | 117       | 117         | Gatecoin.com          | Exchanges         | China   |
| 298  | 000096655   | 116       | 1           | VirWoX.com            | Exchanges         | Austria |
| 299  | 000096665   | 112       | 112         | Bitcoin-24.com        | Old Historic      | —       |
| 300  | 000097019   | 96        | 1           | Cubits.com            | Services/Others   | —       |
| 301  | 000098141   | 85        | 1           | Poloniex.com          | Exchanges         | USA     |
| 302  | 000098218   | 77        | 77          | BetcoinDice.tm        | Old Historic      | —       |
| 303  | 000098094   | 75        | 75          | Bitfury.org           | Pools             | —       |
| 304  | 0000265341  | 59        | 1           | Cubits.com            | Services/Others   | —       |
| 305  | 0000296609  | 53        | 53          | 50BTC.com             | Old Historic      | —       |
| 306  | 0000302266  | 52        | 1           | Cubits.com            | Services/Others   | —       |
| 307  | 0000302626  | 52        | 1           | Cubits.com            | Services/Others   | —       |
| 308  | 0000332619  | 49        | 49          | BtcEur.eu             | Old Historic      | —       |
| 309  | 0000417570  | 40        | 1           | VirWoX.com            | Exchanges         | Austria |
| 310  | 0000434094  | 38        | 1           | Luno.com              | Exchanges         | South Africa |
| 311  | 0000482184  | 35        | 35          | SecureVPN.to          | Services/Others   | —       |
| 312  | 0000489304  | 34        | 1           | Xapo.com              | Services/Others   | —       |
| 313  | 0000539744  | 32        | 32          | MinersCenter.com      | Old Historic      | —       |
| No. | User ID      | #Addr. (1) | #Addr. (2)          | Name              | Category       | Country       |
|-----|--------------|-----------|--------------------|-------------------|----------------|--------------|
| 314 | 000661573    | 28        | 28                | DiceOnCrack.com   | Old/Historic   |              |
| 315 | 000727720    | 26        | 1                 | Cubits.com        | Services/Others|              |
| 316 | 000881383    | 23        | 1                 | Bittrex.com       | Exchanges      | USA          |
| 317 | 001047894    | 20        | 20                | ActionCrypto.com  | Old/Historic   |              |
| 318 | 001068642    | 20        | 1                 | Cubits.com        | Services/Others|              |
| 319 | 001453639    | 15        | 1                 | CoinMotion.com    | Exchanges      | Finland      |
| 320 | 001482058    | 15        | 15                | ASICMiner         | Old/Historic   |              |
| 321 | 001615111    | 14        | 14                | Telco214          | Pools          |              |
| 322 | 001849964    | 12        | 12                | BTradeAustralia.com| Exchanges      | Australia    |
| 323 | 00194672    | 12        | 12                | PinballCoin.com   | Old/Historic   |              |
| 324 | 002236961    | 11        | 1                 | MoonBit.co.in     | Services/Others|              |
| 325 | 002347246    | 10        | 1                 | Cubits.com        | Services/Others|              |
| 326 | 002652769    | 9         | 1                 | Xapo.com          | Services/Others|              |
| 327 | 002867260    | 9         | 1                 | Cubits.com        | Services/Others|              |
| 328 | 003042379    | 8         | 1                 | Xapo.com          | Services/Others|              |
| 329 | 003656659    | 7         | 1                 | Cubits.com        | Services/Others|              |
| 330 | 003733567    | 7         | 1                 | Luno.com          | Exchanges      | South Africa |
| 331 | 004053833    | 7         | 7                 | SimpleCoin.cz     | Exchanges      | Czech        |
| 332 | 004615543    | 6         | 6                 | Lucky8.it         | Gambling       |              |
| 333 | 004799731    | 1         | 1                 | Cubits.com        | Services/Others|              |
| 334 | 005083698    | 6         | 1                 | Cubits.com        | Services/Others|              |
| 335 | 005134566    | 6         | 1                 | Cubits.com        | Services/Others|              |
| 336 | 005316858    | 6         | 1                 | LakeBTC.com       | Exchanges      | China        |
| 337 | 006291300    | 5         | 1                 | Cryptopay.me      | Services/Others|              |
| 338 | 006553479    | 5         | 1                 | Xapo.com          | Services/Others|              |
| 339 | 006787887    | 5         | 1                 | Cubits.com        | Services/Others|              |
| 340 | 006884235    | 5         | 5                 | Exchanging.ir     | Exchanges      | Iran         |
| 341 | 007827688    | 4         | 1                 | Cubits.com        | Services/Others|              |
| 342 | 0080964964   | 4         | 4                 | FoxBit.com.br     | Exchanges      | Brazil       |
| 343 | 008251183    | 4         | 1                 | Cubits.com        | Services/Others|              |
| 344 | 009938059    | 4         | 1                 | Cubits.com        | Services/Others|              |
| 345 | 010541278    | 4         | 1                 | Bitcoind.de       | Exchanges      | Germany      |
| 346 | 010933272    | 4         | 4                 | ePay.info         | Services/Others|              |
| 347 | 011084834    | 4         | 1                 | Cubits.com        | Services/Others|              |
| 348 | 011993584    | 3         | 1                 | Cubits.com        | Services/Others|              |
| 349 | 013456456    | 3         | 3                 | EclipseMC.com     | Pools          |              |
| 350 | 014115232    | 3         | 1                 | Xapo.com          | Services/Others|              |
| 351 | 014366385    | 3         | 1                 | Cubits.com        | Services/Others|              |
| 352 | 014572029    | 3         | 1                 | StrongCoin.com-fee| Services/Others|              |
| 353 | 015920521    | 3         | 1                 | Poloniex.com      | Exchanges      | USA          |
| 354 | 015254424    | 3         | 1                 | Cubits.com        | Services/Others|              |
| 355 | 017326362    | 3         | 1                 | Cubits.com        | Services/Others|              |
| 356 | 018289896    | 3         | 1                 | Huobi.com         | Exchanges      | China        |
| 357 | 018738066    | 3         | 1                 | Cubits.com        | Services/Others|              |
| 358 | 020356858    | 3         | 1                 | Poloniex.com      | Exchanges      | USA          |
| 359 | 021062591    | 3         | 1                 | Cubits.com        | Services/Others|              |
| 360 | 021163228    | 3         | 3                 | ePay.info         | Services/Others|              |
| 361 | 021447975    | 2         | 1                 | Cubits.com        | Services/Others|              |
| 362 | 022923752    | 2         | 1                 | Cubits.com        | Services/Others|              |
| 363 | 023937394    | 2         | 1                 | Cubits.com        | Services/Others|              |
| 364 | 024428379    | 2         | 1                 | Cubits.com        | Services/Others|              |
| 365 | 026928525    | 2         | 1                 | Cubits.com        | Services/Others|              |
| 366 | 033718688    | 2         | 1                 | Xapo.com          | Services/Others|              |
| 367 | 034243835    | 2         | 1                 | VirWox.com        | Exchanges      | Austria      |
| 368 | 034774492    | 2         | 1                 | Cubits.com        | Services/Others|              |
| 369 | 035023024    | 2         | 1                 | MercadoBitcoin.com.br| Exchanges      | Brazil       |
| 370 | 035903871    | 2         | 1                 | Huobi.com         | Exchanges      | China        |
| 371 | 036244970    | 2         | 1                 | Cubits.com        | Services/Others|              |
| 372 | 036353675    | 2         | 1                 | Cubits.com        | Services/Others|              |
| 373 | 036389109    | 2         | 2                 | Dispenser.tf      | Old/Historic   |              |
| 374 | 039716536    | 2         | 1                 | Poloniex.com      | Exchanges      | USA          |
| 375 | 040913069    | 2         | 1                 | Cubits.com        | Services/Others|              |
| 376 | 045999467    | 2         | 2                 | DeepBit.net       | Old/Historic   |              |
| 377 | 046840841    | 2         | 1                 | Cubits.com        | Services/Others|              |

Continue to next page
| No. | User ID   | Addr. (1) | Addr. (2) | Name       | Category     | Country       |
|-----|-----------|-----------|-----------|------------|--------------|---------------|
| 378 | 0047562289| 2         | 1         | Cubits.com | Services/Others | —             |
| 379 | 0049075064| 2         | 1         | Cubits.com | Services/Others | —             |
| 380 | 0050709449| 2         | 1         | Poloniex.com | Exchanges | USA           |
| 381 | 0051958689| 2         | 2         | CoinWorker.com | Services/Others | —             |
| 382 | 0052140982| 2         | 1         | Cubits.com | Services/Others | —             |
| 383 | 0053961615| 2         | 1         | Cubits.com | Services/Others | —             |
| 384 | 0055586746| 2         | 1         | Cubits.com | Services/Others | —             |
| 385 | 0056053804| 2         | 1         | Luno.com   | Exchanges     | South Africa  |
| 386 | 0058656155| 2         | 1         | Xapo.com   | Services/Others | —             |
| 387 | 0058696998| 2         | 1         | Cubits.com | Services/Others | —             |
| 388 | 0058767317| 2         | 1         | Cubits.com | Services/Others | —             |
| 389 | 0059001161| 2         | 1         | Bittrex.com | Exchanges     | USA           |

*End of Table A-3*
Appendix B: Illustrating Adjacent Matrices of $G_t$ or $X_{sd}$

Fig. B.1 illustrates the adjacency matrices for all the networks $G_t$ in Table I, or equivalently the non-negative matrix $X_{sd}$ given by (15).

**Fig. B.1.** Adjacent matrices for all the networks $G_t$ in Table I, or equivalently $X_{sd}$ given by (15). Colors show the strength of links, expressed by the frequencies.
Appendix C: Topic Model of Latent Dirichlet Allocation (LDA)

Topic models in natural language processing and machine learning are probabilistic models for how terms appear in documents in a given corpus. Assumed is the term frequency, irrespective of the sequential order of term occurrences; such an assumption is often called “bag-of-words”. A topic is a latent variable in the model for explaining term frequency occurrences. Latent Dirichlet allocation (LDA) [41] is such a model widely used in a variety of applications. See [37, 42, 43] and references therein for introduction and applications. In this appendix, we briefly summarize the topic model only for the purpose of explicitly relating it with non-negative matrix factorization (NMF).

Documents in a corpus (a collection of documents) are given with a vocabulary of $V$ terms (different words). Document $d$ ($d = 1, \ldots, D$) comprises $N_d$ words, possibly with duplication, each denoted by $w_{dn}$ ($n = 1, \ldots, N_d$). $K$ topics are assumed to be present for the given corpus. A topic is a latent or hidden variable giving a probability distribution for term occurrences. If a document is an article on sports, terms like “swimming” and “gymnastics” are likely to occur, while “market” and “employee” will be unlikely.

Consider the example of an article on “influence of hosting Olympics to economy” in a corpus of newspaper. It would be natural to consider that the article belongs to two topics of sports and economy. In topic models, documents are not assumed to belong to a single topic, but to simultaneously belong to multiple topics, and the topics vary among documents. Topic models aim at modeling such mixed membership.

Document $d$ has a topic distribution, which is a multinomial distribution with the parameters:

$$\theta_d = (\theta_d1, \ldots, \theta_dK), \quad (C·1)$$

where $\theta_{dk} = p(k | \theta_d)$ is the probability that a topic $k$ is assigned to the document $d$, satisfying

$$\theta_{dk} \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \theta_{dk} = 1. \quad (C·2)$$

Then a topic $z_{dn} \in \{1, \ldots, K\}$ is assigned to each word $w_{dn}$ for $n = 1, \ldots, N_d$. Each topic $k$ is a term distribution, which is also a multinomial distribution with the parameters:

$$\phi_k = (\phi_k1, \ldots, \phi_kV), \quad (C·3)$$

where $\phi_{kv} = p(v | \phi_k)$ is the probability that a term $v$ occurs, satisfying

$$\phi_{kv} \geq 0 \quad \text{and} \quad \sum_{v=1}^{V} \phi_{kv} = 1. \quad (C·4)$$

The topic model of LDA can be most readily understood by looking at how it generates words in documents as follows.

1. For each topic $k = 1, \ldots, K$, generate the parameters $\phi_k$ for the term distribution by

$$\phi_k \sim \text{Dir}(\beta), \quad (C·5)$$

where $\sim$ reads that a variable is realized as a sample from the distribution on the right-hand side, and $\text{Dir}(\cdot)$ is the Dirichlet distribution\(^7\). $\beta = (\beta_1, \ldots, \beta_V)$ is a set of hyper-parameters with $\beta_i > 0$.

\(^7\)Dirichlet distribution is defined by

$$\text{Dir}(x | \xi) = \frac{\Gamma\left(\sum_{i=1}^{I} \xi_i\right)}{\prod_{i=1}^{I} \Gamma(\xi_i)} \times \prod_{i=1}^{I} x_i^{\xi_i-1},$$

where $I$ is the number of variables and it is assumed that $x_i \geq 0$ for all $i$ and $\sum_{i=1}^{I} x_i = 1$. $\Gamma(\cdot)$ is the gamma function defined by $\Gamma(u) = \int_0^\infty t^{u-1} e^{-t} dt$ for $u > 0$. Dirichlet distribution is a generalization of Beta distribution (the case $I = 2$).
2. Now for each document \( d = 1, \ldots, D \),
   (a) generate the parameters \( \theta_d \) for the topic distribution by
   \[
   \theta_d \sim \text{Dir}(\alpha),
   \]  
   where \( \alpha = (\alpha_1, \ldots, \alpha_K) \) is a set of hyper-parameters with \( \alpha_k > 0 \).
   (b) For each word \( w_{dn} \) (\( n = 1, \ldots, N_d \))
      (i) generate a topic by
      \[
      z_{dn} \sim \text{Cat}(\theta_d),
      \]  
      where \( \text{Cat}(\cdot) \) is the categorical distribution\(^8\), and then
      (ii) generate a word from the vocabulary by
      \[
      w_{dn} \sim \text{Cat}(\phi_{k=z_{dn}}).
      \]

Note that a topic is assigned to each occurrence of word \( w_{dn} \) at the location \( n \) of the document \( d \), depending on how the document \( d \) has a mixture of topics, but is not associated with each term \( v \) in the vocabulary of \( V \) terms. In the previous example, the term “swimming” occurs with a relatively high probability if the topic of sports is assigned at the location of occurrence, but the term will not be likely to occur if the topic of economy is assigned at the same location.

Given the number of topics, \( K \), and the hyper-parameters, \( \beta \) and \( \alpha \), the model \( M \) can be specified by the parameters \( \{ \theta_d \}_{d=1}^D \) and \( \{ \phi_k \}_{k=1}^K \), while the data \( D \) comprises \( D \) documents, each of which is a set of words \( \{ w_{dn} \}_{n=1}^{N_d} \) for \( d = 1, \ldots, D \). In the Bayesian framework of
   \[
   p(M|D) = \frac{p(D|M) \cdot p(M)}{p(D)},
   \]
the prior distribution \( p(M) \) is given by the Dirichlet distributions of (C·5) and (C·6). It is easy to show that the likelihood \( p(D|M) \) can be written by
   \[
   \log p(D|M) = \sum_{d=1}^{D} \sum_{n=1}^{N_d} \log \sum_{k=1}^{K} \theta_{dk} \cdot \phi_{kw_{dn}},
   \]
which follows from the assumption of bag-of-words and the independence of documents. Estimation of parameters by maximum likelihood or Bayesian frameworks has a difficulty due to the log of sum in (C·10), so there are a variety of technical methods such as EM (expectation and maximization), variational Bayes, and MCMC (Markov chain Monte Carlo) algorithms (see [37, 42, 43] and references therein).

For our purpose, let us turn out attention to how a word is chosen from the \( V \) terms in the vocabulary in the \( D \) documents, given the model parameters. Looking at (C·7) and (C·8), one can see that the probability that a term \( v \) is chosen as a word in the document \( d \), denoted by \( \lambda_{dv} \), is given by
   \[
   \lambda_{dv} = \sum_{k=1}^{K} \theta_{dk} \cdot \phi_{kv},
   \]
where it is assumed that
   \[
   \lambda_{dv} \geq 0 \quad \text{and} \quad \sum_{v=1}^{V} \lambda_{dv} = 1.
   \]
\(^8\)Categorical distribution is a multinomial distribution for a single trial (a roll of a dice), i.e.
   \[
   \text{Cat}(v|\phi) = \phi_v,
   \]
where \( v \in \{1, \ldots, V\} \) and \( \phi = (\phi_1, \ldots, \phi_V) \) with \( \phi_v \geq 0 \) for all \( v \) and \( \sum_v \phi_v = 1 \).
(Note that (C·11) is the factor in the likelihood (C·10)). In terms of matrices, the $D \times V$ matrix, denoted by $\Lambda$, with elements $\lambda_{dv}$ is represented by a product of two matrices; the $D \times K$ matrix $\Theta$ with elements $\theta_{dk}$ times the $K \times V$ matrix $\Phi$ with elements $\phi_{kv}$. Note that it is usually the case that $K \ll V$ and $K \ll D$. One can immediately see that (C·11) is actually a non-negative matrix factorization (NMF) of $\Lambda$ by two generally much smaller matrices:

$$\Lambda = \Theta \cdot \Phi$$

under the conditions of (C·2) and (C·4), from which (C·12) follows.

A document-term matrix is the $D \times V$ matrix, the rows and columns of which correspond to the documents and terms respectively, and each element is the frequency (or any other measure that represents some intensity) for the term occurrences. One can regard the document-term matrix as a realization in the sense that each element at column $v$ and row $d$ represents how the term $v$ was realized in the document $d$ according to the categorical (multinomial) distribution with the parameter $\lambda_{dv}$. See the main text in Section 3.4.

How to determine the number of topics, $K$, is important. There have been a number of works on this issue. A class of studies [38–40] is based on some measure to compute similarities of extracted topics, called coherence. The idea is that an optimal value of $K$ is the point where the overall dissimilarity among topics achieves a maximum, because extracted topics should be different to a certain degree. Other approaches include the idea of perplexity [37], hierarchical Dirichlet process [44], and so forth. It is beyond the present paper to review these approaches. Interested readers are guided to look at them and reviews [42, 43] with references therein. In the main text, we use the idea of coherence [38–40], which we find to work well for our purpose.
Appendix D: Optimal Number of NMF Components

In Section 3.4, we showed that the matrix decomposition of NMF can be regarded as a probabilistic model, which is equivalent to a model of probabilistic latent semantic analysis (PLSA) in natural language processing, in particular, latent Dirichlet allocation (LDA). The number of NMF components is such an analogous parameter as the number of topics in the latter. Then, as a bonus, one can employ the method of estimating the number of topics to our problem of determining the number of NMF components, $K$. We found that three different methods in the literature [38–40] give a consistent result for the optimal value of $K$, and also that one of them [39] is relatively stable in giving the optimal (see Fig. 8).

In Fig. D-1, we summarize the same result for the data in all the other months in the same year. We found that the optimal number $K$ is not volatile in its variation among different months, and also small in its value ranging around 11–18. Table D-1 shows the optimal values for all the months of the year 2019.

Table D-1. Optimal values for the number of NMF components, $K$, estimated from Fig. D-1

| Month | Optimal $K$ |
|-------|-------------|
| 01    | 14          |
| 02    | 16          |
| 03    | 18          |
| 04    | 16          |
| 05    | 16          |
| 06    | 18          |
| 07    | 15          |
| 08    | 12          |
| 09    | 13          |
| 10    | 11          |
| 11    | 13          |
| 12    | 13          |
Fig. D-1. Determining $K$, the number of NMF components, by minimizing the measure of coherence [39]. We performed Monte Carlo simulations with 20 runs for each number of components. Averages (points) and 99% level (gray band, narrow) calculated from standard errors are drawn. Data: all months in the year 2019.