Received 20 March 2017
Revised 19 October 2017
Accepted 16 November 2017

DIGITAL CURRENCIES

The evolution of the bitcoin economy

Extracting and analyzing the network of payment relationships

Paolo Tasca
University College London, London, UK

Adam Hayes
University of Wisconsin Madison, Madison, Wisconsin, USA, and

Shaowen Liu
Deutsche Bundesbank, Frankfurt, Germany

Abstract

Purpose – This paper aims to gather together the minimum units of users' identity in the Bitcoin network (i.e. the individual Bitcoin addresses) and group them into representations of business entities, what we call "super clusters". While these clusters can remain largely anonymous, the authors are able to ascribe many of them to particular business categories by analyzing some of their specific transaction patterns (TPs), as observed during the period from 2009 to 2015. The authors are then able to extract and create a map of the network of payment relationships among them, and analyze transaction behavior found in each business category. They conclude by identifying three marked regimes that have evolved as the Bitcoin economy has grown and matured: from an early prototype stage; to a second growth stage populated in large part with "sin" enterprise (i.e. gambling, black markets); to a third stage marked by a sharp progression away from "sin" and toward legitimate enterprises.

Design/methodology/approach – Data mining.

Findings – Four primary business categories are identified in the Bitcoin economy: miners, gambling services, black markets and exchanges. Common patterns of transaction behavior between the business categories and their users are a "one-day" holding period for bitcoin transactions is somewhat typical. That is, a one-day effect where traders, gamblers, black market participants and miners tend to cash out on a daily basis. There seems to be a strong preference to do business within the bitcoin economy in round lot amounts, whether it is more typical of traders exchanging for fiat money, gamblers placing bets or black market goods being bought and sold. Distinct patterns of transaction behavior among the business categories and their users are flows between traders and exchanges average just around 20 BTC, and traders buy or sell on average every 11 days. Meanwhile, gamblers wager just 0.5 BTC on average, but re-bet often within the same day. Three marked regimes have evolved, as the Bitcoin economy has grown and matured: from an early prototype stage, to a second growth stage populated in large part with "sin" enterprises (i.e. gambling, black markets), to a third stage marked by a sharp progression away from "sin" and toward legitimate enterprises. This evolution of the Bitcoin economy suggests a trend toward legitimate commerce.

Originality/value – The authors propose a new theoretical framework that allows investigating and exploring the network of payment relationships in the Bitcoin economy. This study starts by gathering together the minimum units of Bitcoin identities (the individual addresses), and it goes forward in grouping

JEL classification – E42, L14, O12, O35, P40
This paper forms part of a special section "Digital currencies", guest edited by Paolo Tasca.

The Journal of Risk Finance
Vol. 19 No. 2, 2018
pp. 94-126

1526-5943
DOI 10.1108/JRF-03-2017-0059

them into approximations of business entities, what is called "super clusters", by using tested techniques from the literature. A super cluster can be thought of as an approximation of a business entity in that it describes a number of individual addresses that are owned or controlled collectively by the same beneficial owner for some special economic purposes. The majority of these important clusters are initially unknown and uncategorized. The novelty of this study is given by the pure user group and the TP analyses, by means of which the authors are able to ascribe the super clusters into specific business categories and outline a map of the network of payment relationships among them.

Keywords Business analysis, Bitcoin, Network theory, Blockchain, Digital currencies, Payment traceability

Paper type Research paper

1. Introduction

As the price of Bitcoin has risen above $4,000 recently, the world's largest and most used decentralized cryptocurrency has become a topic of interest among a variety of disciplines – from economics, computer science and payments to public policy, information systems and the law[1]. No longer a curio for hobbyists, Bitcoin is now being taken quite seriously by academics and practitioners around the globe. Even central banks such as the Federal Reserve and Bank of England have started to take notice. Indeed, the cumulative value of all bitcoins has risen above $70bn, with the number of transactions on the Bitcoin blockchain rising exponentially from around 1,000 per day in 2011 to more than 300,000 per day at the moment of writing. At the current exchange rate the notional value of daily turnover approaches $1bn[2].

With a surge in both user base and interest from the outside, understanding what goes on within the Bitcoin economy makes an important contribution. For example, Bitcoin still carries a negative connotation among some who associate the cryptocurrency with illegal activity, made prescient for instance by the take-down of the online black market Silk Road by the FBI in late 2013. What our analysis shows is that by 2015 such illicit exchange made up only a very small proportion of all Bitcoin activity. This does not mean that black market activity has gone away, rather their users and operators have shifted to alternate digital currencies such as ZCash and Monero, as Bitcoin has matured into a legitimate financial institution. At the same time, investors are now seeking to add Bitcoin to their portfolios as a diversifier and a number of financial firms are beginning to accommodate this demand.

It is thus appropriate as an academic pursuit to explore how the Bitcoin economy is populated and extract the map of payment relationships, and to furthermore to trace the evolution of those relationships over time to build up a better understanding of its political economy. This paper takes a step in that direction by identifying the interconnectedness of economic agents that use the Bitcoin payment network to transfer the digital currency among each other internally (meaning within-Bitcoin transactions and not transactions to exchange Bitcoin for other currency). To do this, we start by identifying and clustering together the minimum units of Bitcoin identity, which are the individual "addresses", into what we call "super clusters". We then we tag those clusters using a novel method to de-anonymize economically relevant addresses and sort them into distinct categories. Finally, we describe the dynamics of how these clusters behave over time.

In this context, a super cluster can be thought of as an approximation of a discrete business entity in that it describes a group of Bitcoin addresses that are owned or controlled collectively for some particular economic purpose by the same party[3]. Although exact identities of such super clusters can remain unknown, we are able to allocate many of them to specific business categories – as either an exchange, mining pool, online gambling site and black market or composite of two or more of these categories – by analyzing their specific transaction patterns (TPs), as observed during the period 2009-2015. With this information, we unveil and study the

Bitcoin network of payment relationships both among super clusters and also between super clusters and their users: traders; gamblers; or black market user-dealers[4].

We are subsequently able to identify three distinct regimes that have existed in the Bitcoin political economy, as it has grown and developed. First, a "proof of concept" or "mining-dominated" phase, followed by a "sin" or "gambling/black market-dominated" phase, and finally a "maturation" or "exchange-dominated" phase. The novelty of our study, moreover, is to elaborate and advance a general de-anonymization methodology that allows us to link clusters composed of groups of addresses to identifiable business categories, which we use to map the system's evolution.

It is possible to accomplish such a map of activity and interaction among Bitcoin users because *pseudonimity*, rather than strict anonymity, is a defining characteristic of the Bitcoin network (Reid and Harrigan, 2013). As such, the true identities of users are hidden behind their addresses that work as aliases, but which may be revealed upon transacting with somebody else[5]. In other words, if Alice remits payment to Bob, then their identities will be revealed to one another by virtue of exchanging addresses to send or receive bitcoin.

There are a few approaches suggested for revealing such identities in a systematic manner. One proposed approach for de-anonymization is by mapping Bitcoin addresses to identifiable IP addresses. Kaminsky (2011) proposes that "if we are able to connect to every node, the [IP of the] first node to inform you of a transaction of the source is it". Informed by this idea, Koshy *et al.* (2014) conducted the first trial using this method and managed to map nearly 1,000 Bitcoin addresses to their owners' IPs. This method, however, is greatly limited when transactions are executed through proxy services, which is not an inconsequential caveat. Another approach is to cluster Bitcoin addresses into a single entity and then try to link this entity with a "real" name, as described by Lischke and Fabian (2016); our work here continues along this second track.

There are two general procedures that must be clearly defined at the onset: "clustering" and "labeling". Clustering refers to grouping together all the addresses that belong to the same beneficial owner (i.e. a legal entity or individual person) into a unique assemblage. This approach requires one to apply what we will call either the "input address heuristic" and/or the "change address heuristic", which are described in detail just below[6]. After clustering, one can then apply labeling, which consists of either[7]:

- manually tagging Bitcoin addresses to specific entities by directly participating in Bitcoin transactions with those entities; or
- scraping information from Web pages on the internet where, for any reason, the identity of Bitcoin address holders is made public and can be extracted.

According to the *input address heuristic*, Bitcoin addresses used as inputs either synchronously in the same multi-input transactions or asynchronously in different multi-input transactions (when at least one input address is shared), are grouped together in clusters. In other words, if address x and address y are both inputs to a unique transaction, then we assume addresses x and y must also belong to the same cluster. Furthermore, if both address y and address z belong to some other transaction, we would infer that addresses x, y and z all belong to the same cluster. From the beginning of Bitcoin, Nakamoto (2008) indirectly recognized the power of the input address method by saying that:

> [s]ome linking is still unavoidable with multi-input transactions, which necessarily reveal that their inputs were owned by the same owner. The risk is that if the owner of a [public] key is revealed, linking could reveal other transactions that belonged to the same owner.

Later, Ron and Shamir (2013) extensively discuss the input address method and apply it via the Union-Find graph algorithm in a study of the Bitcoin network through the May 13, 2012.

Within the scope of this heuristic, other projects including Spagnuolo (2013) and Doll *et al.* (2014) try to provide some practical applications of the theory by developing front-end Web services to show, in real-time, identity correspondent to a specific address query. In particular, Spagnuolo (2013) proposes the BitIodine tool, which parses the blockchain and clusters addresses that are likely to belong to the same user or group of users, classifies such users and labels them.

With the *change address heuristic*, a cluster is composed of the input addresses plus the output addresses that are predicted to be change addresses for a transaction. A first proposal approximating this heuristic comes from Androulaki *et al.* (2013), who naively assume that "[i]n the current Bitcoin implementation, users rarely issue transactions to two different users". Presumably, this assumption perhaps once held in the past, but it is no longer the case. Therefore, this initial version of the change address heuristic is relatively fragile compared to the input address heuristic, and aggressive implementations require large amounts of hand tuning to prevent false positives[8]. Meiklejohn *et al.* (2013) reevaluate the change address heuristic and apply it cautiously by identifying only one-time change addresses under the following conditions:

- the transaction is not one that involves new coin generation;
- it is the first appearance of the address and at the same time not the first appearance for all other output addresses (i.e. all the other addresses have been previously used); and
- there is no other address in the output that is the same as the input address (i.e. no self-change address).

The assumption behind this enhanced version of the change address heuristic is that the change address is newly generated by the user's wallet; even the owner may not acknowledge its existence. In contrast, the receiver's address is known in advance and notified to the sender. Thus, also the one-time change address requires significant human adjustment to avoid excessive false positives when:

- the receiver is a new user or creates a new address never used before;
- the transaction output has two receivers' addresses without change address; and
- the sender uses an old address to receive change, or there is no change transaction at all.

Despite some studies that rely on this version of the change address heuristic, e.g. Garcia *et al.* (2014), for the purpose of our study, we opt for using a version of the input address heuristic. Although our method is subject to some false negatives as it only considers eligible for clustering addresses being used as transaction inputs, it is nonetheless robust to false positives. This is crucial as any false positive would compromise the results of the pattern analysis we apply later to ascribe the clusters to particular business categories; in such a case, the clusters themselves would be composed of wrong addresses that would likely follow incorrect behavioral patterns. This pitfall is avoided in our methodology.

False positives could have conceivably become a problem with the introduction, since 2013, of the "coinjoin" practice; refer Kristov Atlas (2015) and Tasca (2015). Coinjoin is an example of a tool used to actively anonymize transactions within a distributed ledger. The principle behind the method is quite simple: if for example, Alice wants to send one bitcoin to Bob, and Carla wants to send one bitcoin to David, a coinjoin transaction could be established whereby the addresses of Alice and Carla are both listed as inputs, and the addresses of Bob and David are listed as outputs in one unique transaction. Thus, when inspecting the 2-to-2 transaction from outside it is impossible to discern who is the sender and who the recipient. In this hypothetical example, we would not be able to tell if it is Bob or David who is the recipient of Alice. If this

sketch of the coinjoin principle were its actual implementation, then the input address method could mistakenly cluster together the addresses of Alice and Carla as if they belonged to the same entity. However, in practice, a coinjoin transaction works in a crucially different way: the coinjoin technique shuffles the addresses of users, but it also creates *new* batches of unique addresses that are subsequently added to the users' addresses and mixed together with these transactions as well. The result is that coinjoin addresses could be reused several times along with several other addresses from different users. Thus, novel unknown large clusters are created that do not belong to any precise business category because their addresses are very likely linked to more than two distinct entities or directly to coinjoin service providers. Those clusters are similar to "black holes" as they "absorb" addresses that should have been enclosed within other clusters having a clear business profile and are not then misattributed to one of our known categories. To sum up this explanation, even in the presence of coinjoin transactions, our method is robust because the likelihood of encountering false positives is for all intents and purposes negligible.

In this study, the result of applying the input address heuristic returns more than 30 million clusters, which reduces to a more manageable 2,850 when considering only those composed of at least 100 addresses and that have received at least one thousand bitcoins from January 2009 through May 2015[9]. We label such clusters "super clusters" because they represent big agents with a strong presence and intensity of economic activity in the Bitcoin system. All together, these super clusters transacted hundreds of billions of dollars worth of notional value over the study period, at the current exchange rate.

We acknowledge that it is impractical to correctly identify all of the 2,850 individual super clusters in the sample. However, our study has a less ambitious aim, which is to ascribe all these super clusters to a given broader business category and explore their network of business relationships, rather than drill down on individual identities. With that in mind, from a list of publicly available pre-identified addresses obtained from the internet we do successfully identify 209 super clusters out of the 2,850 as a seed for attributing unknown clusters to business categories. In other words, this subset of known clusters, which we call the "known group", is used as the benchmark to identify the business category of the remaining 2,641 clusters in the "unknown group". Following from this, we conclude our study by unveiling the network of payment relationships between these 2,850 super clusters, and by exploring the relative interdependence among business categories.

The paper proceeds as follows: Section 2 introduces some preliminary definitions; Section 3 describes the data set we draw upon; Section 4 introduces pure user group (PUG) analysis to classify super clusters in the unknown group using what we know of the known group; Section 5 elaborates on the PUG analysis with a TP analysis, examining transaction inflows and outflows among those super clusters in known group; Section 6 back-tests the results from the PUG analysis; and Section 7 describes the network of entities on the Bitcoin network as discerned from the PUG and TP analyses and develops the progression of three distinct regimes that have existed over the course of the Bitcoin political economy.

2. Preliminary definitions

As explained in Section 1, the building block of our analysis is the concept of clustering Bitcoin addresses. In this section, we provide a formal definition of clustering by omitting unnecessary technical information which may turn out to be redundant and therefore not useful for the scope of our analysis.

We define the set Tx of all the Bitcoin transactions, occurred during the period of our analysis, as $Tx = (tx_1, \ldots, tx_i, \ldots, tx_z)$. To each element tx_i of Tx corresponds the cluster set $c_i = (a_1, a_2, \ldots a_n)_i$ containing all the input addresses $(a_1, a_2, \ldots)$ used in the transaction tx_i.

By using a variant of a Union-Find graph algorithm (**?**), if two or more clusters directly or indirectly (via other clusters) have at least one address in common, we merge those clusters into a single unique one. At the end of the merging process, we get $C = \{c_1, \ldots, c_x, \ldots, c_y, \ldots, c_z\}$ which is the set of all disjoint clusters such that $c_x \cap c_y = \varnothing$ for all $c_x, c_y \in C$. Let $W(C)$ be a finite set $W(C) = \{w_{xy}(c_x, c_y) \mid c_x, c_y \in C, c_x \neq c_y\} \cup \{w_{xx}(c_x, c_x) \mid \forall c_x \in C\}$. Then, $W \subseteq W(C)$ is the set of all (direct) transaction (with loops[10]) between clusters, where w_{xy} is the total quantity of bitcoins transferred from cluster c_x to cluster c_y:

$$w_{xy} = \begin{cases} w_{xy} & \text{if there is a transation from } c_x \text{ to } c_y. \\ 0 & \text{otherwise.} \end{cases}$$

We define a super cluster, $\hat{c}_x$, as any special cluster that belongs to the partition $\hat{C} \subset C$:

$$\hat{C} = \left\{ c_x \in C \middle| \sum_{h=1}^{z} w_{hx}(c_h, c_x) \geq 1,000\, BTC \wedge n(c_x) \geq 100 \right\}. \quad (1)$$

where $n(c_x)$ denotes the number of addresses in cluster x.

According to our definition, a super cluster is any cluster that satisfies the following two thresholds:

(1) having received at least 1,000 bitcoins during our research window; and

(2) comprising at least 100 unique addresses.

The first threshold is necessary to increase the likelihood of excluding inactive entities from the analysis. The second threshold is necessary to exclude as many private individuals as possible, who typically own only one or a few addresses. Together, these thresholds increase the robustness of the TP analysis of the clusters in Section 5, which is based on statistics requiring big enough data. In fact, smaller clusters composed of some tens, or even some hundreds of addresses are only able to generate a trivial amount of transaction data, giving us insufficient information to perform a meaningful analysis.

3. Data set

In our study, we parsed data from the Bitcoin Core over the period of the January 3, 2009 (block 0) through May 8, 2015 (block 355551)[11]. Over this interval, the Bitcoin network proliferates both in terms of number of addresses and in terms of number of transactions. Refer Table I for a summary of our data set. All the data related to Bitcoin transactions are imported into and managed via a MySQL database (see the diagram in Figure A1 in Appendix 1).

By applying the input address heuristic, 75,191,953 unique Bitcoin addresses are grouped into 30,708,660 clusters, of which, about two-thirds are clusters composed of only a single address, as shown in Table II.

Then, by applying the criteria defined in equation (1), 2,850 super clusters are filtered out. Figure 1 shows the network of super clusters $\hat{C}$ and their transactions among each other, as well as with all the remaining clusters in $C \backslash \hat{C}$.

By gathering publicly available address information, we are able to link part of super clusters $\hat{c}_x \in \hat{C}$ to real-world entities (e.g. BTCChina, Kraken, Xapo) which belong to different business categories. Specifically, we gathered 359,776 deciphered addresses from ? and ?. According to their entity information, we could compose a group of deciphered sets of addresses, $P = \{p_1, p_2, \ldots, p_Y, \ldots\} = \underset{Y \in \Gamma}{S} p_Y$. Precisely, $p_\gamma = \{a \mid a$ belong to the known

beneficial owner $\gamma\}$ is the set of addresses that belong to the beneficial owner γ whose identity is publicly available from the internet[12].

Thus, depending on whether a super cluster hold at least one address belonging to any $p_\gamma \in P$ or not, $\hat{C}$ is then decomposed into either a known group, $\hat{C}^K$, or a unknown group, $\hat{C}^U$. Formally:

$$\hat{C} = \hat{C}^K \cup \hat{C}^U \tag{2}$$

with $\hat{C}^K \cap \hat{C}^U = \varnothing$ by definition and:

$$\hat{c}_x \in \begin{cases} \hat{C}^K & \text{if } \hat{c}_x \cap p_\gamma \neq \varnothing \wedge \hat{c}_x \cap \left(P \backslash p_\gamma\right) = \varnothing, \forall_\gamma \in \Gamma \\ \hat{C}^U & \text{if} \begin{cases} \hat{c}_x \cap p_\gamma \neq \varnothing \wedge \hat{c}_x \cap \left(P \backslash p_\gamma\right) \neq \varnothing, \forall_\gamma \in \Gamma \\ \hat{c}_x \cap p_\gamma = \varnothing, \forall_\gamma \in \Gamma. \end{cases} \end{cases} \tag{3}$$

The matching exercise turns out the following result: $n(\hat{C}^K) = 209$ and $n(\hat{C}^U) = 2{,}641$ such that $n(\hat{C}^K) + n(\hat{C}^U) = n(\hat{C}) = 2{,}850$[13]. As a side note, we remark that equation (3) follows a prudential principle that aims to avoid false positives. Namely, any cluster in $\hat{C}$ that has addresses linked to more than one set $p_\gamma \in P$, is considered unknown and confined to the set $\hat{C}^U$[14].

Then, according to their business model, each identified super cluster is allocated into one of the following primary business categories: *exchange* $\hat{C}^{KX}$, *mining pool* $\hat{C}^{KP}$, *online gambling* $\hat{C}^{Ktt}$, *black market*, $\hat{C}^{KB}$. Besides these big four business categories which are populated by economic entities with a clear business profile, there are also few other economic entities with a business models (e.g. bitcoin wallets) heterogeneous among them and disparate from the previous ones. Then, we classify them into the category *others*, $\hat{C}^{KO}$.

Table AI in the Appendix 2 shows us the results, namely, $n(\hat{C}^{KX}) = 104$, $n(\hat{C}^{KP}) = 18$, $n(\hat{C}^{Ktt}) = 45$, $n(\hat{C}^{KB}) = 13$ and $n(\hat{C}^{KO}) = 29$ such that $n(\hat{C}^{KX}) + n(\hat{C}^{KP}) + n(\hat{C}^{Ktt}) + n(\hat{C}^{KB}) + n(\hat{C}^{KO}) = n(\hat{C}^K) = 209$.

Figure 2 shows the payment network of the 209 identified super clusters[15] in $\hat{C}^K$. In the next sections, we will use the information on the super clusters in the set $\hat{C}^K$ together with the information on their interactions with all the other clusters in $C \backslash \hat{C}^K$ to derive the business membership of each unknown super cluster in $\hat{C}^U$.

4. Pure user group analysis

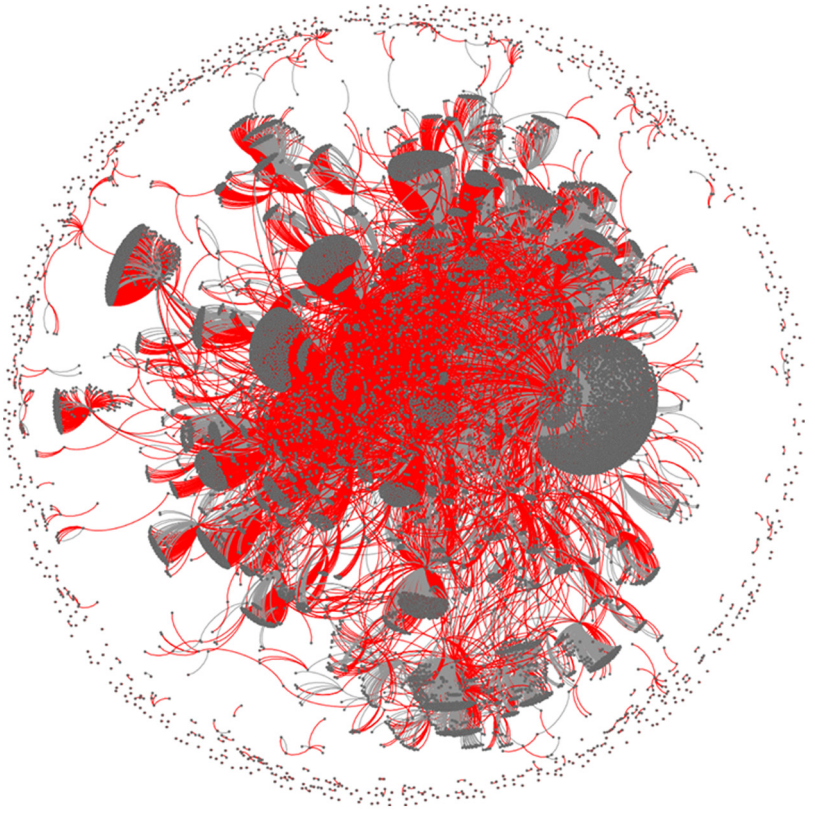
The PUG analysis is carried out to classify (into specific business categories) super clusters in the unknown group, and it is based on the definition and classification of "pure" users. By

Bitcoin core parsed from the January 3, 2009 until May 8, 2015

Max block height	355,551
Total number of transactions	68,030,042
Total number of input	172,743,139
Total number of output	194,476,567
Total addresses identified	75,191,953
Total clusters identified(include at least one address)	30,708,660
Number of clusters with at least two addresses	9,847,999
Total transactions between clusters	88,950,021

Table I. Blockchain database facts

Source: Bitcoin core



Notes: Every red node represents a single super cluster $\hat{c}_x \in \hat{C}$ and every gray node represents a counterpart cluster. For visualization purposes, we set a threshold of at least 1,000 BTC transferred between a super cluster and its counterpart. Therefore, the plot shows 1,957 super clusters out of 2,850 in $\hat{C}$. One may observe that some of the clusters are highly connected to each other, although many are isolated. These isolated entities could be individuals, some highly self-contained businesses, or some clusters that belong to active business entities but which are kept untied from the others, i.e. used independently for purposes different from the main business activity

Figure 1. Network visualization of the interactions of the super clusters in $\hat{C}$ with each other and also with all the remaining clusters in CC

pure users we mean all those clusters populating the Bitcoin economy (except for those already in the known group) that had bilateral transactions with super clusters (in the known group) belonging to only one business category. In other words, for each specific business category, we build a correspondent PUG:

- clusters having transactions only with exchanges in $\hat{C}^{KX}$ are classified in the PUG *traders*;
- clusters having transactions only with gambling services in $\hat{C}^{Ktt}$ are classified in the PUG *gamblers*; and

- clusters having transactions only with black market services in $\hat{C}^{KB}$ are classified in the PUG black market *user-dealers*.

The classification of the clusters into different PUGs is the first step of the PUG analysis. The second step consists of classifying the super clusters in the unknown group into a specific business category in the case they transact *only* with the corresponding specific PUG. For example, those super clusters in the unknown group that had transactions only with traders are classified as exchanges and so on also for the other categories. However, the clusters in the known group identified in the categories mining pools and others follow a peculiar business model. Thus, we do not create the set of PUGs having transactions only with mining pools in $\hat{C}^{KP}$ because the mining pools in the unknown group will be identified via the coinbase analysis (Section 4.2). Similarly, we do not create the set of PUGs having transactions only with others in $\hat{C}^{KO}$ because those clusters do not have a clearly defined business profile. In other terms, to avoid false positives, we will not try to classify super clusters in the unknown group into the category others.

4.1 Pure user group identification

In this first part of the analysis, we consider only the following sets $\hat{C}^{KX}$, $\hat{C}^{Ktt}$ and $\hat{C}^{KB}$. Accordingly, we introduce the following set notation: $U^X \subset C \backslash \hat{C}^K$ is the subset of pure traders that had transactions only with exchanges in $\hat{C}KX$; $U^{tt} \subset C \backslash \hat{C}^K$ is the subset of pure gamblers that had transactions only with gambling sites in $\hat{C}^{Ktt}$; $U^B \subset C \backslash \hat{C}^K$ is the subset of pure black market user-dealers that had transactions only with black markets in $\hat{C}^{KB}$. Formally:

$$U^X = \left\{ c_x \in C \backslash \hat{C}^K \middle| \exists \hat{c}_y \in \hat{C}^{KX} : w_{xy}(c_x, \hat{c}_y) > 0 \vee w_{y,x}(\hat{c}_y, c_x) > 0 \right.$$

$$\left. \sum \wedge w_{xj}(c_x, c_j) = 0 \vee w_{jx}(c_j, c_x) = 0 \sum, \forall cj \in \hat{C}^K \backslash \hat{C}^{KX} \right\}. \quad (4)$$

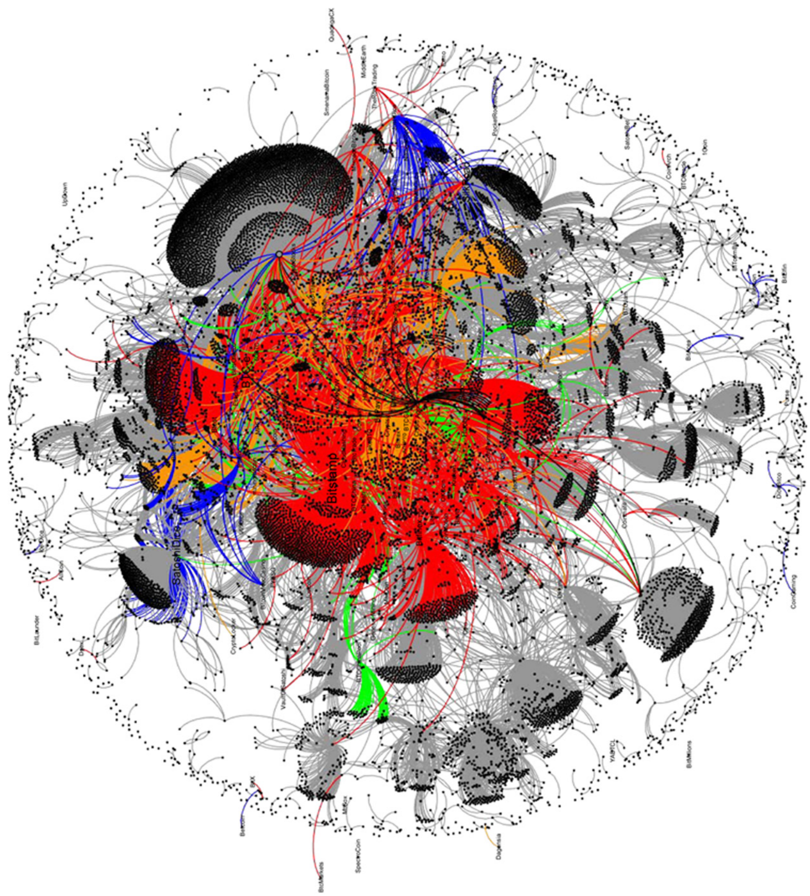
$$U^{tt} = \left\{ c_x \in C \backslash \hat{C}^K \middle| \exists \hat{c}_y \in \hat{C}^{Ktt} : w_{xy}(c_x, \hat{c}_y) > 0 \vee w_{y,x}(\hat{c}_y, c_x) > 0 \right.$$

$$\left. \sum \wedge w_{xj}(c_x, c_j) = 0 \vee w_{jx}(c_j, c_x) = 0 \sum, \forall cj \in \hat{C}^K \backslash \hat{C}^{Ktt} \right\}. \quad (5)$$

Table II. The clusters identified with the input address heuristic are grouped per number of addresses composing them

Input address heuristic: clustering result	
Number of addresses included in each cluster	Number of clusters identified
> 10001	194
$1001 \sim 10000$	1,145
$101 \sim 1000$	12,185
$11 \sim 100$	436,093
$2 \sim 10$	9,398,382
$=1$	20,860,661
Total number of clusters	30,708,660

Source: Bitcoin Core



Notes: As it happens that more than one super cluster may belong to single entities in Γ, we combine them into one node in the network. For visualization purposes, we set a threshold of at least 5,000 BTC being transferred between a super cluster $\hat{c}_x \in \hat{C}$ and its counterpart. Therefore, the plot shows only 94 super clusters out of 209 in $\hat{C}^K$. The gray nodes are the counterparts of each $\hat{c}_x \in \hat{C}$. Each super cluster is colored according to its business category: green for miners, red for exchange, blue for gambling, orange for others, black for black market, purple for composite category and gray for the clusters which are the counterparts. The color of the edge is the same as the source nodes. One may clearly observe some large entities with many counterparts, such as Silkroad (black market), SatoshiDice (online gambling), BitStamp (exchange) and BTC-e (exchange)

Figure 2.
Network visualization of the 209 super clusters in the set $\hat{C}^K$ that have been identified by cross-linking the known addresses in the set P with the addresses in each $\hat{c}_x$ $\hat{C}$

$$U^B = \left\{ c_x \in C \backslash \hat{C}^K \middle| \exists \hat{c}_y \in \hat{C}^{KB} : w_{xy}(c_x, \hat{c}_y) > 0 \vee w_{y,x}(\hat{c}_y, c_x) > 0 \right.$$

$$\left. \sum \wedge w_{xj}(c_x, c_j) = 0 \vee w_{jx}(c_j, c_x) = 0 \sum, \forall cj \in \hat{C}^K \backslash \hat{C}^{KB} \right\}. \quad (6)$$

This first part of the PUG analysis returns the following results: $n(U^X) = 440,434$, $n(U^{tt}) = 415,528$ and $n(U^B) = 74,233$.

The statistics of the bitcoin transactions between pure users and clusters in the known group reveal that the average volume per transaction differs substantially with respect to each business category: The average volume per transaction from/to traders to/from exchanges is 20 BTC; the average volume per transaction from/to gamblers to/from gambling services is 0.5 BTC; and finally, the average volume per transaction from/to user-dealers to/from black market services is 3 BTC (Table III).

4.2 Identification of mining pools

The super clusters in the category mining pool are identified without using the PUG analysis. Indeed, each newly generated Bitcoin block includes a reward to the successful miner: an amount equal to the sum of the block reward (or subsidy), i.e. newly available bitcoins, plus any accumulated fees paid by transactions included in that block. To allocate this sum, a new generation transaction is created whose input, called the "coinbase", contains the reward for the miners. Thus, unlike all other transaction inputs, the coinbase is not linked to any previous output. This feature offers a simple and direct method to identify those clusters belonging to the mining category by filtering out the transactions with "null" input and only one output.

Let $\hat{C}^{U\ Coinbase}$ be the set of clusters (in the unknown group) composed (also, but not only) of addresses with coinbase inputs, $n(\hat{C}^{U\ Coinbase}) = 575$. Not all these 575 clusters, however, can reliably be defined as mining pools; for some of them, mining is not their primary activity and rewards from coinbase transactions represent only a small per cent of their activity. To make sure that the taxonomy is robust, we classify only clusters in the unknown group whose mining rewards occupy more than 80 per cent of its total income, as mining pool, $\hat{C}^{UP}$. The remaining clusters $\hat{C}^{UP-} = \hat{C}^{U\ Coinbase} \backslash \hat{C}^{UP}$ that cannot be defined as mining pools according to our threshold are instead classified via the PUG analysis.

4.3 Classification of unknown super clusters

The principle of PUG classification for unknown clusters is straightforward and works as follows: If one super cluster in the unknown group transacts only with one specific

Table III.
For PUG in different categories, the table summarizes the average transaction amount (BTC) and average transaction interval (minutes)

		Statistics for PUG transaction			
		PUG → $\hat{C}^K$		$\hat{C}^K$ → PUG	
PUG	Num of clusters	Avg tx volume (BTC)	Avg tx interval (min)	Avg tx volume (BTC)	Avg tx interval (min)
U^X	440,434	23.4	20,529.5	17.6	10,685.0
U^{tt}	415,528	0.5	528.3	0.5	387.3
U^B	74,233	2.7	22,151.3	3.4	9,394.0

PUG, then we suspect that this cluster belongs to the business category correspondent to that specific PUG. For example, if during the period January 2009-May 2015, one super cluster in $\hat{C}^U$ records transactions with one or more traders in U^X but not with gamblers in U^{tt} and user-dealers in U^B, it is classified as an exchange. One should note that this does not rule out the possibility for the exchange to transact with any other cluster in C beyond those in U^X. The clusters who transact with multiple PUGs are identified in the composite category, $\hat{C}^{UM}$, which implies those super clusters might have multi-business lines.

Let:

$$\hat{C}^{U\overline{X}} = \left\{ \hat{c}_x \in \hat{C}^U \middle| \exists c_y \in U^X : w_{yx}(c_y, \hat{c}_x) > 0 \wedge w_{xy}(\hat{c}_x, c_y) > 0 \sum \right\} \tag{7}$$

be a broad subset of exchanges in $\hat{C}^U$ that have transactions *not only* with traders.

Let:

$$\hat{C}^{U\overline{t}} = \left\{ \hat{c}_x \in \hat{C}^U \middle| \exists c_y \in U^{tt} : w_{yx}(c_y, \hat{c}_x) > 0 \wedge w_{xy}(\hat{c}_x, c_y) > 0 \sum \right\} \tag{8}$$

be a broad subset of gambling services in $\hat{C}^U$ that have transactions *not only* with gamblers.

Let:

$$\hat{C}^{U\overline{B}} = \left\{ \hat{c}_x \in \hat{C}^U \middle| \exists c_y \in U^B : w_{yx}(c_y, \hat{c}_x) > 0 \wedge w_{xy}(\hat{c}_x, c_y) > 0 \sum \right\} \tag{9}$$

be a broad subset of black market services in $\hat{C}^U$ that have transactions *not only* with user-dealers.

Then, the subset of exchanges in $\hat{C}^U$ that have transactions only with traders in U^X is:

$$\hat{C}^{U\overline{X}} = \left\{ \hat{c}_x \in \hat{C}^{U\overline{X}} \setminus \left(\hat{C}^{U\overline{t}} \cup \hat{C}^{U\overline{B}} \cup \hat{C}^{U\overline{P}} \right) \right\}. \tag{10}$$

Similarly, the subset of gambling services in $\hat{C}^U$ that have transactions only with gamblers in U^{tt} is:

$$\hat{C}^{Utt} = \left\{ \hat{c}_x \in \hat{C}^{U\overline{t}} \setminus \left(\hat{C}^{U\overline{X}} \cup \hat{C}^{U\overline{B}} \cup \hat{C}^{U\overline{P}} \right) \right\}. \tag{11}$$

The subset of black market services in $\hat{C}^U$ that have transactions only with user-dealers in U^B is:

$$\hat{C}^{UB} = \left\{ \hat{c}_x \in \hat{C}^{U\overline{B}} \setminus \left(\hat{C}^{U\overline{X}} \cup \hat{C}^{U\overline{t}} \cup \hat{C}^{U\overline{P}} \right) \right\}. \tag{12}$$

Finally, the subset of multi-business clusters in $\hat{C}^U$ that have transactions with more than one user group is:

$$\hat{C}^{UM} = \left\{ \hat{c}_x \in \left(\hat{C}^{U\overline{X}} \cup \left(\hat{C}^{U\overline{B}} \cup \hat{C}^{U\overline{t}} \cup \hat{C}^{U\overline{P}} \right) \backslash \left(\hat{C}^{UX} \cup \left(\hat{C}^{Utt} \cup \hat{C}^{UB} \cup \hat{C}^{UP} \right) \sum \right\} \right. . \tag{13}$$

Table IV shows that $n(\hat{C}^{UX}) = 310$, $\hat{C}^{Utt} = 755$, $\hat{C}^{UB} = 41$, $\hat{C}^{UP} = 57$ and $\hat{C}^{UM} = 630$.

5. Transaction pattern analysis

In this section, we introduce a TP analysis to study the different TPs of the super clusters in set $\hat{C}^{K}$ (listed in Table AI). The TP analysis is used to garner more insights into stylized facts characterizing the distinct business behaviors of the super clusters. Moreover, the TP analysis is used in Section 6 to measure the accuracy of the PUG analysis by testing the pattern similarity between the clusters in $\hat{C}^{K}$ and those in $\hat{C}^{U}$. In the following, we divide the TP analysis in *inflow* and *outflow* analysis.

5.1 Inflow analysis

The inflow analysis consists of examining the properties of the transactions *toward* any super clusters in the known group $\hat{C}^{K_i}$. We select the transactions in the set:

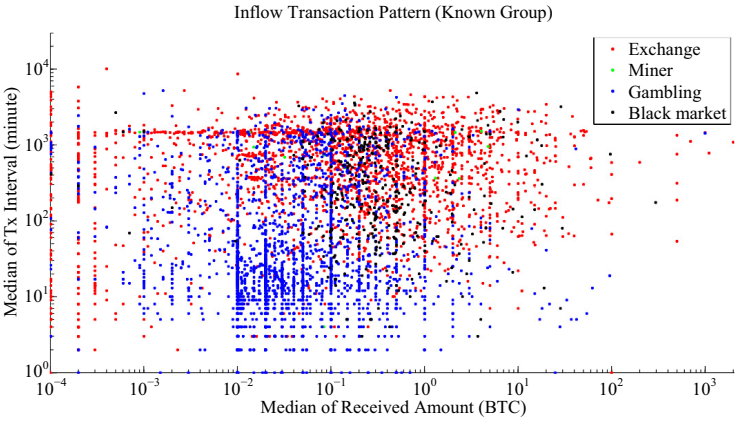
$$\overrightarrow{W}^{K} \subset W = \left\{ w_{yx}(c_y, \hat{c}_x) \in W \middle| c_y \in C \backslash \hat{C}^{K}, \hat{c}_x \in \hat{C}^{K}, n_{yx} \geq 100 \right\} \tag{14}$$

where n_{yx} denotes the number of transactions from cluster c_y to cluster $\hat{c}_x$ during the period of the analysis. According to equation (14), a pair $(c_y, \hat{c}_x) \in \overrightarrow{W}^{K}$ is only considered if there has been at least 100 transactions from c_y to $\hat{c}_x$. This minimum transaction threshold is subjective and shall be set to any value able to ensure that the descriptive statistics calculated are robust. In our case, with $n_{yx} \geq 100$, we obtain that $n(\overrightarrow{W}^{K}) = 11,899$, involving 148 super clusters in x$\hat{C}^{K}$x. After having defined the set of analysis, we calculate the median of transaction volume and the median of time interval in minutes for each pair $(c_y, \hat{c}_x) \in \overrightarrow{W}^{K}$ [16]. Each dot in Figure 3 represents the measurement for one pair $(c_y, \hat{c}_x) \in \overrightarrow{W}^{K}$: red if $\hat{c}_x \in \hat{C}^{KX}$, green if $\hat{c}_x \in \hat{C}^{KP}$, blue if $\hat{c}_x \in \hat{C}^{Ktt}$ and black if $\hat{c}_x \in \hat{C}^{KB}$. The x-axis is the median of the transaction volume and the y-axis is the median of the time interval (in minutes) between inflow transactions for each pair $(c_y, \hat{c}_x) \in \overrightarrow{W}^{K}$.

Table IV. The number of clusters tagged with the PUG method

Tagged cluster in unknown group	
Category	No. of clusters
$n(\hat{C}UX)$	310
$n(\hat{C}Utt)$	755
$n(\hat{C}UP)$	57
$n(\hat{C}UB)$	41
$n(\hat{C}UM)$	630

Notes: To give the reader a complete view, $\hat{C}^{UP}$ is also listed here, which is identified from coin base transactions



Notes: Each dot characterizes one pair of clusters $(c_y, \ddot{c}_x) \in \overrightarrow{W^K}$. The x-axis is the median transaction volume of all transactions between all the pairs of clusters $\in \overrightarrow{W^K}$ during the period January 2009-May 2015. The y-axis is the median transaction interval (in minutes) of the transactions between all the pairs of clusters $\in \overrightarrow{W^K}$

Figure 3.
Inflow TP for the known group

Figure 3 shows some clustering effects; we can see that for each of our four identified business categories there exists specific patterns of transaction behavior. For example, there are clearly plotted in blue, vertical lines at $x = 0.01$, 0.02 and so on. To capture this more clearly, we plot the kernel density in Figure 4[17]. This illustrates a notable characteristic for gambling behavior, that is gamblers tend to place bets with similar, round lot amounts

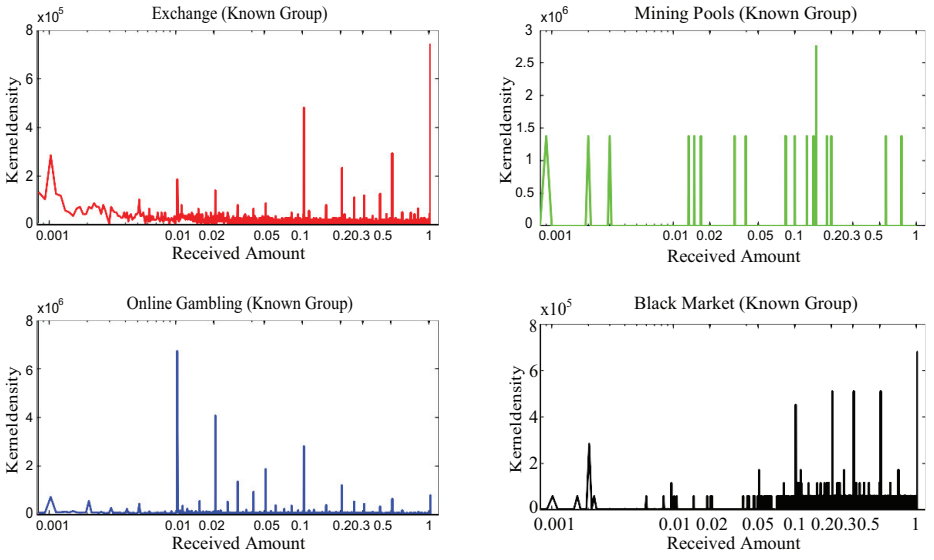


Figure 4.
Kernel density of the inflow amount of bitcoins received by any $c_y \in C \backslash \hat{C}^K$ during the period January 2009-May 2015 by each cluster $\hat{c}_x \in \hat{C}^k$ in the known group

(i.e. 0.1, 0.5, 1.0) again and again, with wagers of 0.01 BTC being placed most frequently. Gamblers may be accustomed to wagering in round amounts in traditional settings using casino or poker chips with specified round values (e.g. $1, $5 or $25), or online using virtual chips. Individuals may carry forward that behavior to bitcoin-based gambling even in instances where the size of bets is determined arbitrarily by the gambler placing bets[18].

With respect to exchanges, although it is less obvious due to some overlap in the plot, we are still able to see some vertical lines in red at $x = 0.1$, 0.5 and 1.0, which indicates that the traders usually deposit into exchanges round amounts of bitcoins, rather than random amounts to presumably exchange them for fiat or alternative digital currency. In other words, it appears that traders may wait until they have accumulated some even amount, most commonly 1.0 BTC, before selling them.

Inflows to black markets show a wider variety of arbitrary transaction size, but still also show marked preference for round lots of bitcoin, notably at amounts of 0.1, 0.2, 0.3, 0.5 and 1.0 BTC. This may suggest that black market sellers explicitly place round lot prices on their items as a matter of doing business. Prescription and illegal drugs are notably sold on black markets, and this indicates that sellers will offer an amount of contraband that corresponds to a round price (say 1.0 BTC), rather than determining what the price would be for a fixed quantity (say for 1 ounce)[19].

Mining pools exhibit a more or less random pattern of inflows, as a mining pool will only be credited with small amounts of bitcoin whenever it finds a new block of bitcoin. When this happens, the pool will generally extract a small profit consisting of either a nominal percentage of the block reward or of the transaction fees associated with that block, or both.

In addition to studying patterns in the amounts of bitcoin inflows, we also consider transaction intervals. We observe a large density of dots, plotted in red, clustering horizontally just above $y = 1{,}000$ in Figure 3, specifically at 1,440 min, which is the number of minutes in one day. What this shows us is that there are a large number of traders who send small amounts of bitcoins to exchanges regularly each day. We suspect that these could be small miners who exchange mined bitcoins for cash on a daily basis, or "day traders" who are active daily but go home flat, having sold out any positions in bitcoin to avoid overnight price volatility. Figure 5 clarifies this effect and shows the kernel density of the intervals between transactions (band = 100 min). The 1,440-min interval is prominent not only for traders to exchanges but also for the other business categories, suggesting that a "one-day" holding period for bitcoin transactions is somewhat typical; a one-day effect where traders, gamblers, black market participants and miners tend to cash out on a daily basis.

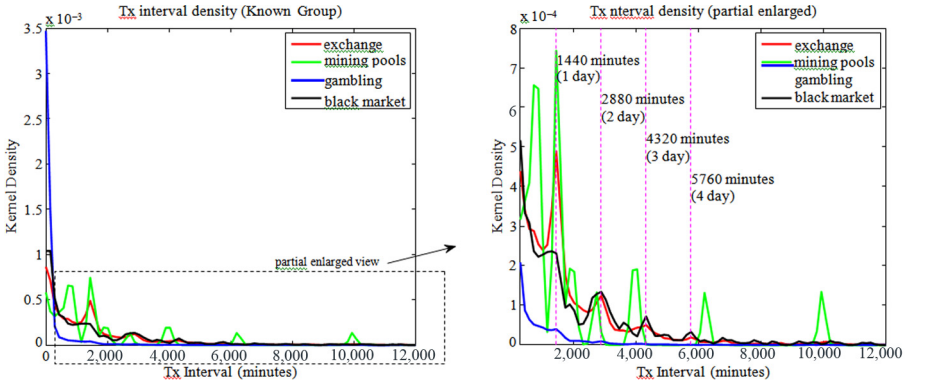


Figure 5.
Kernel density of the inflow transaction intervals between subsequent transactions during the period January 2009-May 2015 for each category in the known group

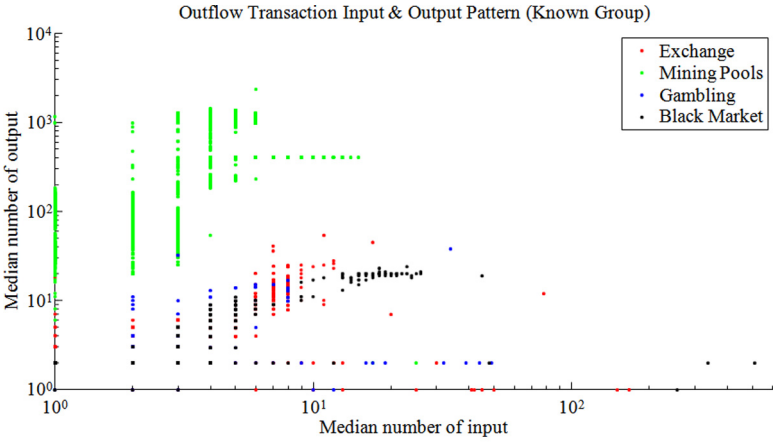
We observe however that gambling has, by far, the shortest interval as well as the highest transaction frequency. This is not difficult to understand, as gamblers can ante or re-bet many times in a matter of minutes.

5.2 Outflow analysis

The outflow analysis consists of examining the properties of the transactions *from* the clusters in the known group $\hat{C}^K$. As for the inflow analysis we measure the median of transaction volume and the median of time interval in minutes. Additionally, we measure the median number of inputs and outputs in the transactions between each pair of clusters. To examine the transaction outflow from the super clusters $\hat{c}_x \in \hat{C}^K$, we select the transactions in the set:

$$\overleftarrow{W}^K \subset W = \left\{ w_{xy}(\hat{c}_x, c_y) \in W \middle| \hat{c}_x \in \hat{C}^K, c_y \in C \backslash \hat{C}^K, n_{xy} \geq 100 \right\} \quad (15)$$

where n_{xy} denotes the number of transactions from cluster $\hat{c}_x$ to cluster c_y during the period of the analysis. According to equation (15), a pair $(\hat{c}_x, c_y) \in \overleftarrow{W}^K$ is considered only if there has been at least 100 transactions from $\hat{c}_x$ to c_y. From our database, we obtain that $n(\overleftarrow{W}^K) =$ 16,188, involving 148 super clusters in $\hat{C}^K$. To extract information about the number of inputs/outputs of the transactions in $\overleftarrow{W}^K$ we plot the median number of inputs/outputs in all the transactions between each pair $(\hat{c}_x, c_y) \in \overleftarrow{W}^K$. Each dot in in Figure 6 represents a value set for pair $(\hat{c}_x, c_y) \in \overleftarrow{W}^K$: red if $\hat{c}_x \in \hat{C}^{KX}$, green if $\hat{c}_x \in \hat{C}^{KP}$, blue if $\hat{c}_x \in \hat{C}^{Ktt}$, and black if $\hat{c}_x \in \hat{C}^{KX}$. The *x*-axis represents the median number of inputs, and the *y*-axis represents the median number of outputs.



Notes: Each dot characterizes one pair of clusters. The x-axis measures the median number of inputs and the y-axis measures the median number of outputs

Figure 6.
Median number of inputs and outputs for all the transactions among each pair $(\hat{c}_x, c_y)\ \overleftarrow{W}^K$ during the period January 2009-May 2015

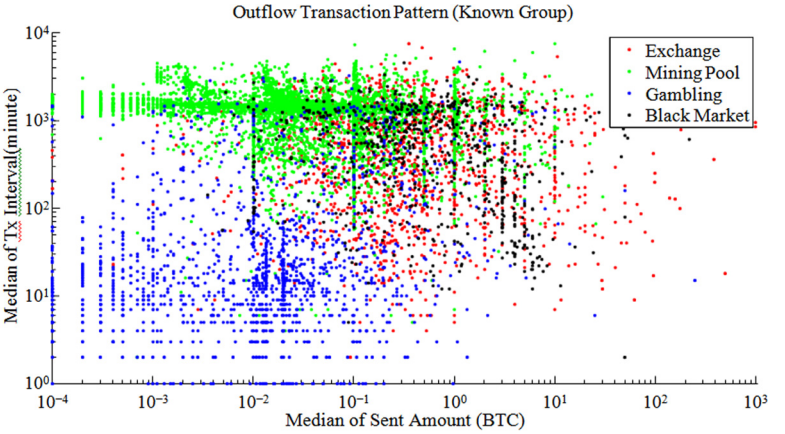
Only the mining pools show a significantly distinct TP from the others. Specifically, the outflow transactions for most of the mining pools are characterized by no more than ten inputs, but at the same time by a large amount of outputs, ranging from tens to thousands. This is consistent with the business model of mining pools: After successfully mining bitcoins, the mining pools will distribute the reward to all the small miners who have contributed some mining effort. So, the number of outputs is much larger than the number of inputs. One could speculate on the size of these mining pools according to the number of outputs in each outflow transaction.

As done for the inflow analysis, for each pair $(\hat{c}_x, c_y) \in \overleftarrow{W}^K$, we calculate the median transaction volume and the median time interval in minutes. The x-axis in Figure 7 represents the median transaction volume for each pair $(\hat{c}^x, c^y) \in \overleftarrow{W}^K$ while the y-axis represents the median time interval (in minutes) between outflow transactions for each pair $(\hat{c}^x, c^y) \in \overleftarrow{W}^K$. The blue dots scattered around the bottom-left area of the plot imply that gambling clusters send relatively small amounts of bitcoin but at a high-frequency to their counterparts[20].

The outflow transaction interval is plotted in Figure 8, which also shows the one-day effect for mining pools. The combination of the results from Figures 5 and 8 reveal a clear stylized fact characterized by many small miners receiving daily rewards from mining pools and then exchanging those rewards for fiat currency on exchange platforms.

6. Pure user group control test with transaction pattern analysis

In this section, we test the results of the PUG classification conducted in Section 4 for clusters in the unknown group by analyzing whether they exhibit pattern similarities with the clusters in the known group. The test is twofold and is based on the translation in matrix



Notes: Each dot characterizes one pair of clusters $(\hat{c}_x, c_y) \in \overleftarrow{W}^K$. The x-axis is the median transaction volume of all transactions between all the pairs of clusters $\in \overleftarrow{W}^K$ during the period January 2009-May 2015. The y-axis is the median transaction interval (in minutes) of the transactions between all the pairs of clusters $\in \overleftarrow{W}^K$

Figure 7.
Outflow TP for the known group

form of the inflow TPs and the outflow TPs involving super clusters in the known group. Each matrix is then compared, via a 2D correlation[21] analysis, with the correspondent one related to the inflow and outflow TPs involving super clusters in the unknown group.

To start, we translate the patterns depicted in Figure 3 into a matrix of transaction volumes and time intervals for all cluster pairs $(c_y, \hat{c}_x) \in {}^{-}\overrightarrow{W}K$ with $\hat{c}_x \in \hat{C}K$. We then create a matrix of transaction volumes and time intervals for all cluster pairs $(c_y, \hat{c}_x) \in {}^{-}\overrightarrow{W}U$ with $\hat{c}_x \in \hat{C}U$ where:

$$\overrightarrow{W}^{U} \subset W = \left\{ w_{yx}(c_y, \hat{c}_x) \in W \middle| c_y \in C \backslash \hat{C}^{U}, \hat{c}_x \in \hat{C}^{U}, n_{yx} \geq 100 \right\} \quad (16)$$

which means that a pair $(c_y, \hat{c}_x) \in \overrightarrow{W}^{U}$ is considered only if there has been at least 100 transactions from c_y to $\hat{c}_x$. For each pair $(c_y, \hat{c}_x) \in \overrightarrow{W}^{Ktt} \subset \overrightarrow{W}^{K} \overrightarrow{W}^{KB} \overrightarrow{W}^{K} \subset \overrightarrow{W}^{K}$ be the subsets of inflow transactions towards exchanges, gambling, and black markets in the known group, respectively. Similarly, let $\overrightarrow{W}^{U} X \subset {}^{-}\overrightarrow{W}^{U}$, $\overrightarrow{W}^{Ktt} \subset \overrightarrow{W}^{KU} \overrightarrow{W}^{UB} \subset \overrightarrow{W}^{U}$ be the subsets of inflow transactions toward exchanges, gambling and black markets in the unknown group, respectively.

Then, the 2D correlations of the inflow TPs between super clusters (in the known and unknown group) and clusters outside the groups are defined as follows: $corr2D(\overrightarrow{W}^{KX}, \overrightarrow{W}^{UX})$ is the 2D correlation between the inflow TPs for the exchanges in the known and unknown groups; $corr2D(\overrightarrow{W}^{Ktt}, \overrightarrow{W}^{Utt})$ is the 2D correlation between the inflow TPs for the gamblers in the known and unknown groups; $corr2D(\overrightarrow{W}^{KB}, \overrightarrow{W}^{UB})$ is the 2D correlation between the inflow TPs for the black markets in the known and unknown groups.

The correlation matrix in Table V shows that the classification of the super clusters according to the PUG analysis is consistent with the results of the TP analysis because the correlations along the main diagonal are greater than the values off-diagonal:

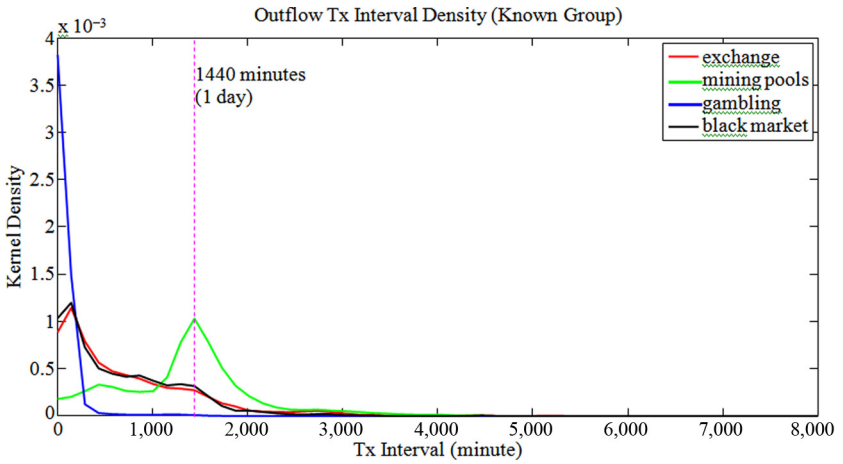


Figure 8.
Kernel density of the outflow amount of bitcoins sent to any $c_y \in C \backslash \hat{C}^{K}$ during the period January 2009-May 2015 by each cluster $\hat{c}_x \in \hat{C}^{k}$ in the known group

$$corr2D\left(\overrightarrow{W}^{KX}, \overrightarrow{W}^{UX}\right) > corr2D\left(\overrightarrow{W}^{KX}, \overrightarrow{W}^{Utt}\right), > corr2D\left(\overrightarrow{W}^{KX}, \overrightarrow{W}^{UB}\right)$$

and

$$corr2D\left(\overrightarrow{W}^{Ktt}, \overrightarrow{W}^{Utt}\right) > corr2D\left(\overrightarrow{W}^{Ktt}, \overrightarrow{W}^{UX}\right), > corr2D\left(\overrightarrow{W}^{Ktt}, \overrightarrow{W}^{UB}\right)$$

and

$$corr2D\left(\overrightarrow{W}^{KB}, \overrightarrow{W}^{UB}\right) > corr2D\left(\overrightarrow{W}^{KB}, \overrightarrow{W}^{UX}\right), > corr2D\left(\overrightarrow{W}^{KB}, \overrightarrow{W}^{Utt}\right)$$

Finally, by following a reverse approach than the one adopted to build the 2D correlation matrix for the inflow TPs, we calculate also the 2D correlation between pairs of outflow transactions involving clusters in the known and unknown group. Table VI shows that also in this case, the classification of the super clusters according to the PUG analysis is consistent with the results of the TP analysis because the correlations along the main diagonal are greater than the values off-diagonal.

7. The bitcoin network

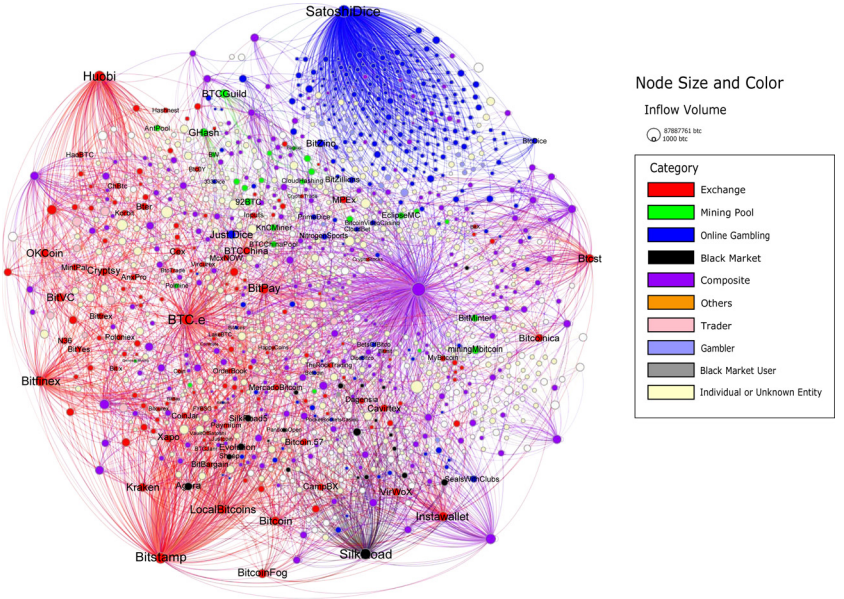
From the PUG analysis, we are able to classify some unknown super clusters into specific business categories. To illustrate the result, Figure 9 plots the payment network between the super clusters in $\hat{C}$ and their counterparts. For the sake of visualisation pourpose, two

Table V.
Correlation of category transaction (inflow) pattern between the known group and unknown group

Correlation matrix – inflow transaction volume/interval matrix		
$\overrightarrow{W}^{UX}$	$\overrightarrow{W}^{Utt}$	$\overrightarrow{W}^{UB}$
$\overrightarrow{W}^{KX}$ *0.8183*	0.2240	0.5090
$\overrightarrow{W}^{Ktt}$ −0.0412	*0.8943*	−0.0100
$\overrightarrow{W}^{KB}$ 0.6615	0.0389	*0.6665*

Table VI.
Correlation of category transaction (outflow) pattern between the known group and unknown group

Correlation matrix – outflow transaction volume/interval matrix		
$\overrightarrow{W}^{UX}$	$\overrightarrow{W}^{Utt}$	$\overrightarrow{W}^{UB}$
$\overrightarrow{W}^{UX}$ *0.4984*	−0.0864	0.3599
$\overrightarrow{W}^{Utt}$ 0.0378	*0.5933*	−0.0705
$\overrightarrow{W}^{UB}$ 0.4260	−0.0632	*0.4509*



Notes: An edge is traced as long as one party of the transaction belongs to our sample group. All clusters are colored according to the categories we explored. For clearer visualization, two thresholds are set here: First, transaction volume between each pair of nodes must be larger than 1,000 BTC. Second, the node's degree must be larger than 2

Figure 9.
Payment network between super clusters in and their counter parts

thresholds are set for plotting: First, we only plot for transactions (edges) with a volume larger than 1,000 BTC; second, the degree of the nodes must be larger than 2.

Figure 10 is a matrix of transactions between those super clusters in $\hat{C}$ ascribed to the major business categories (exchange, mining pool, online gambling, black market and composite). The *y*-axis depicts the sending clusters (grouped by business category) and the *x*-axis depicts the receiving clusters (also grouped by business category). There is no transaction volume limit for plotting this matrix; a dot is plotted as long as a *y*-axis super cluster has ever sent (even once) bitcoins to an *x*-axis super cluster, no matter what the transaction volume is. All the dots are colored according to the category to which the source belong to. For example, all the transactions sent from exchanges are signified by red dots.

We observe that mining pools typically only send coins to other categories, and do not receive any. We also observe that black markets tend to interact most with exchanges and composite services. A more comprehensive analysis of the results shown in Figure 10 is offered by the inflow dependency matrix in Table VII. Table VII(A) lists the bilateral transaction volume between all the pairs of business categories. The number in the cell (i, j) is the amount category i sent to j. For example, cell(6,1) = 6,003,342.66 tells us the category traders sent around six million bitcoins to exchanges. Table VII(B) calculates, for a given category, the percentage of bitcoins received from other categories. For example, cell(6,1) = 26.72 per cent shows us that the 26.72 per cent of the total inflow for the category exchanges comes from traders.

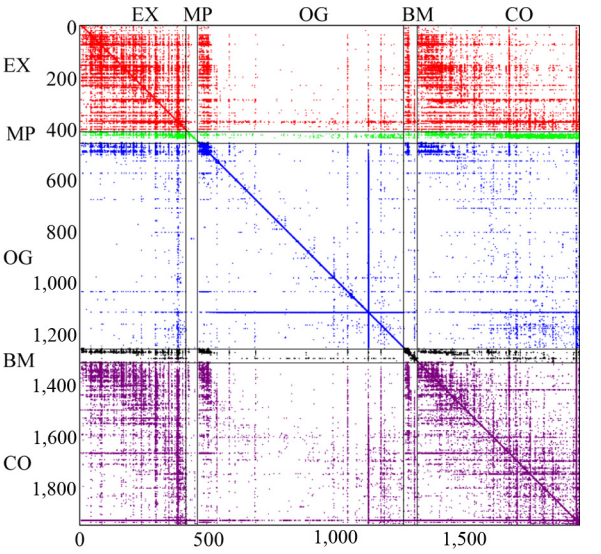


Figure 10.
Transaction matrix between super clusters in different business groups: "EX" for exchange, "MP" for mining Pool, "OG" for online gambling, "BM" for black market, "CO" for composite

Notes: The *y*-axis depicts the sending clusters (grouped by business category) and the *x*-axis depicts the receiving clusters (also grouped by business category). A dot is plotted if during the period January 2009-May 2015, a *y*-axis cluster has ever sent bitcoins to an *x*-axis cluster, no matter what the transaction volume is. All the dots are colored according to the category of source clusters, which is in line with the color rule used in Figure 9

7.1 Evolution of the bitcoin economy

Using the above analysis, we can measure the relative prevalence of each general business category (i.e. mining pool, exchange, online gambling and black market) in our sample, and track their evolution over the study period and visualize shifts in their relative centrality. Figure 11 shows the income inflow from January 2009 through May 2015.

We identify three distinguishable regimes that have occurred in the Bitcoin economy since its inception. The first period runs from approximately January 2009 through March 2012. This "proof-of-concept" period is characterized largely by test transactions among a small number of users, and with very little meaningful commercial activity[22]. Our analysis shows that this initial period is dominated almost entirely by mining, which is what we would expect from a system still devoid of material economic activity.

Next, from approximately April 2012 through October 2013 a second period consisting of "early adopters" appears. This period is characterized by a sudden influx of gambling services and "darknet" black markets; due to the overwhelming prevalence of these arguably nefarious categories, another name for this phase could be the period of "sin"[23]. These types of businesses initially responded to the unique features of Bitcoin such as its relative anonymity (pseudonimty), lack of regulatory and legal oversight, borderless transactions and low transaction costs absent from taxation. This new form of secure digital cash was ideal for the purchase and sale of illicit drugs, stolen items and other contraband

Table VII. Transaction relationship between categories

	Inflow dependency matrix							
	Exchange	Mining pool	Gambling	Black market (BM)	Composite	Trader	Gambler	BM user
A. Inflow transaction matrix (in volume)								
Exchange	12723006.59	5496.23	103072.23	377087.03	3026163.73	7311315.30	658.27	320.85
Mining pool	207146.47	152704.12	18717.09	1050.73	371430.37	1974.96	120.87	93.94
Gambling	66925.21	4440.94	20670014.00	14443.62	668594.19	5632.81	1493367.85	322.65
Black Market(BM)	384240.90	1573.13	30706.69	605954.69	534554.26	458.29	0.00	527171.73
Composite	3079138.63	19936.42	900872.93	858455.30	54573371.03	519874.09	145867.51	57825.19
Trader	6003342.66	2.61	1881.18	279.75	534557.53	0.00	0.00	0.00
Gambler	686.01	0.00	1296775.33	1093.50	210665.85	0.00	0.00	0.00
BM user	1492.33	0.00	2220.96	319168.90	76486.66	0.00	0.00	0.00
B. Inflow transaction matrix (in percentage)								
Exchange	56.63	2.98	0.44	17.31	5.04	93.26	0.04	0.05
Mining pool	0.92	82.92	0.08	0.05	0.61	0.03	0.01	0.02
Gambling	0.29	2.41	89.77	0.66	1.11	0.07	91.05	0.06
Black market(BM)	1.71	0.85	0.13	27.82	0.89	0.01	0.00	90.00
Composite	13.70	10.82	3.91	39.42	90.96	6.63	8.89	9.87
Trader	26.72	0.00	0.01	0.01	0.89	0.00	0.00	0.00
Gambler	0.00	0.00	5.63	0.05	0.35	0.00	0.00	0.00
BM user	0.01	0.00	0.01	14.65	0.12	0.00	0.00	0.00

Notes: The transaction flow in subtable(A) is from row to column. For example, cell(2,1) means mining pools send 207,146 BTC to exchanges. In subtable(B), the percentage is calculated column-wise, such that the figure reflects the inflow ratio for each category. For example, cell(2,1) = 0.92 tells us 0.92% of total income for exchanges is from mining pools

that could not be easily traded elsewhere online, or for gambling from a location where such a practice would be prohibited. Often, users of these "sin" sites would mask their internet traffic via services such as a virtual private network or via the TOR network, encouraging usage growth where the probability of being caught would be minimal (**?**). In fact, our data show that in the January of 2013, gambling and black markets together accounted for fully 51 per cent of all transactional inflows on the Bitcoin blockchain (in our sample).

Figure 12 shows the relative percentage of inflow transactions for each business category from January 2009 through May 2015.

The largest black market at the time was the Silk Road (Figure 9). That service was famously raided and shut down by the FBI in October of 2013, which could help explain the sudden drop in black market activity that brought this period to a close, although this event cannot satisfactorily explain the concurrent drop in gambling activity. The drop in gambling as a percentage of overall bitcoin transactions may have been due to the increase

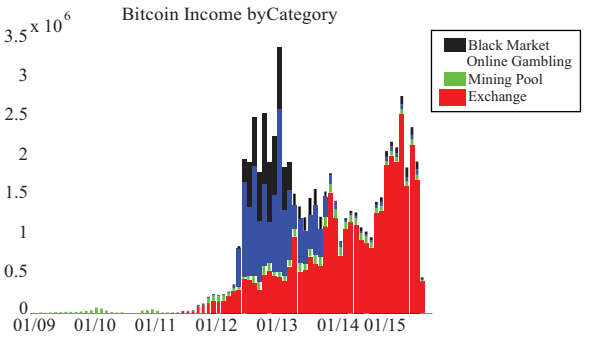
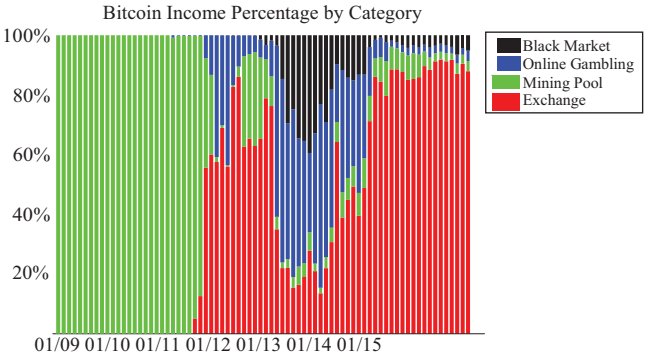


Figure 11.
Stacked plot of the inflow income amount for each business category in our sample (i.e. sum of the bitcoin inflows across all the super clusters in $\hat{C}$ belonging to the same major business category) over the bitcoin network, monthly from January 2009 through May 2015



Notes: Mining dominates initially, then "sin" categories (gambling in blue and black markets in black) rise, but recede over time in favor of exchanges

Figure 12.
Stacked plot of the *relative* income for each business categories as a percentage of total income inflows, monthly from January 2009 through May 2015

in value of one bitcoin, from a few tens of dollars to a few hundreds of dollars during this time. If a gambler tends to bet nominally $100 per day, what used to translate into some dozens of bitcoins instead became fractions of one bitcoin. Indeed, even though the relative amount of gambling has declined, the absolute amount wagered in dollar terms increased modestly over the study period. Still today, over 100 active gambling services currently exist that use bitcoin. It is also worth noting that while the overall amount of business being transacted on "sin" entities has fallen quite significantly, the actual number of black market sites available on the Bitcoin economy has nonetheless grown, with at least four reboots of the Silk Road, and no less than fifty other (now defunct) marketplaces established since January 2014. There were still a dozen or more such darknet marketplaces active around the time the study period ended in mid-2015 (Branwen, 2015).

Still, by November 2013, the amount of inflows attributable to "sin" entities had shrunk significantly to just 3 per cent or less of total transactions. This third period, which we are still currently experiencing, is characterized by a maturation of the Bitcoin economy away from "sin" enterprise and diversifying into legitimate payments, commerce and services. This claim is moreover supported by the ascendancy of the centrality of exchanges in the Bitcoin network. Figure 13 takes the sum of the monthly betweenness centralities of the super clusters in each business category, and it ranks them from January 2012 through May 2015[24]. Each cell is colored according to the category we have identified.

Since January 2014, we see red cells outnumber all the others in each column, which tells us that exchanges are the center of transaction activity.

When a licit merchant or service provider enters the Bitcoin economy and accepts bitcoin as payment, we expect that they will cash out on a steady basis in order to cover business costs and to reduce exposure to bitcoin's price volatility; in doing so, they require the regular use of exchanges. At the same time, investors and other users who see bitcoin as a financial asset would increasingly require exchanges. It is also around this time that external venture capital investment grew in support of Bitcoin-related start-ups and infrastructure, signaling further legitimizing. According to startups in the Bitcoin space raised almost $1bn in three years (Q1 2012-Q1 2015). In 2012, around $2m of VC money made its way to Bitcoin start-ups. In 2013 that number had grown to $95m, followed by $361.5m in 2014 and more than half a billion dollars in 2015.

Mining pools have stayed out of the spotlight in terms of our analysis of inflows and outflows due to the cap on how many bitcoins are created each day. This should not understate the significance of miners and their role in the Bitcoin economy. First, we would

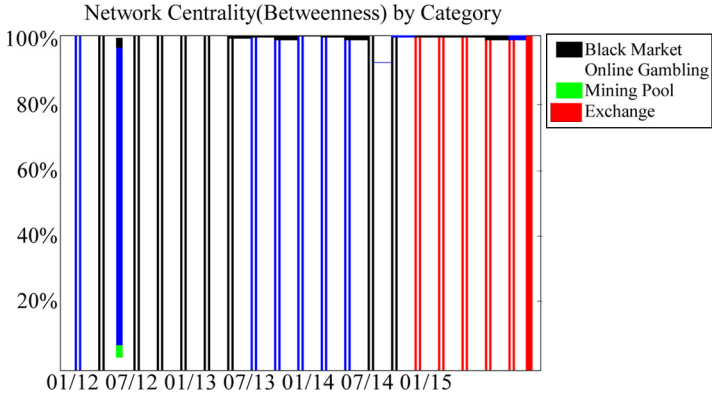


Figure 13. Monthly evolution, from January 2012 through May 2015, of the sum of the (betweenness) centrality measures across all the super clusters in $\mathcal{C}$ belonging to the same major business category in the bitcoin network

not expect mining pools to receive much in the way of income as those who join pools will only extract bitcoins away from pools and not send any *to* them. The pool generally earns income by taking a nominal percentage (1-2 per cent or less) of the block reward and/or by taking in the transaction fees associated with a found block. In terms of outflows, despite the amount of miners active on the network, the rate of unit formation for new bitcoins remains fixed at one block every ten minutes. At the beginning of our study period, the block reward was 50 BTC per block, from March 1, 2009 until November 28, 2012, so on any given day miners collectively produced just 7,200 BTC, a small fraction of total daily transaction volume. After November 28, 2012 and until approximately July 9, 2016, the block reward was reduced to 25 BTC, so that only 3,600 BTC were produced by miners daily, on average. From the July of 2016, the block reward is again reduced by half to 12.5 BTC, or just 1,800 BTC to be produced per day in aggregate through mining. Therefore, even if all participants of mining pools cashed out daily, their contribution to the overall network of payments will always be trivial, and, in fact, will decrease over time as the block reward continues to diminish. At the same time, however, the mining system serves a crucial function as it is the *de facto* "central bank" of the Bitcoin economy, expanding the money supply and validating every transaction. Without a robust and" honest" segment of miners, the fidelity of all other payments in the network would be suspect. In fact, a weak network of miners would leave the Bitcoin economy prone to a so-called 51 per cent attack, where a bad actor could begin to censor transactions by controlling a majority of power that validates transactions. Even though the relative centrality of miners is very small compared to the other business categories, the value they confer on to the network may instead be manifest via the price of bitcoin and miners' marginal profitability (Hayes, 2016).

7.2 Limitations and future directions

This study is not without some important limitations. First, it is obvious that the price of a bitcoin has risen wildly over the course of the study period and afterward, meaning that some of our stylized facts observed between economic actors may no longer apply. For example, the price of a $1,000 quantity of illegal contraband on a black market once corresponded to 50 BTC now commands, say, just 0.25 BTC. While this is certainly likely to diminish the absolute bitcoin volume of many transactions, price fluctuations ought to have no impact on the relative prevalence or centrality of category participants within the Bitcoin economy itself. Indeed, it is the relative predominance of economic activity that we are most concerned with and the pattern of transactions flows between actors. One important note on this, however, that could become an issue for future work is our use of a 1,000 BTC minimum activity threshold for identifying super clusters. As the price of a bitcoin rises further, the number of new entities receiving such large amounts is likely to diminish; therefore, this identification criterion probably ought to be relaxed.

Another concern could be the potentially biased use of coinjoin services by sin actors. One may reasonably argue that only "sinners" would need to use anonymization tools on their transactions, while those with nothing to hide would refrain from paying this added cost (albeit small). As we describe earlier in the paper, the presence of coinjoin mixing does not present the opportunity for false positives using our method; however, if such a bias for coin mixing exists we may be confronted with false negatives excluded from the analysis. If the use of coinjoin increased sharply at the transition to the maturation period, then this is a plausible alternative explanation. A recent in-depth analysis of coinjoin transactions conducted by Möser and Böhme (2017) show that the number of potential coinjoin transactions grew from zero per block prior to late 2013 to relatively stable values between 10 and 15 transactions per block in 2014 and through 2015. During the same period, the

number of total transactions per block grew from around 250 transactions to more than 500 in each block (Blockchain.info, 2017). This suggests the prevalence of coinjoin transactions was at most around 6 per cent of total transactions at the onset of the 2013 transition into the period of maturity and as a percentage of total transactions has fallen substantially since.

A third limitation to this preliminary study is that we do not provide confidence intervals or measure of statistical significance for our correlation matrices, upon which our PUG analysis relies. Our aim herein is meant principally to be descriptive in nature in order to reveal the network of economic clusters and their transaction partners. As we only use four major categories of economic activity (plus a fifth grouping of unknowns), it would be difficult to confuse, for example, a mining pool with a gambling service. However, future work that aims to clarify and complement these economic sectors will no doubt benefit from including robustness checks in terms of relating the statistical significance of correlation coefficients.

Finally, our analysis makes no attempt at stating causality for the transitions between one period to the next and makes no assertion that the period of maturity is destined to continue uninterrupted. While it is unlikely that the Bitcoin economy will revert to a renewed period of sin activity, it is possible that its legitimacy may be undermined for a host of other reasons, foreseen or otherwise. A theoretical foundation drawing on the fields of political economy, history of economics, and economic sociology may be able to provide some insights into these questions of causal effect and trajectory.

As cryptocurrency networks continue to grow in scale and scope, and as they become a more legitimate economic institution, analyses similar to this one ought to be carried out on other blockchains, for example Ethereum or Litecoin. Doing so can lay the groundwork for a comparative political economy of blockchain-based platforms.

8. Conclusions

As the Bitcoin economy grows in size and scope, it becomes increasingly important to better understand the key components and players in that system. However, this task has largely proven cumbersome as many tens of millions of individual addresses exist, which are not obviously linked to any specific individual or business entity and simply represent nondescript public keys in a public-private key pair.

In this paper, we begin to unveil the composition and trajectory of the Bitcoin political economy by analyzing a database composed of millions of individual Bitcoin addresses that we refine down to 2,850 super clusters, each comprised more than 100 addresses and having received at least 1,000 BTC from January 2009 to May 2015. A super cluster is described as an approximation of a business entity in that it describes a number of individual bitcoin addresses that are owned or controlled collectively by the same beneficial owner for some particular economic purpose. These important clusters are, for the most part, initially unknown and uncategorized. However, we can ascribe most of them to one of four specific business categories – mining pools, exchanges, online gambling sites and black markets – by mapping and analyzing the network of actors and pattern of payments among those and a smaller known set of clusters. In particular, we achieve this mapping using a PUG analysis that examines inflows and outflows to and from each cluster, as well as TP analysis to confirm those findings. Our method of de-anonymizing otherwise pseudonymous clusters allows us to not only visualize the Bitcoin network of payments but also to extract stylized facts that describe its internal economy.

We find that there are, in fact, distinct patterns of transaction flows for actors in each business category. For example, flows between traders and exchanges averaged just around 20 BTC over the study period, and traders bought or sold on average every

11 days. Meanwhile, gamblers wagered just 0.5 BTC on average, but re-bet often within the same day. There seems to be a strong preference to do business within the Bitcoin economy in round lot amounts (e.g. 0.1, 0.2, 0.5, 1.0 BTC), whether it is traders exchanging for fiat money, gamblers placing bets or black market goods being bought and sold.

In terms of transaction interval, there is an observable one-day effect for each business category during the study period. For instance, a pattern emerges that many (pooled) miners accumulate mining rewards and subsequently sell those bitcoins on exchanges on a daily basis. This is interesting, as it could suggest most miners are operating to sell their product each day for-profit and are not mining to accumulate and hoard bitcoins for the long term. Whether this observation has any bearing on the price of bitcoin is open to further study. Transaction flows from miners in our sample, however, are a relatively small fraction of total volume compared to the rest of the Bitcoin economy, as miners in aggregate were only able to produce no more than either 3,600 (or 7,200 BTC) per day on average over the study period with a block reward of either 25 (or 50 BTC prior to the block reward halving) due to the limitation of the Bitcoin protocol that enforces a controlled rate of new unit formation at one block every 10 mins.

Notes

1. By convention, we use Bitcoin with a capital "B" to denote the protocol, network, and community, while bitcoin with a small "b" denotes the digital currency and units of that currency.
2. Refer Blockchain.info (2017).
3. In principle, a single entity may have control over more than one distinct super cluster if the common ownership of some of their addresses is not evident from the data.
4. As well as other unknown individuals
5. A Bitcoin address is an identifier of 26-35 alphanumeric characters that is derived from the public key through the use of one-way cryptographic hashing. The algorithms used to make a Bitcoin address from a public key are the Secure Hash Algorithm (SHA) and the RACE Integrity Primitives Evaluation Message Digest (RIPEMD), specifically SHA256 and RIPEMD160, refer Antonopoulos (2014).
6. In the Bitcoin network, the output of a transaction is used as the input of another transaction. If the input is larger than the new transaction output the client generates a new Bitcoin address, and sends the difference back to this address. This is known as change. From the Bitcoin wiki: Take the case of the individual transaction 0a1c0b1ec0ac55a45b1555202daf2e08419648096f5bcc4267898d420dffef87, where a previously unspent output of 10.89 BTC was spent by the client. 10 BTC was the actual payment amount, and 0.89 BTC was the amount of change returned. The client cannot spend just 10.00 BTC out of a 10.89 BTC payment anymore than a person can spend 1 *out of a* 20 bill. The entire 10.89 BTC unspent output became the input of this new transaction and in the process produced are two new unspent outputs which have a combined value of 10.89 BTC. The 10.89 BTC is now "spent" and effectively destroyed because the network will prevent it from ever being spent again. Those unspent outputs can now become inputs for future transactions.
7. This is of course a very time consuming and inefficient activity, e.g. Meiklejohn *et al.* (2013) by participating in 344 transactions was able to manually tag 1,017 addresses.
8. A false positive exists when an address is wrongly included in a cluster (i.e. all addresses are not controlled by the same entity), and a false negative when an address should be in a cluster but is not.

9. We elaborate on the rationale behind these filtering criteria later in the paper.
10. Indeed, fork-merge patterns and self-loops represent a frequent scenario in the Bitcoin economy, *e.g.***?** and **?**.
11. Bitcoin Core was, by far, the most dominant version of the Bitcoin blockchain over the study period.
12. For example, p_{Huobi} is the set of addresses associated to Huobi with $n(p_{Huobi}) = 37{,}756$.
13. See Figure A2 in Appendix 1 for a visualization of the problem we aim at solving.
14. As an example, one of the biggest clusters holding about 6 million addresses which probably should have been included in $\hat{C}^K$ is instead included in $\hat{C}^U$ because although it has 2 million addresses linked to the MtGox exchange, it has one address linked to bitcoin 24.
15. Super clusters linked to the same real-world entity are merged into one node in the network.
16. For example, if the median value of intervals is 60 min, this means that counterparts tends to send to $\hat{c}_x$ bitcoins every 60 min.
17. In order to capture exact density on point, a very precise width is needed. In this case, we set the width 0.00000001 BTC (or 1 satoshi).
18. This is true, for example, in SatoshiDice, the largest bitcoin-based gambling service.
19. This practice is common in transactions involving small amounts of street drugs where a "dime bag" is whatever quantity $10 buys and a "nickel bag" whatever $5 buys.
20. This feature is consistent with the results in the former inflow analysis.
21. For deeper insight into the detail algorithm please see **?**.
22. One notable exception is on the 22 May 2010 in a purchase made by Laszlo Hanyecz, a software developer who paid a fellow BitcoinTalk online forum user 10,000 BTC for two Papa John's pizzas. At today's prices that is the equivalent of $2.25m per pizza!
23. The authors use the terminology "sin" colloquially for illustrative purposes only, and do not attribute any moral or ethical judgment to the word in the context of this paper.
24. For consistency, we also check other centrality measures like weighted degree and closeness, but the result does not change. We refer the reader to **?** for more details on network centrality measures.

References

Androulaki, E., Karame, G.O., Roeschlin, M., Scherer, T. and Capkun, S. (2013), "Evaluating user privacy in bitcoin", *Financial Cryptography and Data Security*, Springer, pp. 34-51.

Antonopoulos, A.M. (2014), *Mastering Bitcoin: Unlocking Digital Cryptocurrencies*, O'Reilly Media.

Blockchain.info (2017), available at: http://blockchain.info (accessed 21 August 2017).

Branwen, G. (2015), "Black market risks", available at: www.gwern.net/Black-market%20survival (accessed 15 June 2016).

Doll, A., Chagani, S., Kranch, M. and Murti, V. (2014), Btctrackr: Finding and display- ing clusters in bitcoin.

Garcia, D., Tessone, C.J., Mavrodiev, P. and Perony, N. (2014), "The digital traces of bubbles: Feedback cycles between socio-economic signals in the bitcoin economy", *Journal of the Royal Society Interface*, Vol. 11 No. 99, pp. 20140620-20140623.

Hayes, A. (2016), "Cryptocurrency value formation: an empirical study leading to a cost of production model for valuing Bitcoin", Telematics & Informatics.

Kaminsky, D. (2011), "Black ops of tcp/ip", Black Hat USA.

Koshy, P., Koshy, D. and McDaniel, P. (2014), "An analysis of anonymity in bitcoin using p2p network traffic", International Conference on Financial Cryptography and Data Security, *Springer*, pp. 469-485.

Kristov Atlas (2015), "Weak privacy guarantees for SharedCoin Mixing service", available at: www.coinjoinsudoku.com/advisory (accessed 1 June 2015).

Lischke, M. and Fabian, B. (2016), "Analyzing the bitcoin network: the fi four years", *Future Internet*, Vol. 8 No. 1, p. 7.

Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G.M. and Savage, S. (2013), "A fistful of bitcoins: characterizing payments among men with no names", *Proceedings of the 2013 Conference on Internet Measurement Conference, ACM*, pp. 127-140.

Möser, M. and Böhme, R. (2017), "Anonymous alone? Measuring Bitcoin's second-generation anonymization techniques", *IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE*, pp. 32-41.

Nakamoto, S. (2008), "Bitcoin: a peer-to-peer electronic cash system", *Consulted*, Vol. 1, pp. 2012.

Reid, F. and Harrigan, M. (2013), *An Analysis of Anonymity in the Bitcoin System*, Springer.

Ron, D. and Shamir, A. (2013), Quantitative analysis of the full bitcoin transaction graph, Springer, *Financial Cryptography and Data Security*, pp. 6-24.

Spagnuolo, M. (2013), "Bitiodine: extracting intelligence from the bitcoin network".

Tasca, P. (2015), "Digital currencies: principles, trends, opportunities, and risks", ECUREX Research WP, 7 September.

Further reading

Barton, F., Himmelsbach, D., Duckworth, J. and Smith, M. (1992), "Two-dimensional vibration spectroscopy: correlation of mid-and near-infrared regions", *Applied Spectroscopy*, Vol. 46 No. 3, pp. 420-429.

Blockchain.info (2015), available at: http://blockchain.info (accessed 1 June 2015).

Cormen, T.H., Leiserson, C.E., Rivest, R.L. and Stein, C. (2009), "Data structures for disjoint sets", *Introduction to Algorithms*, pp. 561-585.

Dingledine, E.A. (2004), *The Second-Generation Onion Router*, Naval Research Lab.

Hanneman, R. and Riddle, M. (2014), *Introduction to Social Network Methods*, University of California, Riverside, available at: http://faculty.ucr.edu/hanneman/nettext

Walletexplorer (2015), available at: www.walletexplorer.com/ (accessed 1 June 2015).

Corresponding author

Paolo Tasca can be contacted at: p.tasca@ucl.ac.uk

Appendix 1

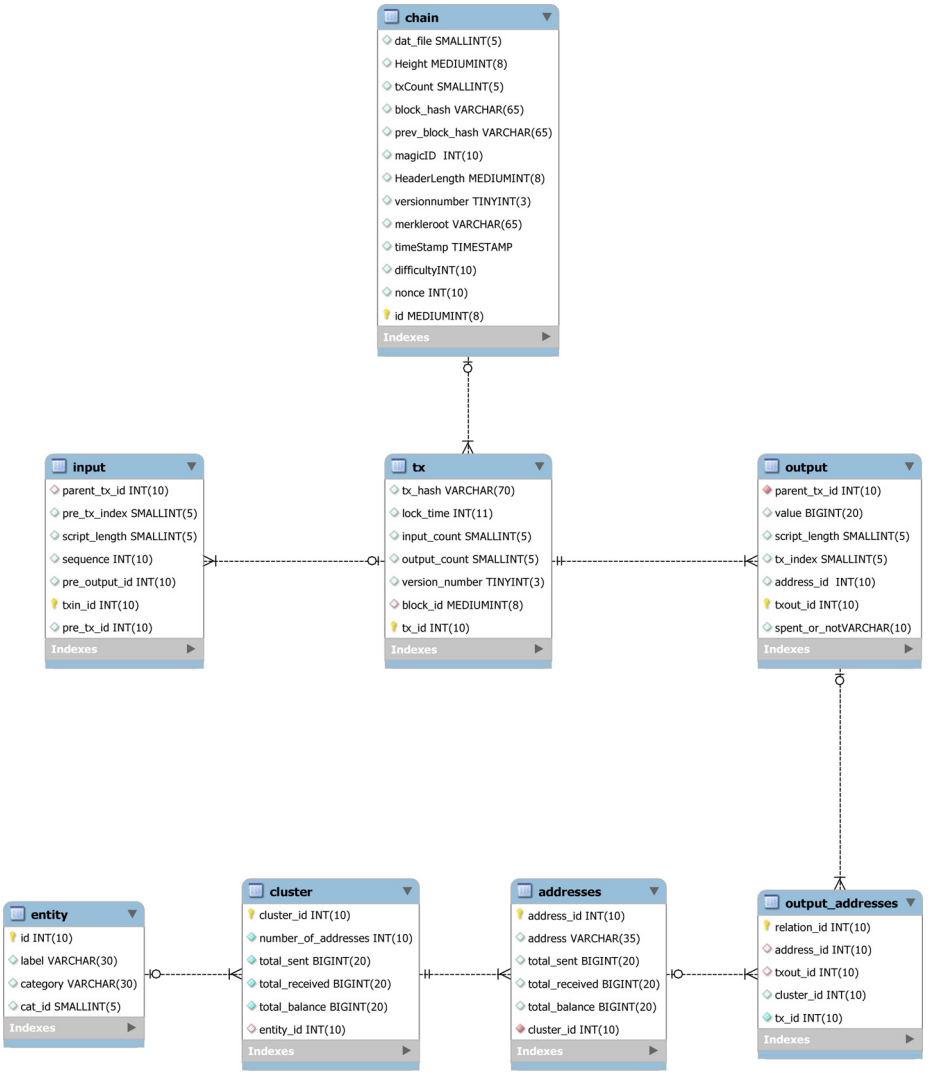


Figure A1. Structure of the MySQL database created from the data described in Table I

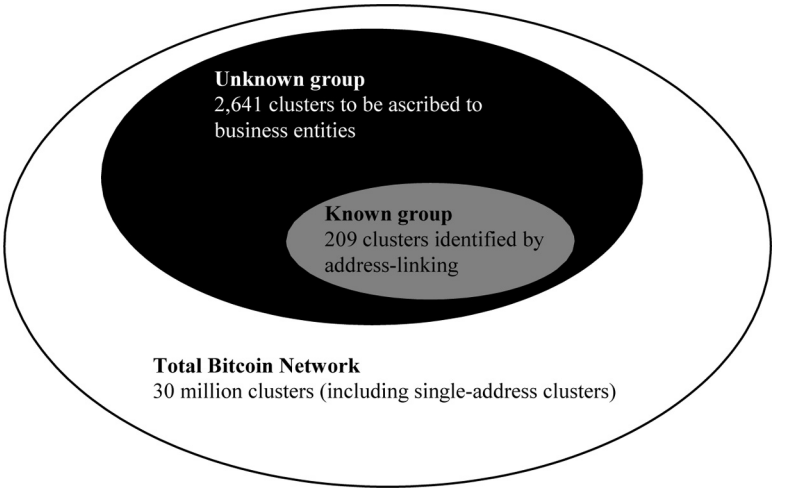


Figure A2.
The total number of clusters in the bitcoin network is about 30 millionNotes: Our research focuses on 2,850 large clusters that include at least 100 addresses and that also have received at least 1,000 bitcoins from January 3, 2009 to May 8, 2015 (Known Group + Unknown Group)

Appendix 2. List of identified super clusters in the set $\hat{C}^K$

Exchange									
Cluster ID	Entity name								
414294	VirWoX	24850005	CoinSpot	45437770	BitYes	9549829	BitMillions	25855276	Evolution
725951	Cavirtex	25671458	CoinTrader	45936035	Huobi	10352795	Betcoins	28893773	BlackBank
1152538	CampBX	25686021	Poloniex	46523945	CleverCoin	11592006	BTCOracle	32188483	CannabisRoad
1477742	MercadoBitcoin	25764718	LiteBit	47947669	BtcTrade	12554372	Coinroll	34450135	PandoraOpen
1538322	BTCChina	26526196	AllCoin	48224085	BitVC	14592351	Just-Dice	39307422	MiddleEarth
1591210	Bitcash	26768188	VaultOfSatoshi	49491730	Coinmate	15935043	BITcoinBa	53431789	Nucleus
1640270	BTC-e	27058962	MintPal	49497346	LocalBitcoins	17858445	YABTCL	58138309	Abraxas
1745068	Bitstamp	27336957	Indacoin	51159703	Bter	18104553	BitZillions		
1872280	Bitcoin	27430132	C-Cex	52772381	ChBtc	18242583	Ice-Dice		**Others**
2403973	TheRockTrading	27516236	1Coin	63125407	BTCChina	18844372	SatoshiRoulette	**Cluster ID**	**Entity name**
3160452	OrderBook	27883925	Bittrex	64576584	Exmo	19074264	Peerbet	51591631	HaoBTC
3169372	BitBargain	28195895	Paymium	64714050	HitBtc	20709464	Betcoin	2481605	BitcoinFog
5128017	LocalBitcoins	29314526	AllCrypt	67252210	Bter	21368294	AnoniBet	38948179	BitLaunder
5946497	HappyCoins	29907632	Dgex	74399779	MtGox	22181037	NitrogenSports	17685858	CryptoLocker
6299268	Cryptonit	30131409	CoinMotion			22815568	CoinGaming	1400957	MPEx
6606601	MtGox	30139877	Bter		**Mining pools**	23210454	SatoshiBet	2062018	Bitcoinica
6960785	Bitfinex	30804162	CoinArch	**Cluster ID**	**Entity name**	24545072	999Dice	3165186	Bitcoinica
7522909	Bitcoin-24	30852641	BTCChina	177451	Eligius	24857474	BitcoinVideoCasino	32965397	UpDown
8058186	Justcoin	31778769	Coin-Swap	2400970	mining.bitcoin	26783278	PocketRocketsCasino	1406234	Btcst
8764670	FYBSG	32344865	BitBay	2440660	BitMinter	28382823	BitAces	17144983	Purse
11025414	BitX	32394318	Bter	4886325	EclipseMC	32814149	BitStarz	21601241	Bylls
11196419	SmenarnaBitcoin	33419156	CoinCafe	5272039	GHash	33495508	Betcoin	14543862	Bitbond
11749226	Cryptorush	34085743	BX	7530073	BTCGuild	37042731	CloudBet	36933042	BTCJam
12637441	McxNOW	34277949	BtcExchange	8388553	50BTC	38624871	PrimeDice	39317993	BitLendingClub
12797521	Korbit	35226292	MeXBT	11551066	50BTC	39363482	DiceNow	17815289	BTCt
13228368	Vircurex	35431781	Zyado	12547187	mining.bitcoin	41129839	DiceBitco	1075785	BitPay
13539065	Crypto-Trade	35636277	QuadrigaCX	13455133	KnCMiner	43427199	PrimeDice	16248472	CoinPayments
13549778	Cryptsy	36674288	MaiCoin	18761724	CloudHashing	44125199	SatoshiMines	65645195	BitPay
13549778	Cryptsy	36837273	HitBtc	21224287	BTCChinaPool	45607266	FortuneJack	1582623	Bitmit
14777694	Coins-e	37013580	Matbea	23855294	Polmine	48934666	SecondsTrade	13255854	CryptoStocks
14833131	AnxPro	37776533	Btc38	34581906	Genesis-Mining	50523669	Betcoin	454407	Instawallet
15004560	BitKonan	38951758	Ccedk	45656162	AntPool	52248120	SatoshiDice	869503	MyBitcoin
16030982	OKCoin	39963036	796	48150806	mining.bitcoin	57476416	BitcoinVideoCasino	8341192	Dagensia
17494455	Huobi	40161739	LakeBTC	58048160	AntPool	58900551	PrimeDice	14011339	CoinJar
		41193900	Bitso	61166475	BW	64148592	BitAces	14359270	Xapo

(*continued*)

Table AI.
List of the super clusters in C^K . One entity could own and control more than one cluster

Table AI.

17518823	CoinMkt	41323542	SpectroCoin			65420930	SwCPoker	14773742	Inputs
17747783	Kraken	41433875	OKCoin	**Gambling**		73161189	PrimeDice	31631652	BitcoinWallet
18055670	Cex	41555907	BTC-e	**Cluster ID**	**Entity name**			51620287	OkLink
18847146	BtcMarkets	41614840	BTC-e	184867	Just-Dice		**Black Market**	59825409	GoCelery
19681395	Bitcoin	41923963	Hashnest	1687007	BetsOfBitco	**Cluster ID**	**Entity name**		
20789150	Coinomat	42369494	Cryptsy	2254800	SealsWithClubs	4401158	SilkRoad		
21373812	Bleutrade	42879690	C-Cex	3486952	SatoshiDice	9563241	Sheep		
21653414	Bitfinex	43277175	Bit-x	4169604	BitZino	19517829	PandoraOpen		
23421684	Coin	43970673	Bter	4831753	BtcDice	20627442	SilkRoad2		
23672561	Masterxchange	43974172	Bter	8339663	BitElfin	22735225	Agora		
24089310	Igot	45333046	Bitcurex	9510403	Playt	22917766	BlueSky		

Notes: The cluster IDs are generated internally from MySQL database, and each cluster has one unique cluster ID. Entities are classified according to their business objective. We focus on the biggest four categories (exchange, mining pool, gambling, black market). Entities with exposure to more than one category, such as HaoBTC (both wallet and mining pools) are categorized as "composite"