Bitcoin Transaction Networks: an overview of recent results

Nicoló Vallarano,[1] Claudio Tessone,[2,3,*] and Tiziano Squartini[1]

[1] *IMT School for Advanced Studies, Piazza S.Francesco 19, 55100 Lucca (Italy)*
[2] *UZH Blockchain Center, University of Zürich, Rämistrasse 71, 8006 Zürich (Switzerland)*
[3] *URPP Social Networks, University of Zürich, Andreasstrasse 15, 8050 Zürich (Switzerland)*

(Dated: May 4, 2020)

Cryptocurrencies are distributed systems that allow exchanges of native (and non-) tokens among participants. The complete historical bookkeeping and its wide availability opens up an unprecedented possibility, i.e., that of understanding the evolution of their network structure while gaining useful insight on the relationships between user behaviour and cryptocurrency pricing in exchange markets. In this contribution we review some of the most recent results concerning the structural properties of *Bitcoin Transaction Networks*, a generic name referring to a set of different constructs: the *Bitcoin Address Network*, the *Bitcoin User Network* and the *Bitcoin Lightning Network*. A common picture that emerges out of analysing them all is that of a system growing over time, which becomes increasingly sparse, and whose structural organization at the mesoscopic lebevel is characterised by the presence of a statistically-significant *core-periphery* structure. Such a peculiar topology is matched by a highly unequal distribution of bitcoins, a result suggesting that Bitcoin is becoming an increasingly *centralised* system at different levels.

INTRODUCTION

A cryptocurrency is an online payment system for which the storage and the verification of transactions - therefore, the safeguard of the system consistency itself - are *decentralised*, i.e. do not require the presence of a trusted third party. This result can be achieved by securing financial transactions through a clever combination of cryptographic technologies [1].

Bitcoin, the first and most popular cryptocurrency, was introduced in 2008 by Satoshi Nakamoto [2]: it consists of a decentralised peer-to-peer network to which users connect to exchange the property of the account units of the system, i.e. *perform bitcoins transactions*. Each transaction becomes part of a publicly available ledger, the *blockchain*, after having been validated by the so-called *miners*, i.e. users that verify the validity of issued transactions according to the consensus rules that are part of the Bitcoin protocol [3, 4]. A new block, containing transactions that where known to the miner since the last block, is 'mined' - on average - every 10 minutes, thereby adding new transactions to the blockchain. Thus, these transactions are 'confirmed', in turn enabling users to spend the bitcoins they received through them[1]. The cryptography protocols Bitcoin rests upon aim at preventing the so-called *double-spending problem*, i.e. the possibility for the same digital token to be spent more than once in absence of a central party that guarantees the validity of the transactions [1, 2]: remarkably, the transaction-verification mechanism Bitcoin relies on allows its entire transaction history to be openly accessible, a feature that, in turn, allows researchers to analyse it in different network representation.

The gain in popularity of Bitcoin led its community to face new problems such as the lack of *scalability* of the transaction-verification method: The first one concerns the (relatively low) maximum number of transactions that can be verified per second - especially if compared with the mainstream competitors such as centralised payment networks. The large *concentration of mining power* in mining pools (implying a decrease in decentralisation of verification in the network) and the tendency of users to *hoard*. In order to solve the aforementioned problems, threatening the overall functioning of Bitcoin as a medium of exchange, new instruments were adopted. Proposed in 2015 [5], the *Bitcoin Lightning Network* (BLN) is a 'Layer 2' protocol that can operate on top of Blockchain-based (Bitcoin-like) cryptocurrencies by creating bilateral channels for *off-chain* payments which are, then, settled concurrently on the blockchain once the channel gets closed. As both the transaction fees and the blockchain confirmation are no longer required, the network is spared from avoidable burden; moreover, the key features of Bitcoin, i.e. its *decentralised architecture*, its *political organisation* and its *wealth distribution* are no longer sacrificed, while the circulation of the native assets is enhanced.

Although Bitcoin is almost ten years old, researchers have started to investigate the Bitcoin structural properties only recently: in [6], the authors consider the network of transactions between addresses at the weekly time scale,

* claudio.tessone@business.uzh.ch

[1] In practice, the so-called '6 confirmations' rule is followed: once a transaction is included in a block which is followed by at least six additional blocks [33], the transaction can be safely considered as confirmed.

showing the emergence of power-law distributions and that the number of incoming transactions reflects the wealth of nodes; in [7], the authors consider the network of transactions between users at the macroscale, in order to check for its small-worldness; in [8], the authors investigate the network of international Bitcoin flows, identifying socio-economic factors that drive its adoption across countries. Several studies also focus on the problem of Bitcoin users de-anonymisation [9–13] while others analyse the interplay between social interactions and the movements of the Bitcoin price [14, 15].

With the present work, we aim at summing up the results of three papers [16–18]. In [16], the authors analyse the local properties of two different Bitcoin representations, i.e. the *Bitcoin Address Network* (BAN) and the *Bitcoin User Network* (BUN) and inspect the presence of correlations between (exogenous) price movements and (endogenous) changes in the topological structure of the aforementioned networks. In [17], the mesocale structure of the BUN is under scrutiny: particular attention is devoted to the identification of the best network model able to describe it; besides, the same exercise as above is carried out, i.e. the comparison between the evolution of purely structural properties and the appearance of price bubbles in a cyclical fashion. Lastly, in [18], the authors inspect the evolution of the BLN topology, pointing out that it is becoming an increasingly centralised system and that the 'capital' is becoming increasingly unevenly distributed.

DATA

As previously said, Bitcoin relies on a decentralised public ledger, the blockchain, that records all transactions among Bitcoin users. A transaction is a set of input and output addresses: the output addresses that are 'unspent', i.e. not yet recorded on the ledger as input addresses, can be claimed, and therefore spent, only by the owner of the corresponding cryptographic key. This is the reason why one speaks of *pseudonimity*: an observer of the blockchain can see all unspent *addresses* but cannot link them to the actual owners.

The Bitcoin Address Network (BAN)

The BAN is the simplest network that can be constructed from the blockchain records: from a technical point of view, it is a directed, weighted graph whose nodes represent addresses; the direction and the weight of links are provided by the input-output relationships defining the transactions recorded on the blockchain. The BAN has been considered across a period of 9 years, i.e. from 9th Janury 2009 to 18th December 2017 at the end of which the data set consists of 304.111.529 addresses, exchanging a total number of transactions amounting at 283.028.575. In terms of traded volume, the transactions between addresses amount at 4.432.597.496 bitcoins.

The Bitcoin User Network (BUN)

Since the same owner may control several addresses [11], one can define a network of 'users' whose nodes are *clusters of addresses*. These clusters are derived by implementing different *heuristics*, provided by the state-of-the-art literature [9, 10, 19, 20]. The 'user networks' we obtain should not be considered as a perfect representation of the actual networks of users but, rather, an attempt to group addresses while minimising the presence of false positives. The BUN has been considered across the same period of the BAN (i.e. of 9 years, from 9th Janury 2009 to 18th December 2017) at the end of which the data set consists of 16.749.939 users, exchanging a total number of transactions amounting at 224.620.265. In terms of traded volume, the transactions between users amount at 3.114.359.679 bitcoins.

The Bitcoin Lightning Network (BLN)

The BLN is constructed in a fashion that is similar to way the BAN is defined: it is a directed, weighted graph whose nodes are the addresses exchanging bitcoins on the 'Layer 2'. Three different representations of the BLN have been studied so far, i.e. the daily one, the weekly one and the daily-block one: while a daily/weekly snapshot includes all channels that were found to be active during that day/week, a daily-block snapshot consists of all channels that were found to be active at the time the first block of the day was released (hence, the transactions considered for the daily-block representation are a subset of the ones constituting the daily representation). The BLN was considered across a period of 18 months, i.e. from 14th January 2018 to 13th July 2019, at the end of which the network consists of 8.216 users, 122.517 active channels and 2.732,5 transacted bitcoins.

Notation

Although the information about the magnitude of transactions is available, the BAN and the BUN have been analysed as binary, directed networks; as such, they are completely specified by their binary, asymmetric adjacency matrices $\mathbf{A}^{(t)}_{\mathrm{BAN}}$ and $\mathbf{A}^{(t)}_{\mathrm{BUN}}$, at time t. The generic entry $a^{(t)}_{ij}$ is equal to 1 if at least one transaction between address (user) i and address (user) j takes place, i.e. bitcoins are transferred from address (user) i to address (user) j, during the time snapshot t and 0 otherwise. The BLN, on the other hand, is a weighted, undirected network, represented by a symmetric matrix $\mathbf{W}^{(t)}_{\mathrm{BLN}}$ whose generic entry $w^{(t)}_{ij} = w^{(t)}_{ji}$ indicates the total amount of money exchanged between i and j, across all channels, at time t; here, we will mainly focus on its binary projection $\mathbf{B}^{(t)}_{\mathrm{BLN}}$, whose generic entry reads $b^{(t)}_{ij} = b^{(t)}_{ji} = 1$ if $w^{(t)}_{ij} = w^{(t)}_{ji} > 0$ and $b^{(t)}_{ij} = b^{(t)}_{ji} = 0$ otherwise.

RESULTS

The Bitcoin Address and User networks

Let us start by reviewing the results concerning the BAN and the BUN at the weekly time scale. Similar results are observed for the BAN and the BUN at the daily time scale [16].

Basic statistics. Let us start by commenting one the evolution of some basic statistics characterising the BAN and the BUN. As fig. 1 shows, both the number of nodes N and the number of links $L = \sum_i \sum_{j(\neq i)} a_{ij}$ increase steadily in time, irrespectively from the considered representation; the link density $d = \frac{L}{N(N-1)}$, however, decreases, meaning that the system becomes sparser. The dependence of d from N can be better specified from a mathematical point of view, upon noticing that the average degree $\overline{k^{in}} = \overline{k^{out}} = \frac{\sum_i \sum_{j(\neq i)} a_{ij}}{L} = \frac{L}{N}$ is constant over time [16]; hence, it follows that $L \propto N$ and $d \sim N^{-1}$. As a general comment, notice how the basic statistics have started to evolve in a more stationary fashion since middle 2011 (see also [12]).

Degree distributions. Generally speaking, both out- and in-degrees are characterised by heavy-tailed distributions, indicating that a large number of low-connected nodes co-exists with few hubs whose degree is several order of magnitudes larger. A visual inspection of the functional form of the degrees distributions suggests the latter ones to follow a power-law [16, 21]; to test this hypothesis the authors in [16] employed an algorithm based on a double Kolmogorov-Smirnov statistical test [22**?**]: what emerges is that the hypothesis above cannot be rejected, at a 0.05 confidence level, for almost half of the considered snapshots.

Of particular interest is the evolution of the out-degrees standard deviation, especially for what concerns its informativeness about exogenous events. As an example, let us consider the failure in February 2014 of Mt. Gox, a quasi-monopolist exchange market at the time. Such an event deeply affected the overall Bitcoin structure: the percentage of snapshots for which the null hypothesis (i.e. the out-degrees distribution follows a power-law) can be rejected amounts at $\simeq 50\%$ *before* February 2014 while it drops to $\simeq 25\%$ *afterwards*.

The authors in [16] also argue that the presence of heavy-tailed distributions may be explained by a mechanism similar to the preferential attachment one: new, or occasional, users 'preferentially' connect to already well-connected nodes (exchange markets, utility providers, etc.), thus leading to the formation of super-hubs. Elsewhere it has been argued that the related mechanism known as 'fittest-gets-richer' or 'good-gets-righer' [23] may be also at work, the computational resources of a node playing the role of its fitness [7].

Bitcoin structure VS Bitcoin price. The result concerning the evolution of the out-degrees distribution suggests that the Bitcoin network structure indeed brings the signature of exogenous events. Since Bitcoin is a cryptocurrency, a natural step is that of studying the presence of correlations between purely topological quantities and its *price*.

The simplest analysis to carry out is that of scattering the network size and the network link density versus the Bitcoin price (in USD). As shown in fig. 2, a clear trend appears, indicating that the price and the size N (the link density d) are overall positively (negatively) correlated throughout the entire Bitcoin history. Notice, however, the trend inversion that can be appreciated immediately after the Mt. Gox failure: it is a consequence of the prolonged price decrease observed in 2014-2016, during which the network size has increased of (almost) one order of magnitude.

To further confirm the presence of a double regime, the authors in [16] have inspected the correlation between the moments of out-degrees distribution and the Bitcoin price over time. To this aim, the so-called Price and its Moving Average (RPMA) indicator has been adopted, defined as

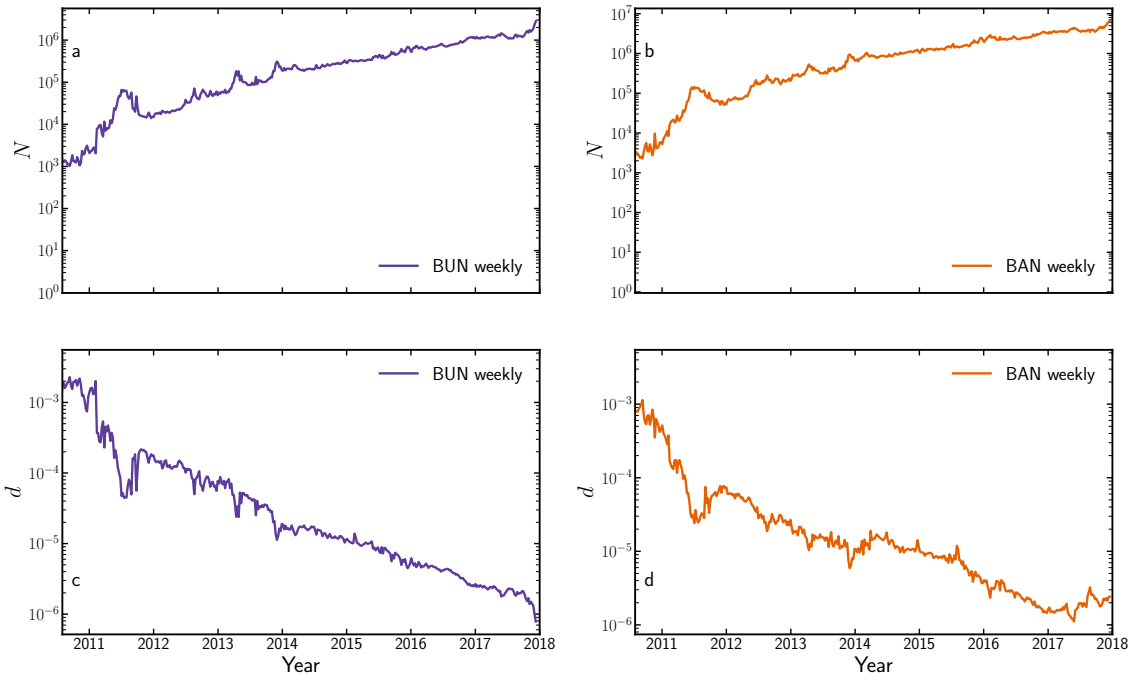

FIG. 1. Evolution of basic statistics, i.e. the number of nodes (left panels) and the link density (right panels) for two Bitcoin network representations, i.e. the BAN (bottom panels) and the BUN (top panels) at the weekly time scale, from July 2010 to 18th December 2017 (i.e. for networks with at least 200 nodes). Although the network size increases, it becomes sparser (irrespectively from the considered representation). Similar results are observed for the BAN and the BUN at the daily time scale. See also [16].

$$\mathrm{RPMA}_t = 100 \log_{10} \left(\frac{P_t}{\frac{1}{\tau} \sum_{s=t-1-\tau}^{t-1} P_s} \right) \tag{1}$$

with τ representing a tunable temporal parameter. As shown in [16], the standard deviation and the kurtosis diverge as the network size grows larger than the value observed in correspondence of the Mt. Gox failure, thus confirming the 'two regimes hypothesis'. Moreover, as fig. 2 shows, larger values of the aforementioned moments (observed *after* the Mt. Gox failure) correspond to price drops, while temporal snapshots corresponding to smaller values of the same quantities seem to be characterised by price increases.

A multivariate Granger test [24] has been also carried out to unveil possible lagged correlations hidden in the data (see fig. 5 in [16]). To this aim, data have been split in two sub-samples, i.e. 2010-2013 and 2014-2017, and the number of nodes N, the number of links L and the higher moments of the empirical (out- and in-) degrees distributions have been put in relation with the log-returns of the Bitcoin price (in USD), within each sub-sample. To sum up, when the BUN is considered at the weekly time scale, a positive feedback loop occurs between N and the price log-returns, whereas at the daily time scale a price increase predicts an increase of the number of nodes N but the viceversa is no longer true. The causality structure is consistent within the two sub-samples.

Analysis of the BUN mesoscale structure. Let us now revise the results concerning the mesoscale structure of the BUN. A recently proposed method [25] based on the *surprise* score function was adopted by the authors of [17] to assess the statistical significance of a peculiar mesoscale organization, known as *core-periphery* structure. According to the interpretation proposed in [25], revealing the core-periphery structure by minimising the surprise means individuating the partition that is least likely to be explained by the null model known as *Random Graph Model* (RGM) with respect to the null model known as *Stochastic Block Model* (SBM) - see also Appendix A. As fig. 3 shows, a core-periphery

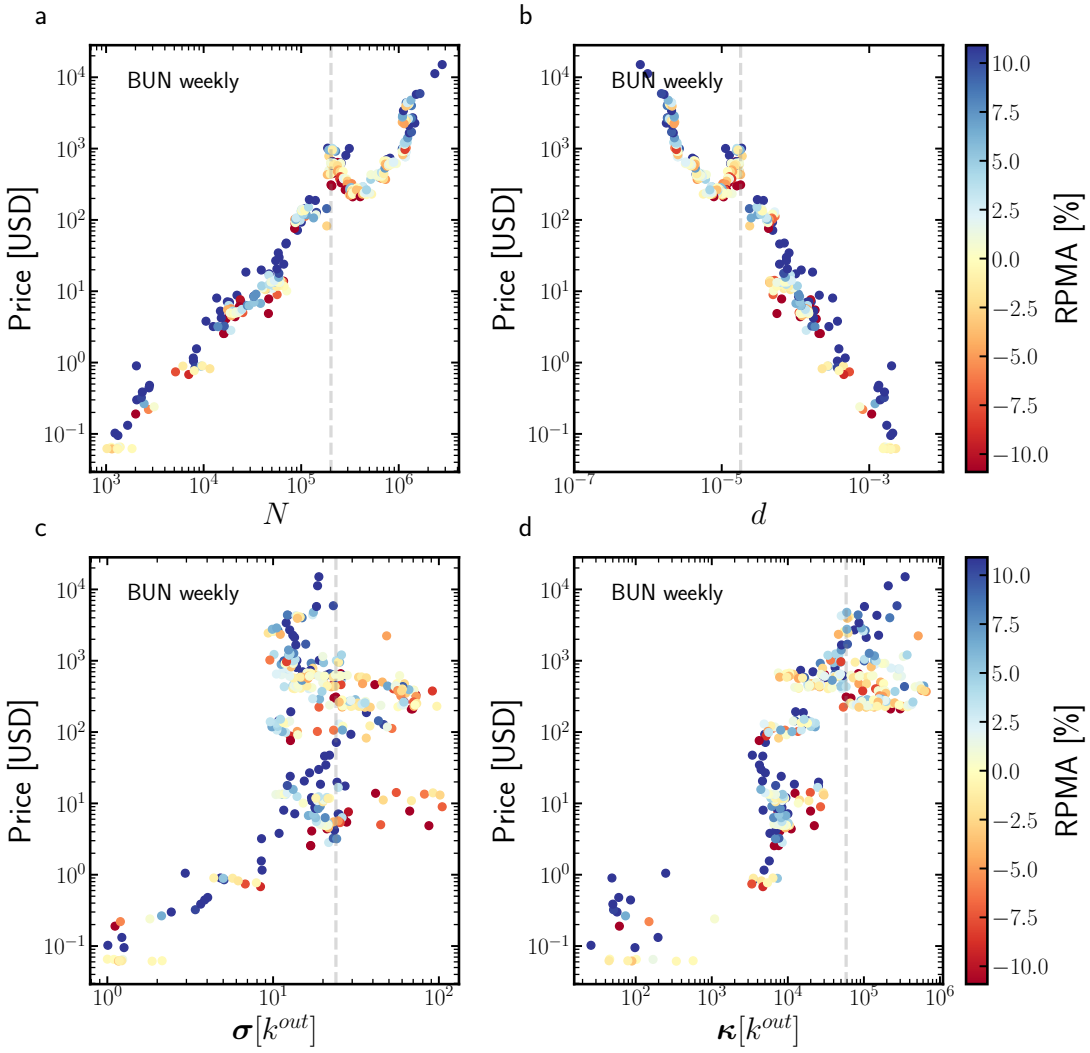

FIG. 2. Correlation between the Bitcoin price in USD, the basic statistics (number of nodes and link density - top panels) and the moments of the out-degrees distribution (bottom panels) for the BUN at the weekly time scale. Additionally, each dot representing an observation is coloured according to the value of the Ratio between the current Price and its Moving Average (RPMA) indicator. The vertical, dashed line coincides with the bankruptcy of Mt. Gox. Purely structural quantities are correlated with exogenous quantities as the Bitcoin price; see, for example, the evolution of the out-degrees standard deviation whose larger values (observable *after* the Mt. Gox failure) correspond to price drops. See also [16].

structure is indeed present: more precisely, during the biennium 2014-2015 the core size amounts at $\simeq 30\%$ of the total network size; after 2016, instead, it seems to shrink back to 2010-2013 values. According to the interpretation provided above, this finding, in turn, means that the BUN is indeed characterised by subgraphs with very different link densities, an evidence that cannot be explained by a model defined by just one global parameter as the one characterising the RGM.

A deeper inspection of the BUN core-periphery structure reveals it to be even richer: in fact, the core portion of the BUN is, actually, the strongly-connected component (SCC) of a *bow-tie* structure whose remaining portions (e.g. the IN and OUT components) compose the BUN periphery [17]. More specifically, while the SCC is the set of nodes that are mutually reachable (i.e. a directed path from any node to any other node, within the SCC, exists), the IN and OUT components are respectively defined as the set of nodes from which the SCC can be reached and the set of

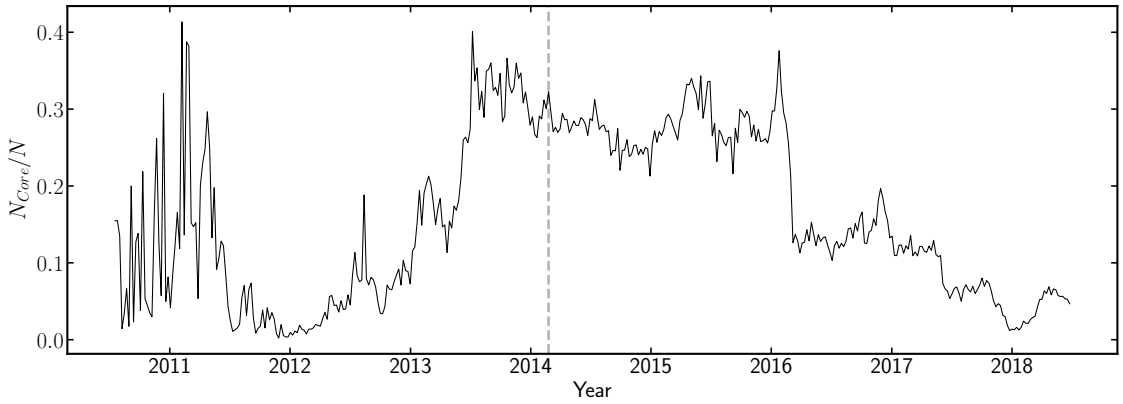

FIG. 3. Evolution of the percentage of nodes belonging to the core portion of the BUN at the weekly time scale. During the biennium 2012-2013 the core portion of the BUN steadily rises until it reaches $\simeq 30\%$ of the network; afterwards, during the biennium 2014-2015, it remains quite constant; then, during the last two years covered by our data set (i.e. 2016-2018), the core portion of the BUN shrinks and the percentage of nodes belonging to it goes back to the pre-2012 values. The vertical, dashed line coincides with the bankruptcy of Mt. Gox.

nodes that can be reached from the SCC. Hence, the picture provided by the evolution of the core-periphery structure can be further refined as follows: since 2016 both the SCC and the OUT-component shrink while the IN-component becomes the dominant portion of the network [17].

An additional analysis, aimed at better quantifying the extent to which a generic, purely topological quantity X and the Bitcoin price are related, can be carried out by plotting the evolution of the temporal z-score

$$z_X^{(t)} = \frac{X^{(t)} - \overline{X}}{s_X} \tag{2}$$

where $\overline{X} = \sum_t \frac{X^t}{T}$ is the mean over a sample of values covering the period T before time t - in our case, the year before t - and $s_X = \sqrt{\overline{X^2} - \overline{X}^2}$ is the corresponding standard deviation. For example, the choice $X = \sigma[k^{out}]$ allows price drawdowns to be revealed and, in some cases, anticipated [16]: in the triennium 2010-2012, and after 2017, the price grows as $z^{(t)}_{\sigma[k^{out}]}$ increases while drawdowns appear in periods during which $z^{(t)}_{\sigma[k^{out}]}$ decreases. Other possible choices are $X = N_{core}$ and $X = r$, i.e. the number of core nodes and the network reciprocity, defined as $r = \frac{\sum_i \sum_{j(\neq i)} a_{ij} a_{ji}}{\sum_i \sum_{j(\neq i)} a_{ij}}$, i.e. as the percentage of links having a 'partner' pointing in the opposite direction. The evolution of the temporal z-score for the two aforementioned quantities is shown in fig. 4. Overall, the two trends show some similarities, being characterised by peaks in correspondence of the so-called *bubbles*, i.e. periods of 'unsustainable' price growth [26]: interestingly, such periods are characterised by values of the inspected topological quantities which are significant also in a statistical sense, as the value of the corresponding temporal z-score proves (in fact, $z^{(t)} \geq 2$ in both cases). Moreover, peaks are also revealed in the triennium 2014-2016, thus signalling some kind of 'activity' missed by purely financial indicators (e.g. the RPMA).

The Bitcoin Lightning network

Let us now move to review the results concerning the BLN. In what follows we will focus on the daily-block snapshot representation.

Basic statistics. As observed for the BAN and the BUN, both the number of nodes N and the number of links $L = \sum_i \sum_{j(>i)} b_{ij}$ of the BLN increase steadily in time while it becomes sparser. Interestingly, however, the evolution of the BLN link density seems to point out the presence of two regimes: as fig. 5 shows, during the first phase (i.e. $N \leq 10^3$) L increases linearly in N and the link density is well described by the functional dependence $d \sim N^{-1}$; afterwards, the link density decrease slows down, seemingly indicating that L has started to grow in a super-linear

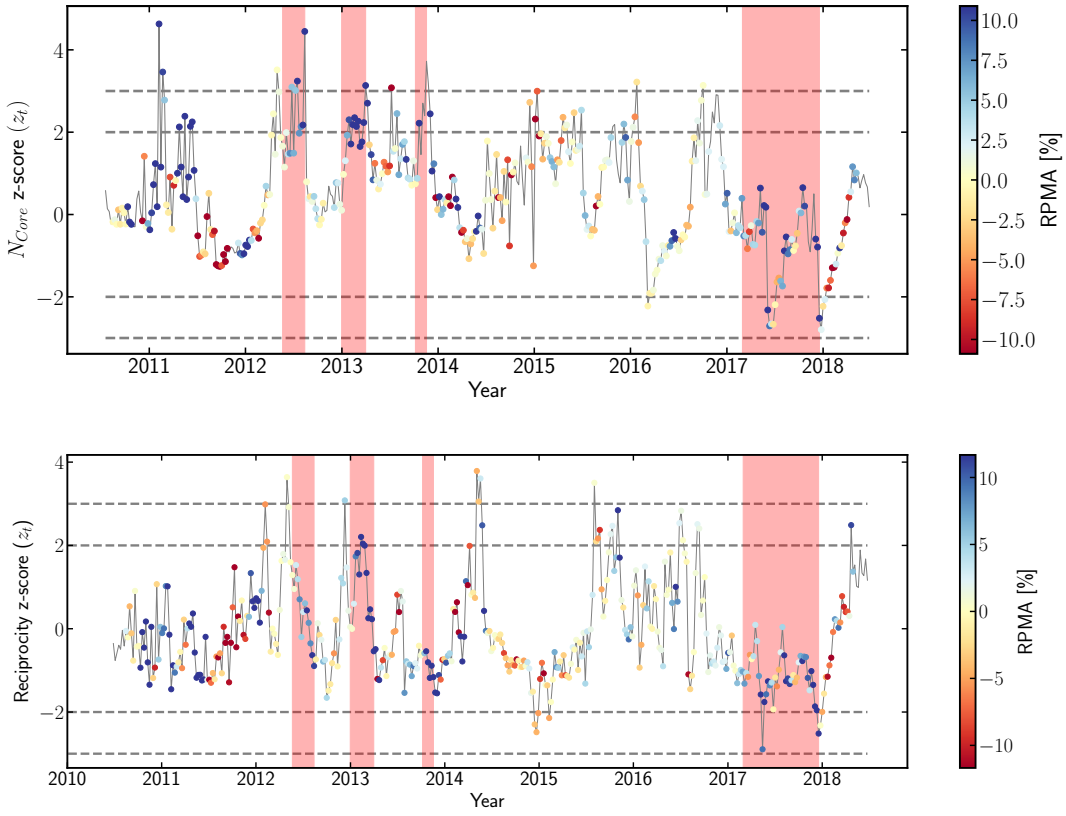

FIG. 4. Evolution of the temporal z-score for the number of core nodes (top panel) and for the reciprocity (bottom panel), for the BUN weekly representation. Shaded areas individuate the so-called *bubbles*, i.e. periods of price increase according to [26]. Additionally, each dot representing an observation is coloured according to the value of the Ratio between the current Price and its Moving Average (RPMA) indicator. Overall, the two trends show some similarities as peaks are clearly visible in correspondence of the so-called *bubbles*, identified by the shaded areas (see also [26]). Interestingly, these values are significant in a statistical sense, as the temporal z-scores reach values $z^{(t)} \geq 2$. See also [17].

fashion with respect to N.

Analysis of the BLN mesoscale structure. In order to inspect the evolution of the BLN 'centralisation', the authors in [18] have considered two different sets of quantities. First, they have computed the Gini index

$$G_c = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} |c_i - c_j|}{2N \sum_{i=1}^{N} c_i} \tag{3}$$

for four centrality measures, i.e. the *degree*, *closeness*, *betweenness* and *eigenvector* centrality (respectively indicated with the symbols $c_i = k_i^c, c_i^c, b_i^c, e_i^c$ - see also Appendix B) and plotted it versus the number of nodes. Interestingly, G_c increases for three measures out of four, pointing out that the values of centrality are more and more unevenly distributed [18].

Additionally, they have also computed the so called *centralisation indices*, encoding the comparison between the structure of a given network and that of a reference network, i.e. the 'most centralised' structure - see also Appendix B. For what concerns the degree, closeness and betweenness centrality, it is the *star graph*: for what concerns the eigenvector index, the star graph does not represent the maximally-centralised structure, although it is kept for the sake of homogeneity with the other quantities. The evolution of the centralisation indices indicates that the BLN is not evolving towards a star graph (indeed, a too simplistic picture to faithfully describe the BLN topology) but towards a suitable generalization of it, i.e. the core-periphery structure [18] (see also later).

The authors in [18] have also benchmarked the observations concerning the evolution of the centrality and the centralisation indices with the predictions, for the same quantities, output by the maximum-entropy null model known

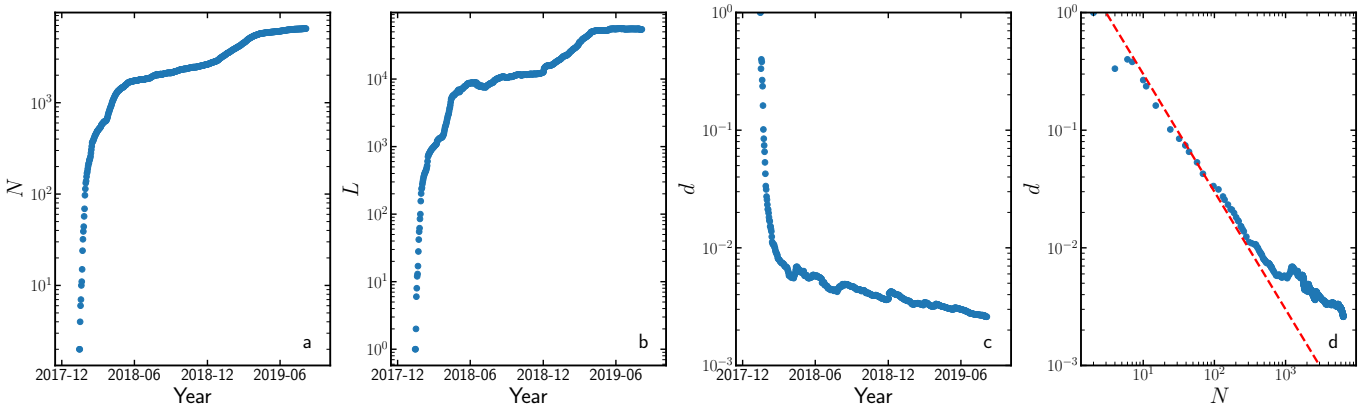

FIG. 5. Evolution of the total number of nodes N, total number of links L and link density $d = \frac{2L}{N(N-1)}$ for the BLN daily-block snapshot representation. As for the BAN and the BUN, the position $d \sim N^{-1}$ well describes the link density dependence on N, at least for the snapshots for which $N \leq 10^3$. See also [18].

as *Undirected Binary Configuration Model* (UBCM - see also Appendix C). To this aim, they have explicitly sampled the ensembles of networks induced by the UBCM [27, 28] and compared the ensemble average of the quantities of interest with the corresponding empirical values. From a merely technical point of view, the authors adopt an iterative, reduced algorithm to solve the system of equations defining the UBCM, i.e.

$$k_i(\mathbf{A}) = \sum_{j(\neq i)=1}^{N} \frac{x_i x_j}{1 + x_i x_j}, \forall\, i \quad \Longrightarrow \quad x_k^{(n)} = \frac{k(\mathbf{A})}{\sum_{k'(\neq k)} f(k') \left[\frac{x_{k'}^{(n-1)}}{1 + x_k^{(n-1)} x_{k'}^{(n-1)}} \right]}, \forall\, k \tag{4}$$

a choice allowing them to solve it within tens of seconds even for configurations with millions of nodes [17] - see also Appendix C. As fig. 6 shows, such a comparison reveals that the UBCM tends to overestimate the values of the Gini index for the degree, the closeness and the betweenness centrality and to underestimate its values for the eigenvector centrality. This seems to point out a non-trivial (i.e. not reproducible by just enforcing the degrees) tendency of well-connected nodes to establish connections among themselves - likely, with nodes having a smaller degree attached to them: such a disassortative structure could explain the less-than-expected level of unevenness characterising the other centrality measures, as each of the nodes behaving as the 'leaves' of the hubs would basically have the same values of degree, closeness and betweenness centrality.

For what concerns the analysis of the centralisation indices, fig. 6 shows that the UBCM underestimates both the betweenness- and the eigenvector-centralisation indices: in other words, a tendency to centralisation 'survives' even after the information encoded into the degrees is properly accounted for, letting the picture of a network characterised by some kind of more-than-expected 'star-likeness' emerge. This observation can be better formalised by analysing the BLN mesoscale structure via the optimization of surprise: as observed for the BUN, a core-periphery structural organization, whose statistical significance increases over time, indeed emerges [18] (see also fig. 7).

In [17], the authors have also adapted the iterative, reduced algorithm cited above to the resolution of the *Directed Binary Configuration Model* (DBCM - see also Appendix C).

A quick look at the weighted structure of the BLN. Having a quick look at the weighted structure of the BLN leads to two notable observations: both the *total amount of exchanged bitcoins* and the *unevenness of their distribution* increase. This trend is confirmed by the evolution the Gini coefficient whose value reaches 0.9 for the last snapshots of our data set. On average, across the entire period, about the 10% (50%) of nodes holds the 80% (99%) of the bitcoins at stake in the network [18].

DISCUSSION

The public availability of the complete Bitcoin transaction history allows to quantify the evolution of structural quantities that characterise different Bitcoin Transaction Networks, and to inspect the inter-dependency between them and the dynamics of Bitcoin price - an analysis that intends, also, to gain insight on the *behaviour* of Bitcoin

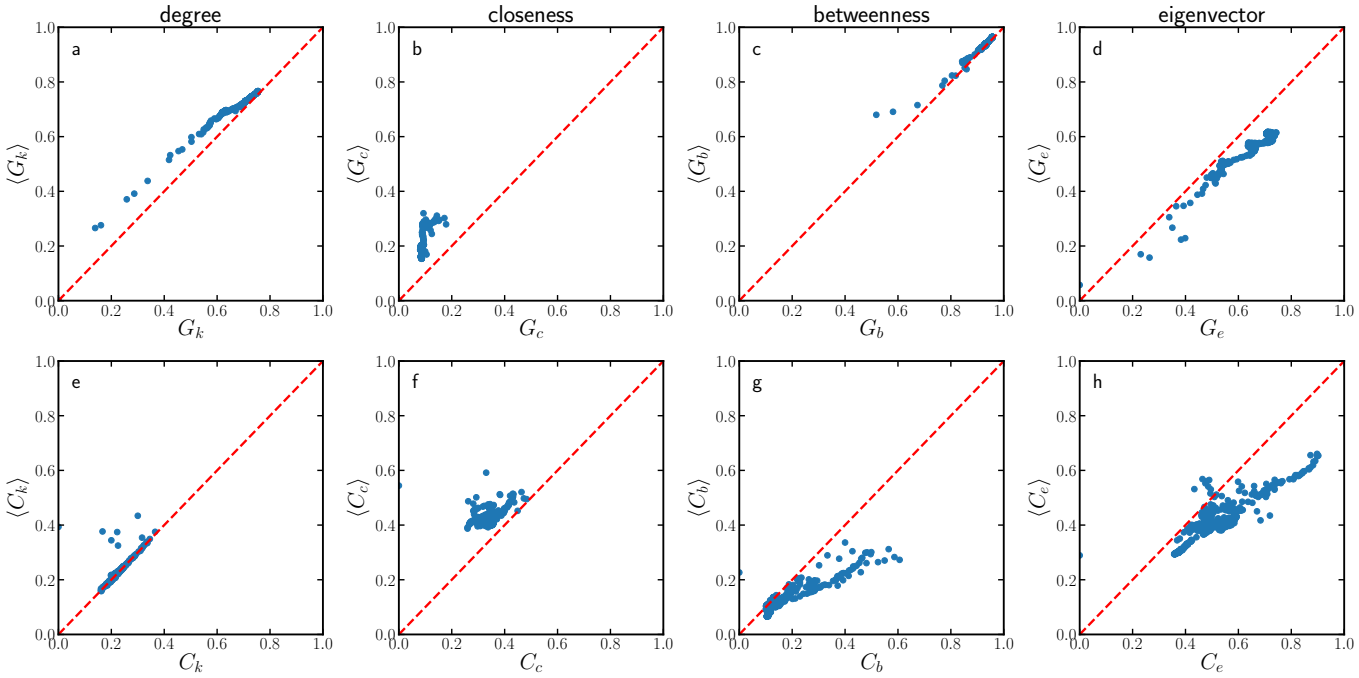

FIG. 6. Top panels: comparison between the observed Gini index for the degree, closeness, betweenness and eigenvector centrality (x-axis) and their expected value, computed under the UBCM (y-axis) for the BLN daily-block snapshot representation. Bottom panels: comparison between the observed degree-, closeness-, betweenness-and eigenvector-centralisation measures and their expected value computed under the UBCM. Once the information contained into the degree sequence is properly accounted for, a (residual) tendency to centralisation is still visible, letting the picture of a network characterised by some kind of more-than-expected 'star-likeness' emerge. See also [18].

users; still, the understanding of the mechanisms shaping the joint evolution of the three quantities above remains far from being complete.

This paper aims at providing an overview of the most recent results on the topic achieved in the last years. One of the main messages concerns the possibility to retrieve signals of exogenous events by analysing the blockchain-induced transaction networks: the best example is provided by the failure of Mt. Gox in 2014, an event deeply affecting the Bitcoin Address Network and the Bitcoin User Network structure (i.e. irrespective from the chosen representation). From this point of view, out-degrees have been found to be particularly informative properties: higher moments of the out-degrees distribution (as the standard deviation, the skewness and the kurtosis) diverge as the network size grows larger than the value observed in correspondence of the Mt. Gox failure; besides, the out-degrees heterogeneity rises during periods of price decline and vice-versa.

Such a result is further refined by a Granger causality analysis, revealing that during the triennium 2010-2012 an increase of the out-degrees standard deviation *causes* a price decline [16]. These results, in turn, suggest a sort of behavioural explanation for the price dynamics displayed during the early stages of Bitcoin: during periods in which the price continuously increases, always larger numbers of traders are, in turn, attracted to the system; the former ones, likely performing only few transactions, link to the network hubs (usually exchange markets) that gain a large number of connections over the course of the weeks and cause the price to rise even more.

Interestingly, the analysis of the Bitcoin Lightning Network reveals the same trends observed for the BAN and the BUN appear, as the emergence of a statistically-significant core-periphery structure, of an uneven distribution of the nodes' centrality and wealth, etc.: these results suggest tendency of the Bitcoin 'Layer 2' network to become less distributed. This process has the undesirable consequence of making this off-chain payment network less resilient to failures, malicious attacks.

All the results reviewed in this article ultimately - and consistently - point out a tendency to centralisation that has been observed in the Bitcoin structure at different levels [18, 29], an evidence that deserves to be investigated in greater detail.

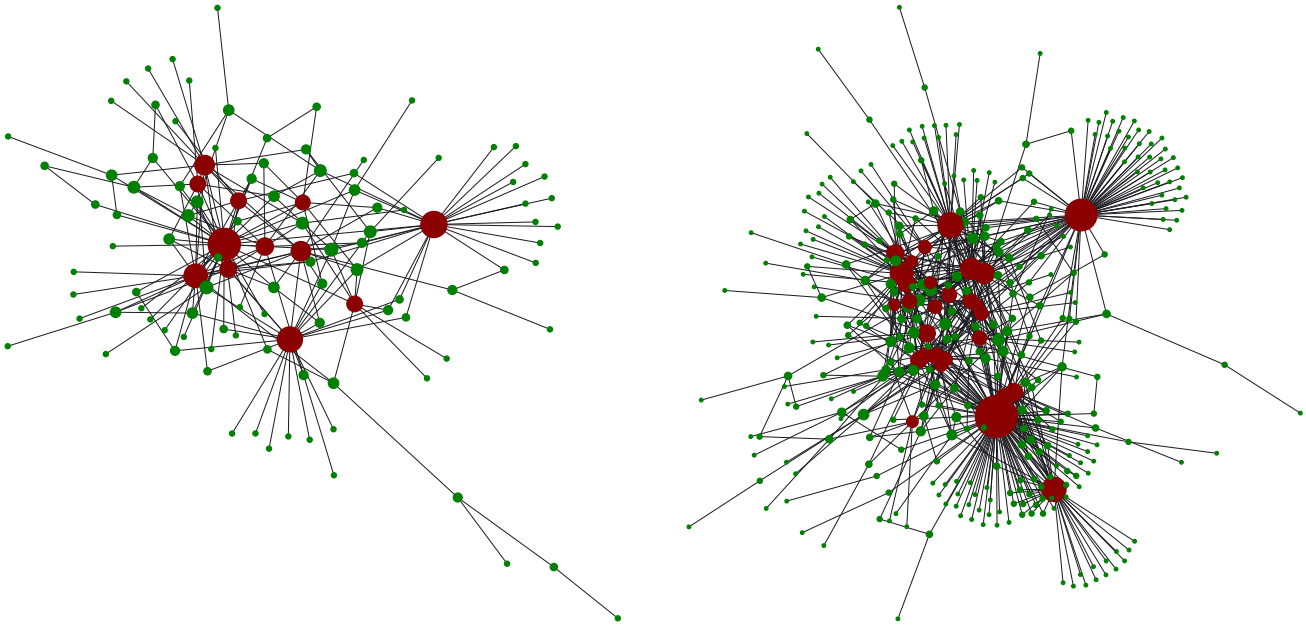

FIG. 7. Core-periphery structure of the BLN daily-block representation on day 17 (left panel) and on day 35 (right panel), with core-nodes drawn in red and periphery-nodes drawn in green. See also [18].

ACKNOWLEDGEMENTS

C.J.T. acknowledges financial support by the University of Zurich through the University Research Priority Program on Social Networks. T.S. acknowledge support from the EU project SoBigData-PlusPlus (grant no. 871042). The authors acknowledge A. Bovet, C. Campajola, F. Mottes, V. Restocchi, J.-H. Lin for useful discussions.

AUTHOR CONTRIBUTIONS

All authors wrote, reviewed and approved the manuscript.

COMPETING INTERESTS

The authors declare no competing financial interests.

DATA AVAILABILITY

Data concerning the entire Bitcoin transaction history is publicly available at the address *https://www.blockchain.com/*. By synchronising the desktop client Bitcoin Core, available at *https://bitcoin.org/en/download*, anybody can have a local copy of the entire transaction history.

[1] A. M. Antonopoulos. *Mastering Bitcoin: Programming the Open Blockchain* (O'Reilly Media, Inc., 2017).
[2] S. Nakamoto. Bitcoin: A peer-to-peer electronic cash system (2008).
[3] H. Halaburda, M. Sarvary. *Beyond Bitcoin: The Economics of Digital Currencies* (Palgrave Macmillan, 2016).
[4] F. Glaser. *Proceedings of the Hawaii International Conference on System Sciences 2017 (HICSS-50)* (2017).
[5] J. Poon, T. Dryja. The Bitcoin Lightning Network: scalable off-chain instant payments (2016).

[6] D. Kondor, I. Csabai, G. Vattay. Do the rich get richer? An empirical analysis of the Bitcoin transaction network. *PLoS ONE* **9** (2014).

[7] M. A. Javarone, C. S. Wright. From Bitcoin to Bitcoin Cash: a network analysis. *arXiv:1804.02350v2* (2018).

[8] F. Parino, M. G. Beiró, L. Gauvin. Analysis of the Bitcoin blockchain: socio-economic factors behind the adoption. *European Physical Journal Data Science* **7**, 38 (2018).

[9] E. Androulaki, G. O. Karame, M. Roeschlin, T. Scherer, S. Capkun. *International Conference on Financial Cryptography and Data Security* (Springer, 2013), pp. 34–51.

[10] M. Harrigan, C. Fretter. *Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), 2016 Intl IEEE Conferences* (IEEE, 2016), pp. 368–373.

[11] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, S. Savage. *Proceedings of the 2013 conference on Internet measurement conference* (ACM, 2013), pp. 127–140.

[12] M. Ober, S. Katzenbeisser, K. Hamacher. Structure and anonymity of the Bitcoin transaction graph. *Future Internet* **5**, pp. 237–250 (2013).

[13] F. Reid, M. Harrigan. *Security and Privacy in Social Networks* (Springer, 2013), pp. 197–223.

[14] D. Garcia, C. J. Tessone, P. Mavrodiev, N. Perony. The digital traces of bubbles: feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of The Royal Society Interface* **11**, 99 (2014).

[15] A. El Bahrawy, L. Alessandretti, A. Baronchelli. Wikipedia and digital currencies: interplay between collective attention and market performance. *Frontiers in Blockchain* (2019).

[16] A. Bovet, C. Campajola, F. Mottes, V. Restocchi, N. Vallarano, T. Squartini, C. J. Tessone. The evolving liaisons between the transaction networks of Bitcoin and its price dynamics. *arXiv:1907.03577* (2019).

[17] N. Vallarano, C. J. Tessone, T. Squartini. Exploring the Bitcoin mesoscale structure. *in preparation* (2020).

[18] J.-H. Lin, K. Primicerio, T. Squartini, C. Decker, C. J. Tessone. Lightning Network: a second path towards centralisation of the Bitcoin economy. *arXiv:2002.02819* (2020).

[19] P. Tasca, A. Hayes, S. Liu. The evolution of the Bitcoin economy: extracting and analyzing the network of payment relationships. *Journal of Risk Finance* **19**, pp. 94–126 (2018).

[20] D. Ron, A. Shamir. *International Conference on Financial Cryptography and Data Security* (Springer, 2013), pp. 6–24.

[21] A. Baumann, B. Fabian, M. Lischke. *WEBIST* (2014).

[22] H. Bauke. Parameter estimation for power-law distributions by maximum likelihood methods. *European Physical Journal B* **58**, 167–173 (2007).

[23] G. Bianconi, A.-L. Barabasi. Bose-einstein condensation in complex networks. *Physical Review Letters* **86** (2001).

[24] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, pp. 424–438 (1969).

[25] J. v. L. de Jeude, G. Caldarelli, T. Squartini. Detecting core-periphery structures by surprise. *Europhysics Letters* **125**, 68001 (2019).

[26] S. Wheatley, D. Sornette, T. Huber, M. Reppen, R. N. Gantner. Are Bitcoin bubbles predictable? Combining a generalized Metcalfe's law and the LPPLS model. *Swiss Finance Institute Research Paper* (2018).

[27] J. Park, M. E. Newman. Statistical mechanics of networks. *Physical Review E* **70**, 066117 (2004).

[28] T. Squartini, D. Garlaschelli. Analytical maximum-likelihood method to detect patterns in real networks. *New Journal of Physics* **13**, 083001 (2011).

[29] A. Gervais, G. Karame, S. Capkun, V. Capkun. Is Bitcoin a decentralized currency? *IEEE Security & Privacy* **12**, pp. 54–60 (2014).

[30] C. Nicolini, A. Bifone. Exploring the limits of community detection strategies in complex networks. *Scientific Reports* **6** (2016).

[31] D. Garlaschelli, M. I. Loffredo. Maximum likelihood: extracting unbiased information from complex networks. *Physical Review E* **78**, 015101 (2008).

[32] N. Dianati. A maximum entropy approach to separating noise from signal in bimodal affiliation networks. *arXiv:1607.01735* (2016).

[33] C. Decker, R. Wattenhofer. *IEEE P2P 2013 Proceedings* (IEEE, 2013), pp. 1–10.

APPENDIX A - DETECTING CORE-PERIPHERY STRUCTURES BY SURPRISE

The 'generalised' star graph structure also known as *core-periphery structure* is defined by a densely-connected core of nodes surrounded by a periphery of loosely-connected vertices. In order to check for its presence, we implement a recently-proposed approach [25], prescribing to minimise the score function known as *bimodular surprise* and reading

$$\mathscr{S}_{\parallel} = \sum_{i \geq l^{*}_{\bullet}} \sum_{j \geq l^{*}_{\circ}} \frac{\binom{V_{\bullet}}{i}\binom{V_{\circ}}{j}\binom{V-(V_{\bullet}+V_{\circ})}{L-(i+j)}}{\binom{V}{L}} \tag{5}$$

which is nothing else that the multinomial version of the *surprise*, originally proposed to carry out a *community detection* exercise [30]. The presence of three different binomial coefficients allows three different 'species' of links to be accounted for: the binomial coefficient $\binom{V_\bullet}{i}$ enumerates the number of ways i links can redistributed *within* the first module (e.g. the core portion), the binomial coefficient $\binom{V_\circ}{j}$ enumerates the number of ways j links can redistributed *within* the second module (e.g. the periphery portion) and the binomial coefficient $\binom{V-(V_\bullet+V_\circ)}{L-(i+j)}$ enumerates the number of ways the remaining $L-(i+j)$ links can be redistributed *between* the first and the second module, i.e. over the remaining $V-(V_\bullet+V_\circ)$ node pairs; the values i and j are bounded by the values $V_\bullet$ and $V_\circ$, although the sum $i+j$ can range between $l_\bullet^*+l_\circ^*$ and the minimum between L and $V_\bullet+V_\circ$.

From a technical point of view, $\mathscr{S}_\parallel$ is the p-value of a multivariate hypergeometric distribution, describing the probability of $i+j$ successes in L draws (without replacement), from a finite population of size V that contains exactly $V_\bullet$ objects with a first specific feature and $V_\circ$ objects with a second specific feature.

APPENDIX B - CENTRALITY AND CENTRALIZATION MEASURES

Indices measuring the centrality of a node aim at quantifying the 'importance' of a node in a network, according to some specific topological property. Among the measures proposed so far, of particular relevance are the *degree centrality*, the *closeness centrality*, the *betweenness centrality* and the *eigenvector centrality*:

- the *degree centrality* of node i coincides with the degree of node i, i.e. the number of its neighbours, normalised by the maximum attainable value, i.e. $N-1$:

$$k_i^c = \frac{k_i}{N-1}. \tag{6}$$

  From the definition above, it follows that the most central node, according to the degree variant, is the one connected to all the other nodes;

- the *closeness centrality* of node i is defined as

$$c_i^c = \frac{N-1}{\sum_{j(\neq i)=1}^{N} d_{ij}} \tag{7}$$

  where d_{ij} is the topological distance between nodes i and j, i.e. the length of the shortest path(s) connecting them. From the definition above, it follows that the most central node, according to the closeness variant, is the one lying at distance 1 from each other node;

- the *betweenness centrality* of node i is given by

$$b_i^c = \sum_{s(\neq i)=1}^{N} \sum_{t(\neq i,s)=1}^{N} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{8}$$

  where σ_{st} is the total number of shortest paths between node s and t and $\sigma_{st}(i)$ is the number of shortest paths between nodes s and t that pass through node i. From the definition above, it follows that the most central node, according to the betweenness variant, is the one lying 'between' any two other nodes;

- the *eigenvector centrality* of node i, e_i^c, is defined as the i-th element of the eigenvector corresponding to the largest eigenvalue of the binary adjacency matrix (whose existence is ensured by the Perron-Frobenius theorem). According to the definition above, a node with large eigenvector centrality is connected to other 'well connected' nodes [21].

The centrality indices defined above provide a rank of the nodes of a network. Sometimes, however, it is useful to compactly describe a certain network structure in its entirety. To this aim, a family of indices, known as *centralisation indices*, has been defined. In mathematical terms, any centralisation index reads

$$C_c = \frac{\sum_{i=1}^{N}(c^* - c_i)}{\max\{\sum_{i=1}^{N}(c^* - c_i)\}} \tag{9}$$

where $c^* = \max\{c_i\}_{i=1}^{N}$ represents the empirical, maximum value of the chosen centrality measure (i.e. computed on the network under consideration) and the denominator is calculated over a benchmark graph, defined as the one providing the maximum attainable value of the quantity $\sum_{i=1}^{N}(c^* - c_i)$. The most centralised structure, according to the degree, closeness and betweenness centrality is the *star graph*, in correspondence of which one finds that

- $\sum_{i=1}^{N}(k^* - k_i^c) = (N-1)(N-2)$;
- $\sum_{i=1}^{N}(c^* - c_i^c) = \frac{(N-1)(N-2)}{2N-3}$;
- $\sum_{i=1}^{N}(b^* - b_i^c) = \frac{(N-1)^2(N-2)}{2}$.

For what concerns the *eigenvector-centralisation* index, the star graph does not represent the maximally centralised structure; however, for the sake of comparison with the quantities above, the authors in [18] have calculated it on a star graph as well, in correspondence of which one finds that $\sum_{i=1}^{N}(e^* - e_i^c) = (\sqrt{N-1}-1)(N-1)/(\sqrt{N-1}+N-1)$.

APPENDIX C - AN ITERATIVE METHOD TO SOLVE NULL MODELS

The significance of any result can be assessed only after a comparison with a properly-defined benchmark (or null) model. To this aim, one can consider the *Exponential Random Graph* (ERG) framework. Generally speaking, the problem to be solved in order to define a benchmark model within such a framework reads

$$\max_P \left\{ S[P] - \sum_{i=0}^{M} \theta_i \left[\sum_{\mathbf{A}} P(\mathbf{A}) C(\mathbf{A}) - \langle C_i \rangle \right] \right\} \tag{10}$$

where

$$S[P] = -\sum_{\mathbf{A}} P(\mathbf{A}) \ln P(\mathbf{A}) \tag{11}$$

is *Shannon entropy* and $\vec{C}(\mathbf{A})$ is an M-dimensional vector of constraints representing the information defining the benchmark - notice that $C_0 = \langle C_0 \rangle = 1$ sums up the normalization condition of the probability distribution $P(\mathbf{A})$. The solution to the problem above reads

$$P(\mathbf{A}, \vec{\theta}) = \frac{e^{-H(\mathbf{A},\vec{\theta})}}{Z(\vec{\theta})} \tag{12}$$

with $Z(\vec{\theta}) = \sum_{\mathbf{A}} P(\mathbf{A}, \vec{\theta})$ representing the *partition function* and $H(\mathbf{A}, \vec{\theta}) = \vec{\theta} \cdot \vec{C}(\mathbf{A})$ representing the *Hamiltonian*, i.e. the functions summing up the normalization condition and the imposed constraints, respectively.

The Undirected Binary Configuration Model (UBCM). In case the Undirected Binary Configuration Model (UBCM) is chosen as a benchmark, the Hamiltonian reads

$$H(\mathbf{A}, \theta) \equiv \vec{\theta} \cdot \vec{k}(\mathbf{A}) = \sum_{i=1}^{N} \sum_{j(>i)=1}^{N} (\theta_i + \theta_j) a_{ij} \tag{13}$$

a position leading to the probability function $P(\mathbf{A}) = \prod_i \prod_{j(>i)} p_{ij}^{a_{ij}} (1-p_{ij})^{1-a_{ij}}$ with $p_{ij}^{\mathrm{UBCM}} \equiv \frac{e^{-(\theta_i+\theta_j)}}{1+e^{-(\theta_i+\theta_j)}} \equiv \frac{x_i x_j}{1+x_i x_j}$. The unknown parameters can be estimated by invoking a second maximization principle, i.e. the maximization of the likelihood function. The latter is defined as

$$\mathcal{L}(\vec{x}) = \ln P(\mathbf{A}|\vec{x}) \tag{14}$$

and needs to be optimised with respect to the vector of unknown parameters $\vec{x}$. Remarkably, whenever the probability distribution is exponential (as the one deriving from Shannon entropy maximization), the likelihood maximization leads to the system $\langle \vec{C} \rangle = \vec{C}(\mathbf{A})$ to be solved, that in the UBCM case reads

$$k_i(\mathbf{A}) = \sum_{j(\neq i)=1}^{N} \frac{x_i x_j}{1 + x_i x_j}, \; \forall \, i. \tag{15}$$

In order to solve the system above, the iterative recipe

$$x_i^{(n)} = \frac{k_i(\mathbf{A})}{\sum_{j(\neq i)=1}^{N} \left[\frac{x_j^{(n-1)}}{1 + x_i^{(n-1)} x_j^{(n-1)}} \right]}, \; \forall \, i \tag{16}$$

can be employed; naturally, such a recipe needs to be initialised: the values $x_i^{(0)} = \frac{k_i(\mathbf{A})}{\sqrt{L}}$, $\forall \, i$ can be chosen, i.e. the solution to the system of equations defining the UBCM in the sparse case. Let us notice that the computation of the system above can be further sped up, by assigning to the nodes with the same degree k the same value of the hidden variable x, i.e.

$$x_k^{(n)} = \frac{k(\mathbf{A})}{\sum_{k'(\neq k)} f(k') \left[\frac{x_{k'}^{(n-1)}}{1 + x_k^{(n-1)} x_{k'}^{(n-1)}} \right]}, \; \forall \, k \tag{17}$$

where the sum runs over the *distinct* values of the degrees and $f(k)$ is the number of nodes whose degree is k [31].

The Directed Binary Configuration Model (DBCM). In case the Directed Binary Configuration Model (DBCM) is chosen as a benchmark, the Hamiltonian reads

$$H(\mathbf{A}, \alpha, \beta) \equiv \vec{\alpha} \cdot \vec{k}^{out}(\mathbf{A}) + \vec{\beta} \cdot \vec{k}^{in}(\mathbf{A}) = \sum_{i=1}^{N} \sum_{j(\neq i)=1}^{N} (\alpha_i + \beta_j) a_{ij} \tag{18}$$

a position leading to the probability function $P(\mathbf{A}) = \prod_i \prod_{j(\neq i)} p_{ij}^{a_{ij}} (1 - p_{ij})^{1 - a_{ij}}$ with $p_{ij}^{\mathrm{DBCM}} \equiv \frac{e^{-(\alpha_i + \beta_j)}}{1 + e^{-(\alpha_i + \beta_j)}} \equiv \frac{x_i y_j}{1 + x_i y_j}$. As for the UBCM, the unknown parameters can be estimated by invoking the maximization of the likelihood function. In the DBCM case, it leads to the system

$$k_i^{out}(\mathbf{A}) = \sum_{j(\neq i)=1}^{N} \frac{x_i y_j}{1 + x_i y_j}, \; \forall \, i \tag{19}$$

$$k_i^{in}(\mathbf{A}) = \sum_{j(\neq i)=1}^{N} \frac{x_j y_i}{1 + x_j y_i}, \; \forall \, i. \tag{20}$$

In order to solve the system above, the iterative recipe (originally proposed in [32] and further refined in [17])

$$x_i^{(n)} = \frac{k_i^{out}(\mathbf{A})}{\sum_{j(\neq i)=1}^{N} \left[\frac{y_j^{(n-1)}}{1 + x_i^{(n-1)} y_j^{(n-1)}} \right]}, \; \forall \, i \tag{21}$$

$$y_i^{(n)} = \frac{k_i^{in}(\mathbf{A})}{\sum_{j(\neq i)=1}^{N} \left[\frac{x_j^{(n-1)}}{1 + x_j^{(n-1)} y_i^{(n-1)}} \right]}, \; \forall \, i \tag{22}$$

can be employed. As for the UBCM case, the solutions to the system of equations defining the DBCM in the sparse case, i.e. $x_i^{(0)} = \frac{k_i^{out}(\mathbf{A})}{\sqrt{L}}$, $\forall\, i$ and $y_i^{(0)} = \frac{k_i^{in}(\mathbf{A})}{\sqrt{L}}$, $\forall\, i$, can be chosen to initialise the recipe above. The computation of the system above can be further sped up, by assigning to the nodes with the same (pair of) out- and in-degrees (k_i^{out}, k_i^{in}) the same pair of values (x, y) [31].