Data Science for investigating the block-chain in Bitcoin

Introduction: The white-paper on Bitcoin appeared in November 2008, written by a computer programmer using the pseudonym "Satoshi Nakamoto". His invention is an open-source, peer-to-peer digital currency (being fully electronic, with no physical manifestation). Money transactions do not require a third-party intermediary, with no traditional financial-institution involved in transactions. Therefore, the Bitcoin network is completely decentralized, with all the parts of transactions performed by the users of the system.

The buyer and seller directly interact (peer-to-peer), but without using their real identities, and no personal information is transferred from one to the other. However, unlike a fully anonymous transaction, there is a transaction record. A complete transaction record of every bitcoin and every Bitcoin user's encrypted identity is maintained on a public ledger, called the *block-chain*. For this reason, Bitcoin transactions are thought to be *pseudonymous*, not anonymous.

The actors in the Bitcoin network are the *users* who own a *wallet* associated with a couple of private/public cryptographic keys. In Bitcoin, a private key is usually a 256 bit random number, and by using the *Elliptic Curve Digital Signature Algorithm* (*ECDSA*), a 512 bit public key can be obtained from it. Afterwards, from the public key it is possible to obtain a Bitcoin *address*, e.g., applying a hashing function on it. Users use these keys to sign the transactions they generate in order to transfer their money to other users; transactions are then broadcast to the Bitcoin peer-to-peer network.

Miners keep the block-chain consistent, complete, and unalterable: they repeatedly verify and collect newly broadcast transactions into a new group of transactions, called a block. Each block contains information that chains it to the previous block in the block-chain, that is a hash of the previous block. Thank to this field, a block (and consequently the block-chain) is computationally impractical to be modified, since every block after it would also have to be regenerated.

Motivations: Criminals often find ways to exploit legitimate technologies for nefarious uses. Bitcoin is not different: due to its pseudo-anonymity and un-traceability, it is currently used for criminal activities such as money laundering, buying and selling illegal goods or services. Bitcoin has been the de-facto currency of the Dark Web, the "hidden", since the pioneering marketplace *Silk Road*. According to the FBI, Silk Road made a total of 1.2 billion dollars between 2011 and 2013. However, new *Darknet markets* proliferated: these digital markets primarily are black markets, selling or brokering transactions involving drugs, cyber-arms, weapons, counterfeit currency, stolen credit card details, forged documents, unlicensed pharmaceuticals, steroids, other illicit goods, as well as fully-legal products. The life of Darknet markets is often turbulent and short: sometimes they are closed by police forces, sometimes they are cracked, sometimes they perform an "exit scam" stealing escrowed bitcoins. Only in 2014 (last confirmed information), 43 new markets opened and 46 closed. The daily sales volume of six large-scale dark markets reached up to 650.000$ in 2014.

Laundering money using bitcoin is attractive because of its low fees, instantaneous transactions, and virtually anonymous. For instance, bitcoins coming from illegal activities, e.g., proceeds of previously-mentioned markets, can be mixed with "clean" money by using *mixing services*. These services are third parties used to break the connection between a Bitcoin address sending coins and the address(s) they are sent to. The destination address of every mixed transaction receives the same amount of money from possibly many different addresses.

Project: We are applying to the *Microsoft Azure Data Science Research Award* in order to obtain cloud computing resources to support our data-intensive research project. The aim behind our research is to analyze in deep all the scenarios introduced before: the underlying idea is to filter out insignificant information from the block-chain with the purpose to better analyze it to find, for instance, mixing services. The raw block-chain is more than 90Gbyte at the moment, increasing more than 3Gbyte per month, and when represented in a database it occupies some Terabytes. Examining such large amount of data demands for some computational power that can be provided only by Cloud Computing resources.

We have already developed an application based on *Visual Analytics* (VA), which can be tested online at this link: http://www.dmi.unipg.it/blockchainvis/ (type in "user" and "test" as respectively username and password). At a first step, the tool highlights transaction islands, i.e., the sub-graphs disconnected from the super-graph, which represents the whole block-chain. Then it is possible to apply further filters on, e.g., the interval of dates, blocks, transaction values, or number of addresses used in transactions, using VA techniques. Such views highlight specific nodes (e.g., roots and leaves of flows, or miners) and exclude useless and confusing information: e.g., part of the transactions representing changes that will be returned to her can be hidden. A longer draft about such an application can be found here: http://tinyurl.com/gtnxctq.

Hence, we already have some preliminary results for the line of research we are proposing, and these results seem promising. However, we need more computational power than what provided by common hardware: at the moment we are not able to deal with the full block-chain, but only with a third of its size, due to the size of data. This prevents us from designing a deeper analysis, as explained in the following.

By analyzing the block-chain and correlating it with this publicly available meta data, it is possible to find out how much an address is used for e.g., mixing activities, if it was used for scamming users in the past, if and how it is related to other addresses and entities. Addresses can be algorithmically grouped in clusters that correspond with entities that control them. Collapsing addresses into clusters compacts and simplifies the huge transaction graph, creating edges between users that correspond to aggregate transactions. In summary, our current approach aggregates addresses manually, and the whole process is not automated. Automation is what we would like to achieve thank to the power of Microsoft Cloud Computing.