Personal Data Collection and Using Blockchain to Protect Privacy

Joe Putera

18217035

Program Studi Sistem dan Teknologi Informasi

Sekolah Teknik Elektro dan Informatika

Institut Teknologi Bandung

Bandung, Indonesia

-Abstract

In this modern day and age, the internet is an essential part of many's lives. We use the internet to interact with people, to entertain ourselves, to seek information for education, to trade commodities across the world, to look for jobs, and to share information. The internet has essentially become a second world for many of us. But within this new virtual world, dangers lurk its corners as it does in the real world. Many of them aren't readily visible or recognizable, yet they exist still within the depth of the internet. Many online services that one can access for free gathers data of its users. Amassing huge amounts of personal data to be used for their own benefit and to the users' dismay. In this paper, we will discuss why the practice of personal data collection might be detrimental to users of online services, and how blockchain technology might be the solution to this problem.

Key words*: personal data collection, online services, blockchain*

1.Introduction

1.1. Background

Humankind as a whole has developed a dependency on the internet and the devices that connect us to the internet. This phenomenon seems fairly benign to our society if not actually improves it. But as Friedrich Nietzsche once said “If thou gaze long enough into the abyss, the abyss gaze also into thee.” This sentence can apply to many things but the internet in particular embodies this concept really well. Through the internet, one can connect with others and browse for information with ease but at the very same time also runs the risk of exposing oneself to others. This becomes problematic considering that your personal data might be used by irresponsible parties for malicious purposes or their own gain. While most malicious attempts to steal your data can be avoided by simply not accessing the seedier parts of the internet, some reputable sites and organizations may also be responsible for your data leakage.

We are now living in an era where most online services rely on advertising as their main source of income. This is the main explanation for how most online services can be accessed and utilized for free [1]. And in this era of online advertising, a concept known as “targeted marketing” becomes a very lucrative deal for marketing companies to cut advertising costs considerably while increasing their profit. Targeted marketing is done by sending advertisements only to groups of people identified to be interested or more willing to consume services or commodities offered in the marketing campaign [2]. While the idea itself is pretty harmless at face value, the actual method of how potential consumers are identified is where the concern lies. These huge companies would collect extensive

amounts of data of their service users and how they utilize and manipulate this data is completely unknown to the user even though such a process might be detrimental to the users' privacy.

Nowadays, people have been growing more conscious of their rights for privacy, yet big organizations they interact with still amass a huge database of their personal data. What is even more concerning is the fact that these people have little to no power over how their personal data is being used and shared by these big organizations. These organizations may seem reliable and transparent with your data, but recent incidents might prove otherwise. Few of the more known examples of this are Facebook's data leakage to Cambridge Analytica, a privacy breaking bug in Google+'s API which Google hid for around half a year, and Uber's data breach which results in information of 57 million Uber users falling into hackers' hands [3]. While some of these incidents may not seem like a malicious attempt at your data by the organizations' themselves, it does not change the fact that people's datas are at risk without them even being able to do anything about it. But is there a solution available for this?

1.2. Essay's Structure

This paper will discuss why one should care about personal data collection, how blockchain technology might be the solution to this predicament, and speculations of what's to come if no precaution is taken.

Each segment of this paper will be written with the assumption that the reader has no prior knowledge of information security or information technology in general, and aims to explain background knowledge necessary to understand the main topic of using blockchain to safeguard your personal data and privacy. The

language and style of writing used will attempt to explain these technical topics as simple and as easy to understand as possible, as the existence of a writing understood by none would serve no purpose.

The topics covered would center around the usage of blockchain technology to safeguard your personal data. But before tackling the main issue at hand, this paper will first explore the reason why protecting your personal data is important, and then this paper will give ample explanation of what blockchain is and how we can utilize it for securing your personal data.

2. Discussion

2.1. Overview of personal data

As we have mentioned above, personal data is the current day's gold to those who can utilize it. Your personal data is a list of information which consists of anything that is related to you as an individual. Common examples of it includes:

1. Email address
2. Social security number
3. Phone number
4. Birth date
5. Etc

Those are the things that usually come to mind when mentioning personal data in any context [3]. But as we move into a progressively more digital era, your personal data is not limited to physical items or traditional number/letter type of information. When you delve into the internet, you leave traces that can be linked

to you as an individual. Things like social media accounts will show information of your hobby, of what happened in your life through status updates, and of the people you connect with. Online entertainment platforms like Netflix and Youtube can build a profile of you and the things you watch and recommend something uncannily accurate to what you like. These are the traces you leave that allow companies to harvest your data and profit from it via targeted advertising.

Although nothing seems to be too worrisome about online services profiting from their business practices, personal data gathering can inherently cause privacy infringement, personal and professional embarrassment, restricted access to labour markets, restricted access to best value pricing, and many others [1]. Companies may also sell datas they have collected to a third party, leaving your data as an individual exposed. Targeted advertising on the other hand offers a different set of problems altogether, ranging from mild discomfort of users seeing advertisements that uncannily target their needs, to limiting users' diversity of options in the market space due to the targeted nature of the advertising/marketing. We'll delve a little bit into the topic with a real case scenario that has already made the news.

The example we'll take a look at is the Google+'s data leak. In March of 2018, Google found a bug in its Google+ API that allows third-party apps to access private data from its millions of users [4]. In Google+, one can choose to allow third-party apps to have access to one's own public information, private information (information shown only to listed friends), and most importantly the public information of one's listed friends. The bug we mentioned however allows these third-party apps to access not only the public information of one's friends, but also their private information. Allowing third-party apps to learn of users' home address, contact info, and other vital information.

At this point the bug is worrying enough as it is, but the worst part of this case is not the mistake Google had made, it's the cover-up. Google had decided to not publicize this bug even though they later on admit in October of 2018 that they had known of this bug's existence since March of 2018 [4]. At around this time, Facebook also had a data leak scandal of their own and this might prompt Google to hide their problem, waiting for the backlash on Facebook to die down. This event might not be as huge as it was had Google decided to speak to users about this problem once they found it.

One of the many proposed solutions to this predicament is to make use of blockchain to save data in verifiable open ledger securely without any centralized trusted third-party for it. The work that will be explained here uses both blockchain storages and off-blockchain storages to create a personal data management system that focuses on preserving privacy [5].

2.2. Cryptographic Hash Function

Before we can tackle the topic of blockchain which would be the central tool for this paper's proposed solution, we need to explain something a bit simpler. Something that is central to the idea of a blockchain. That something is cryptographic hash functions [6, 7].

Hash Functions are any function that works to map data of arbitrary size to fixed-sized values. Values returned by the function is called a *hash* or a *digest.* For a given input, the hash function will produce a hash which consists of strings of characters. This hash can then be used as a key to retrieve the data of the original input. The hash produced might look like gibberish at a glance, but its real defining

quality is the fact that the usage of hashes shorten the time it takes to look up data although at the cost of needing additional storage place to store these hashes.

Cryptographic hash functions are an even more sophisticated hash function that is suitable for cryptography (the practice and study of techniques to secure communication). The main reason that this function works well with cryptography is because inverting this function to obtain the input from a given output is nigh impossible. Hashes obtained from this kind of function also has the quality of being completely unique for each given input. This is why it is commonly used to secure communications or digitally sign things.

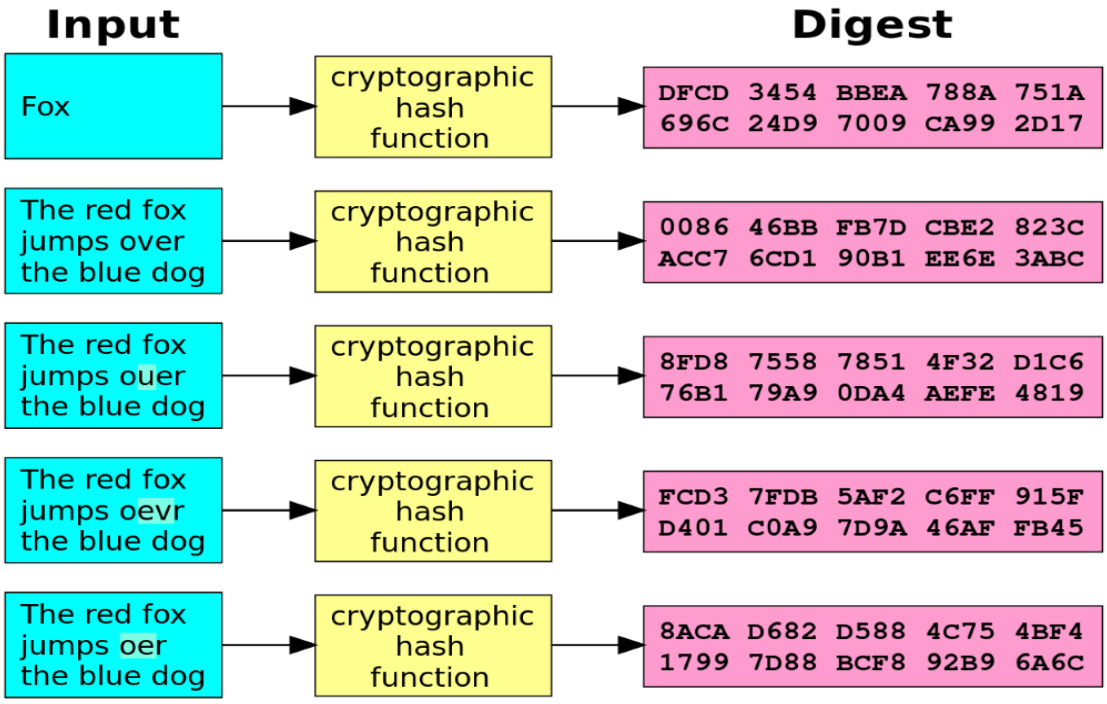

Figure 1. Illustration of how Cryptographic Hash Function works [8]

Cryptographic hash functions output a unique hash for every unique input. The same input would produce the same output, but changing the input even

slightly will significantly change the digest. As reverse engineering the output into an input is almost improbable even with knowledge of the function used, the only viable way to find the message inputted into a hash function would be by brute forcing every possible input to find a matching hash. Brute-forcing is also improbable in some cases such as the SHA-256 hash function which has 2^256 possibilities to force one's way through.

All in all, it is a pretty handy tool for cryptography. But what does this whole hash function thing have to do with our topic? The answer to that you will find in the next segment where we'll explain what a blockchain is.

2.3 Blockchain

Blockchain is currently one of the most talked and discussed topic within the realm of academics. It is the technology which became the basis of what is currently known as *crypto currency*. This piece of technology was first invented by an entity named Satoshi Nakamoto (It is unknown whether Satoshi Nakamoto is a single individual or actually a group of individuals since the name Satoshi Nakamoto itself is actually a pseudonym and the entity itself had long disappeared since 2010), the creator of the widely known Bitcoin.

Blockchain is basically a public ledger able to be accessed by anyone using it. This ledger is then distributed amongst its users and is not kept in any centralized position/address/server. This ledger will then be managed by a network of devices. Each device connected to this network is called a *node*. Every node has the same hierarchy within the network with each and every one of these nodes being able to contribute to the ledger.

Every transaction in the ledger is grouped into *blocks* which will then be run through a hash function. Each block would contain transactional data, the hash, the hash of the preceding block. These blocks will then need a *proof of work* to be accepted as part of the public ledger.

Proof of work is a method used by many crypto currencies to counter spam or DoS (Denial of Service) attacks aimed at the blockchain. A proof of work is essentially a piece of data that is hard to produce but easy to validate. For many blockchains, a proof of work is needed to be able to submit a block to the public ledger. A blockchain *miner* would then try to find this proof of work. The first miner to find the proof of work is then rewarded with the ability to authorize the transaction and in return for their service, they will earn a small amount of cryptocurrency. Transactions recorded into a block would be added into the main blockchain ledger where it would then be broadcasted to every other users, allowing them to update their ledger to its most updated form.

In the event that there are 2 different versions of a blockchain which conflict with one another, the blockchain with the longest chain (one which contains more blocks) would be accepted as the "correct" ledger. In this way, tampering with any part of the ledger would require you to tamper with every block following after it since a single change in the ledger would alter the hash of the block which would then not match with the hash store in the next block about the previous block. The blockchain "forger" would then be pushed out of the system if they don't have at least more than 50% of the blockchain network's processing capability since the forger wouldn't be able to compete for the longer blockchain otherwise. In this way, a blockchain will actually be more resistant to tampering the longer its chain

is. This means that blockchains are inherently almost invincible to any kind of tampering.

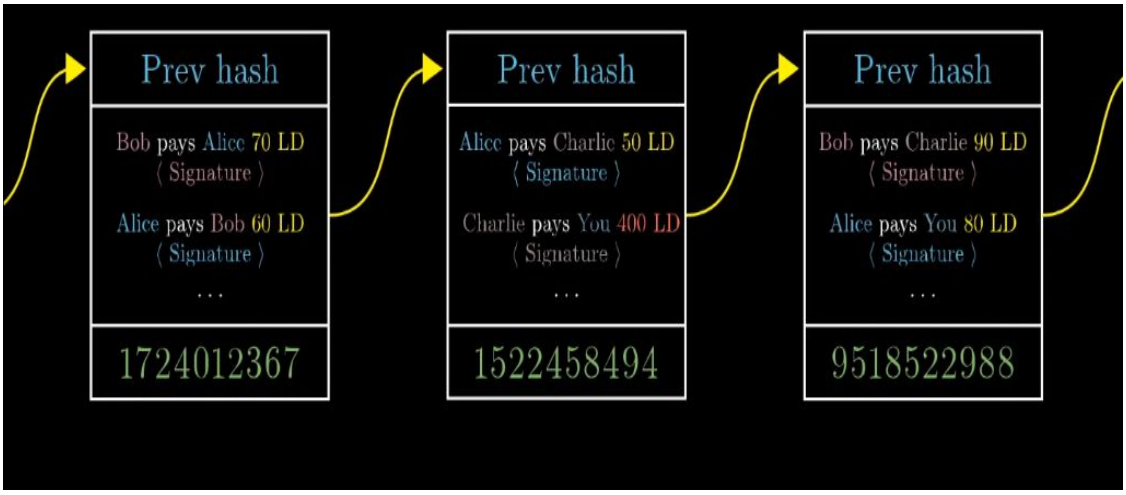

Figure 2. Illustration of blockchain [9]

Blockchain also uses cryptographic techniques to protect the information of its users where users are represented as a virtual address which correspond to a public key generated by the user. This enables users to retain their privacy to a high degree. This is also the reason why plenty of shady dealings usually use crypto currencies as their trading currency of choice.

As mentioned above, blockchain technology offers plenty of advantages. The main benefits it gives though are its resilience to a system-wide failure since every node within the network has a copy of the ledger, its resistance against fraud in transactions since forging is very improbable due to blockchain's trust algorithm, and its independence from using a centralized trusted third-party (e.g. the bank when transferring bank account balance) [10].

These advantages unfortunately do not come without some drawbacks attached to them. One of the more vital flaws of this is the efficiency problem blockchains have. The process of finding a proof of work is basically a lottery game that consumes plenty of time. This process limits the amount of transactions the whole network can do in a given amount of time. The amount of resources needed just to find a single proof of work (like the electricity needed by miners' processors for example) is relatively large. Storage would also be a problem since these blockchain ledgers can grow in size really fast even though every node needs to have a copy of the latest one. For example, Bitcoin's ledger size as of the 22th of April 2020 is 255 Gigabytes [11].

2.4 Proposed Solution

Throughout this paper, we have explored why and how important data privacy is and also explored the necessary background knowledge to understand the system Guy Zyskind, Oz Nathan, and Alex 'Sandy' Pentland proposed as a solution to our privacy issues. The system proposed will be explained with the use of mobile platform as an example but it can definitely be implemented for other platforms.

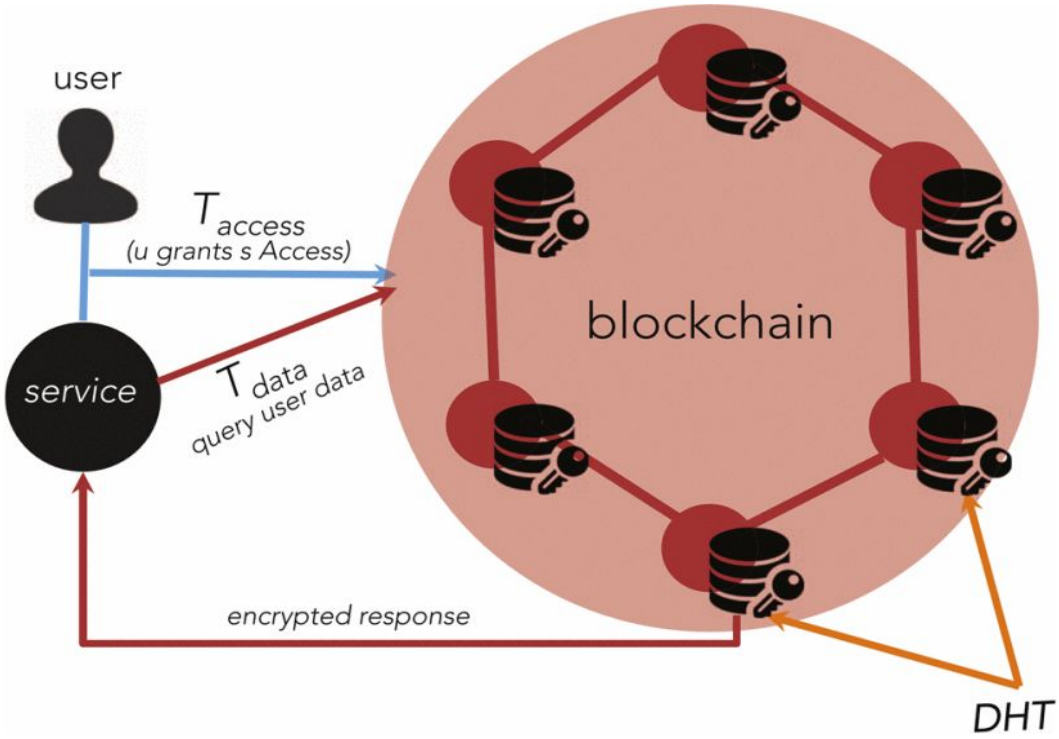

Figure 3. Illustration of Proposed System [5]

We will begin with an overview of the proposed system. As illustrated in Figure 3, the three entities comprising our system are mobile phone users, interested in downloading and using applications; *services*, the providers of such applications who require processing personal data for operational and business-related reasons; and *nodes*, entities entrusted with maintaining the blockchain and a distributed private key-value data store in return for incentives (small amount of cryptocurrencies as mentioned before). Note that while users in the system normally remain (pseudo) anonymous, we could store service profiles on the blockchain and verify their identity [5].

The system is designed to accept 2 types of transactions which are:

1. Taccess, used to manage access control
2. Tdata, used for data storage and retrieval.

To further illustrate the process, consider the following example. A user installs an application that uses this platform to preserve their privacy. As the user signs up for the first time, a new shared identity is created and sent with associated permissions to the block chain in a Taccess transactions. Data collected on the phone (e.g., sensor data such as location, fingerprint, voice, etc.) is encrypted using a shared encryption key and sent to the blockchain in a Tdata transaction, which will route it to an off-blockchain key-value store, while only keeping a pointer to the data on the public ledger (the pointer is the SHA-256 hash of the data) [5].

Both the service and the user can now query the data using a Tdata transaction with the pointer(key) associated to it. The blockchain then verifies whether the digital signatures belong to the user or the service or not. For service, its permission to access data to access data is checked as well. The user can then change the permissions granted to the service anytime the user wishes to do so by issuing a Taccess transaction with new sets of permissions. Developing a mobile dashboard that gives an overview of one's data and the ability to change permissions becomes an easy task to do and is akin to developing centralized-wallets, like Bitcoin's Coinbase [5].

The off-blockchain key-value store is an implementation of Kademlia [12], a distributed hash table, and interface to the blockchain. The hash table is then maintained by a network of nodes, who will approve read/write transactions. Datas are randomized across the nodes and then replicated to ensure that it is always available.

3.Conclusion

In general, personal data shouldn't be entrusted to other entities you cannot account for. Especially so if they are handling your personal data irresponsibly or maliciously misusing it. The responsibility of keeping one's privacy and personal data should fall in their own hands and it is imperative that proper measures must be independently taken by individuals to safeguard their privacy.

If no precautions are taken, we might be looking at a bleak future where freedom of expression and privacy might be forever muted. Such a thought would seem pessimistic and somewhat delusional. Almost akin to plots and settings of dystopian fictional stories. But if nothing is to be done with our current situation, those fictional tales might as well be our apparent destiny. As more big organizations gain power, their control over our daily life increases.

The proposed solution that we discuss here might be the answer to our current predicament and the key to a brighter future within the internet space and the real world. This solution is not without its own set of downsides as the consequence of large scale implementation has yet to be seen and can only be somewhat predicted. But blockchain technology seems like one of the more promising solutions and will likely continue to be so for quite some time.

Reference

1. Chaudhry, A., Crowcroft, J., Howard, H., Madhavapeddy, A., Mortier, R., Haddadi, H., & McAuley, D. (2015). Personal data: thinking inside the box. In Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives (CA '15). Aarhus University Press, Aarhus N, 29–32.

2. Targeted Marketing (n.d.). accessed 20 April 2020. https://www.marketing-schools.org/types-of-marketing/targeted-marketing.html.
3. Data Privacy Concerns: An Overview for 2019 (2019). accessed 20 April 2020.https://medium.com/@the_manifest/data-privacy-concerns-an-overview-for-2019-2ccea79aa6f8.
4. Cyphers, B., & Gebhart, G. (2018). The Google+ Bug is More About The Cover-up Than The Crime. accessed on 21 April 2020. https://www.eff.org/deeplinks/2018/10/google-bug-more-about-cover-crime.
5. Zyskind, G., Nathan, O., & Pentland, A. (2015). Decentralizing Privacy: Using Blockchain to Protect Personal Data. 2015 IEEE Security and Privacy Workshops. San Jose, CA. 2015. pp. 180-184.
6. Halevi, S., & Krawczyx, H. (2007). Strengthening Digital Signatures via Randomized Hashing. https://webee.technion.ac.il/~hugo/rhash/rhash.pdf
7. Knuth, D. (1973), The Art of Computer Science, Vol. 3, Sorting and Searching, p.527. Addison-Wesley, Reading, MA., United States.
8. Cryptographic Hash Function (n.d.). accessed on 21 April 2020. https://en.wikipedia.org/wiki/Cryptographic_hash_function.
9. 3Blue1Brown (2017). But how does bitcoin actually work?. accessed on 20 April 2020. https://www.youtube.com/watch?v=bBC-nXj3Ng4&t=1047s.
10. Blockchain Advantages and Disadvantages (n.d.). accessed on 21 April 2020.https://academy.binance.com/blockchain/positives-and-negatives-of-blockchain.

11. Blockchain Size (n.d.). accessed on 22 April 2020. https://charts.bitcoin.com/btc/chart/blockchain-size#5ma4.
12. Maymounkov, P., Mazieres, D. (2002). Kademlia: A peer-to-peer information system based on the xor metric in Peer-to-Peer Systems. Springer. pp. 53-65, 2002.