CyberVein

A Dataflow Blockchain Platform

Whitepaper - Version 1.0

Jack Ge, Jerry Ning and Arthur Yu

whitepaper@cybervein.org

May 9, 2018

Executive Summary

With humanity moving deeper into the Information Age, modern societies produce increasingly large amounts of data, while becoming ever more dependent on its on-demand availability, as well as its integrity. As individuals, organizations and even Nation States, we grew accustomed to and reliant on our ability to query information on any subject, at any time, and to freely use it as we see fit. In return, as a society we have de facto agreed to tolerate an increasing degree of transparency in terms of the metadata that describes our private and public lives.

This pay-off arrangement between data producers, accumulators, and consumers has proven itself immensely instrumental in terms of economic growth, public discourse, and the overall improvement of our quality of life. However, above this “New Deal of Data” looms the shadow of an inherent market failure that prevents our Information Economy from reaching its full potential, while allowing actors in privileged positions in the data economy to exploit our dependency on information to their own advantage.

While one side to this “New Deal of Data” finds itself subjected to an automatically and increasingly involuntarily extraction of information through countless applications, sensors, and algorithms dispersed throughout Cyber- and “Meat”-space alike, the other party to the deal expects to be blindly trusted when it comes to the handling of this enormous pool of information. Even more significantly, following a severe lack of economic incentives, the most valuable pool of data in existence has never actually reached the public domain to begin with. This pool includes the results of countless academic research projects, as well valuable economic data produced and accumulated by private and public entities alike.

As a result, the overwhelming majority of the fuel that is supposed to power our Information Economy remains siloed in isolated databases, where it is prone to malevolent manipulation, and where its economic and cultural potential remains unutilised, if not downright weaponized.

At first glance it appears that Blockchain Technology is an ideal tool to solve both of the problems stated above: the questionable trust-relationship between data producers, handlers, and consumers; as well as the lack of economic incentives to maintain vast, resource intensive data sets in the public domain.

In terms of facilitating trust and data-integrity, Distributed Ledger Technology has already proven itself as immensely useful. The trillion dollar Bitcoin, Altcoin, and token markets have demonstrated beyond any reasonable doubt that it is possible and useful to store crucial information publicly, while ensuring that its integrity remains intact without any authority enforcing cumbersome security measures.

The same crypto-economy with its tokenization schemes has also demonstrated that through game-theoretical means, the pooling and sharing of computational resources as well as information

can be elegantly incentivised with spectacular results. Alas, to date fundamental technological hurdles remain which prevent blockchain technology from playing a significant role in terms of securing the integrity of complex databases, while incentivising the pooling and sharing thereof.

Blockchains were originally introduced to secure a single and fairly simple data stream, namely the sequence of monetary transactions in a peer-to-peer (P2P) cash system. These transactions are recorded in a monolithic public ledger, which is just that - an ever growing account of who sent what to whom. This account can then serve as a "single source of truth" on which all network participants can blindly agree upon.

Attempting to expand this "single source of truth" concept to complex databases that are capable of recording and processing the vast amounts of data produced by 7.6 billion humans and their cultural and economic activities would, for obvious reasons, result in a ghastly failure if based on contemporary blockchains such as Ethereum and its progenies. To accommodate for such a revolutionary new implementation of blockchains, CyberVein proposes a radically novel approach to Distributed Ledger Technology, one able to record multiple datastreams in structured, highly complex decentralized databases which can be processed, queried, and manipulated in real-time by several parties in a trust-less global network. To achieve this, CyberVein improves on Directed Acyclic Graph architectures by introducing a resource conserving consensus algorithm, as well as a smart contracting language and virtual machine, optimised for the handling of massive amounts of data. Additionally, the CyberVein contracting language comes with built-in monetization functions which allow owners of data to define its value and to condition access to it through direct payments or any other way they see fit.

CyberVein databases are stored and maintained as independent smart contracts, permissioned to contract parties, yet potentially accessible to other network participants. All entries, modifications, and amendments to CyberVein databases are stored as appended smart contract transactions. This means that a CyberVein database automatically contains all previous versions of itself, expanding immutability to the temporal dimension.

Consequently, the CyberVein network is built as a fractal network of independent yet interlocked databases on which users only sync data that is relevant to them, protecting the network from congestion.

The result is a public network of immutable databases, comparable to the Internet as public network of networks. On this network information is protected from tampering, just like Bitcoin transactions on the blockchain. No one, including the owners of a database, can corrupt its records, delete it, or tamper with its processing history. This allows data to be securely interconnected, and transformed into structured knowledge, while economic incentives for data sharing are baked into the blockchain protocol itself.

With this CyberVein creates an entirely new market environment in which data sharing among competing entities is not only made possible, but economically incentivised through crypto-economical and game-theoretical means. With this, CyberVein allows for the maintenance of publicly accessible data-sets which can be used by anyone and corrupted by no one, functioning as the infrastructure for increasingly data-dependent information economies and smart infrastructures.

1 Introduction

It is said that information is the oil of the 21st century—which is as right as it is wrong. We often refer to this quote when talking about the large and glamorous Silicon Valley companies with their enormous user bases. Nevertheless, the restricted purpose of these platforms is targeting users for influencing buying or voting behaviour. Compared to oil, data gains its value from being diverse and highly interconnected, whereas the quality of oil instead lies in uniform absence of impurity and is independent of the amount stored.

Besides social media platforms, that are actually advertising and influencing networks, countless smaller databases exist. They often contain very specific knowledge that is valuable to specialised groups or geographic locations, e..g, genetic categorisation of species that occur on a specific archipelago.

Apart from so called KYC (Know Your Customer) functionalities that are often used in the financial industry as well as a process aware smart contract language, the blockchain industry is especially missing means for storing and consuming structured data. Key-value and binary-blob mechanisms lack the feature of storing relations between entities and navigating them. A more dynamic aspect is the continuous creation of data that flows through a network, for example vehicles populating the road network of a city or supply voltages in a power grid.

An inherent problem with big data is the analysis of data chunks. Currently there is no easy way of sharing data sets and joining computing power in order to process them. In blockchain, there are two conflicting goals regarding sharing data sets:

- In order to secure data on the blockchain, it needs to be shared on a sufficiently large number of nodes in the network.
- Keeping a blockchain's history lean and clean requires storing as little data as possible.

From the technical point of view, CyberVein is enhancing the crypto space with an important dimension that it is still missing: a data centric paradigm that acknowledges industries in the digital age are driven by vast amounts of structured data.

In the remainder of the section we discuss use cases of CyberVein, ranging from *research*, *smart cities*, *supply chains*, *dataflow networks*, *artificial intelligence* to *DNA sequencing*. For each use case, we outline the current state of the art, its downside and how CyberVein can help to solve current issues for each specific use case.

1.1 Academic Research

Scientific discoveries are the result of the interplay between observation of specific events and formulation of theories to explain them beyond their concrete occurrence. New and more precise technology has vastly increased our means to collect data, for example satellites for large scale examination of weather phenomena [2] or desertification [11], but also particle decay in particle accelerators [18].

Theories can only be validated with conclusive data which makes precise measurement inalienable. That in turns leads to an increase of required data storage capacities. Storing the data is only one aspect for research. Evaluation of collected data and creating models that can be simulated is equally important. Most of the bigger universities have mainframes available for such tasks [40].

International large scale research collaborations also post several further challenges. Besides communication, the success of a research project also depends on sharing results as well as sufficient computing power to obtain meaningful results. The more extensive a project, the more important it is to distribute computing tasks among the participating research institutes. There could also be the case where an institute does not provide computing power but very specific expertise to a collective effort but still depends on another participant to provide the computing resources.

Another challenge for researchers is the validation of findings presented in academic papers. Apart from making the data permanently available, it also has to be stored in an immutable manner to prevent data manipulation. Immutability is of exceptional importance in areas like climate and pharmacological research where scientific rigor and economic interests often contradict each other [4][14].

Besides some weakly enforced internal university guidelines, no apparent incentive exists for researchers to maintain, store and publish data used in scientific publications so that others can use it to verify their findings long-term[15]. Ideally, a research paper would contain a hash of the data set used. With continuous research, not only immutability is important but also access control and tracing the history of updates to the data set.

Recently, the research fields of Psychology and Pharmacology are facing a reproducibility crisis where previous findings could not be verified by reproducing findings of prior publications [3][35]. A direct negative effect is the loss in value and insight for the scientific community where recent results from other publications are used as building blocks to support their own line of reasoning. Apart from that, academia has developed into an economy where too much attention is paid to the quantity—instead of the quality—of the published material [36]. CyberVein provides a new approach to establish publishing data used in research papers as a new baseline for empiric research by providing the required technological and economical framework. Publishing the data used alongside the research paper will allow for more in depth analysis when attempting to reproduce results and comparison to the steps taken during statistical analysis.

1.2 Smart Cities

As a side product of the progressing digitalization of our daily lives as well as the growth of the Internet of Things (IoT) [30][34], smart cities, that are managed based on electronic data collected by sensor become more relevant. Reliable statistical data enables cities, municipalities and local authorities to discover trends and their causes as well as imbalances and problems [21].The more precise the information, the easier it is for authorities and decision makers to allocate resources where they have the biggest and most positive impact. This pertains to more static, i.e. comparably slower changing, information like cost of living, energy/water supply, real estate prices and public transportation but also to more dynamic data like status of the power grid, traffic congestion during rush hours, air quality and public transportation delays.

Current challenges are synchronisation between authorities and unification of data sets in order to gain a common view of information forming the current state of the city. The bigger a city and the stronger its growth, the more important it is to channel and orchestrate forces acting on it, in order to maintain infrastructure throughout the city's area. Infrastructure is also one of the key factors that decide whether or not companies and highly specialised experts are attracted to a region. Therefore data is not only necessary for maintaining infrastructure but also to develop it in a way that is consistent with the anticipated and forecasted development of the city in a cost efficient way.

Synchronising and sharing data in a reliable way would not only benefit the authorities of a city but also companies to find the most suitable site. Above that, other cities can learn from the data of a city that successfully applied a new concept or technology or managed to handle a very specific challenge.

Furthermore, in order to automate processes and facilities within an area with separate and standalone devices, information needs to be collected from and instructions sent to them. Storing information in a robust way and allowing for recovery in case of failure are prerequisites for large scale automation. A power grid, for example, has to be controlled so that equipment and facilities are not damaged. Even if an incident can not be prevented, the collected data might help to reveal its cause afterwards. From an automation viewpoint, the city can be perceived as a feedback system that incorporates external data in order to keep certain variables like water flow rate or line voltage in predefined ranges. The upper bound for reaction times is the latency of the used technology. Latency is an inherent feature of distributed systems. For large systems to be functionally effective and efficiently maintainable its components need to be erected on the same foundation/architecture.

While IoT blockchains like IOTA [37] already exist, they only focus on a very limited use-case and do not support CyberVein's sidechain features. Besides that, IOTA is missing smart contract capabilities. CyberVein's blockchain system provides the perfect platform for authorities and companies in the realm of IoT and smart cities that require a blockchain-based solution.

1.3 Supply Chain

While the term *value chain* refers to all processes that lead to the creation of a product a company is selling, *supply chain* describes a connection of value chains: Beginning with the raw materials to the customer who is buying the final product [9]. Complexity arises from the number of participants in a supply chain and their need for economical employment of resources. Since unused storage capacities impose fixed costs, companies seek to optimise the utilisation of those already available. This requires planning ahead so that resources are available when needed while avoiding extended storage of unused material and final products. Making correct decisions in this setting is highly dependent on reliable information from participants up- and downstream along the supply chain.

A paradox aspect arises from the need to protect trade secrets, e.g., valuable suppliers, fast, cheap and reliable transport companies, in an industry where collaboration and information exchange are prerequisites. In order to have goods delivered to the next party in the supply chain, information on the starting point and the extent of the transfer needs to be shared with - at least - the transportation company. The far more common case is outsourcing information and process management to 3rd or 4th party logistics (3PL and 4PL) companies—it becomes unclear how data is protected against leaks on their platforms [27], with whom it is shared and how it is used by the 3PL/4PL company [33]. Ideally, necessary information could be passed along the supply chain without being rerouted through 3rd/4th party systems and losing control over who can access the data.

Another problem in supply chain management is the use of paper documents to provide necessary information to companies downstream. While handling paper documents (e.g. confirming the transfer of a cargo container from ship to truck) already slows down the process, additional delays occur when documents are lost or handled incorrectly. A single point of truth where data is stored immutably also improves trust among involved parties. Currently a very questionable form of 'immutability' results from the fact that data is stored in the information system of one participant that all others have to trust (3PLs/4PLs).

While many view blockchains as the next step of supply chain management [29] most solutions still have privacy issues and other problems. CyberVein's isolated sidechain concept ensures that data is only shared with qualified partners while providing a tamper-proof storage at the same time.

1.4 Dataflow Networks

The term Big Data [22] is often associated with large chunks of data being statically processed in data centers. Large volumes of data indeed arise from a multitude of data sources transmitting information perpetually. This dynamic and temporal quality is related to the connection between data and information. Big data always seeks to discover structure and insight (information) that is assumed to be implicitly present in the data set, only waiting to be discovered [16][43]. The value of information depends on its up-to-dateness which in most cases decreases over time. For example, a two days old weather forecast is of limited importance to an airline.

A characteristic of dataflow networks is the perpetual creation of data by users and their lack of control over these since control might be exercised by third parties siloing the data. Smartphones create data through applications that collect information [7], cars collect information on location but also through readings of internal sensors [26] and browsers create data that is stored in cookies. Unfortunately, for the user it is not clear what kind of data is collected and how (transparency). In addition, it is also unclear which third parties get access and might control its creation. Apart from privacy issues, the companies collecting this data use it as a revenue stream without the user receiving finacial benefit. CyberVein provides an infrastructure where no single entity has exclusive control over a data set and where valuation of information is a built-in feature. Wide range adoption of the technology for this use case depends on users putting emphasis on transparency, thus forcing owners of large amounts of personalised information to adept.

1.5 Artificial Intelligence

Artificial intelligence is becoming increasingly important. Most people might think of AlphaZero [39], IBM's Watson and DeepBlue [8][20] as well as speech recognition [19]. Nevertheless, knowledge representation, expert systems and multiagent systems belong to this category as well [38].

Training a model for classification or prediction requires data. Before data can be fed into an AI algorithm, it needs to be preprocessed, including reduction of dimensionality.

Although not being a new concept, neural nets gained popularity under the term deep learning recently [24]. Neural nets are composed from layers that contain neurons. Each neuron of one layer is connected to every other neuron of the layers above and below by weighted edges. As this constitutes a graph structure, it is especially suitable for being stored in a graph database. The same applies to the data that is used for training the neural nets (models)—it is highly structured.

Besides structure, the size of training data sets and models poses a challenge, current blockchain systems are unable to meet. Using a distributed system, training data and models can be shared and updated more easily.

Another area of Artificial Intelligence that depends on structured storage of data are ontologies. These are directed graphs containing concepts, relations and individuals. Typical relations between concepts are specialisation and generalisation. Individuals are instances of concepts and further knowledge about them can be inferred by the relations of the concepts. For health care expert systems, ontologies are used to represent knowledge about diseases and symptoms in order to make much more precise diagnoses. For some diseases, some symptoms contradict each other in a very specific way, for others, certain symptoms always occur combined. Keeping all possible combinations in mind is impossible for physicians but feasible for IT systems.

Ontologies are also useful for multi agent systems [17]. In a multi agent system, several agents try to achieve a common goal, e.g. robots moving goods in a warehouse, trying to fulfill a common schedule. In order to coordinate their efforts and maintain a common vision of the overall state of the environment (including agents), these agents need to exchange queries, information and instructions in a common format. Here, ontologies enable the modelling of discourse specific knowledge like domain entities, services, interaction protocols and workflows in an explicit and programming language agnostic way.

CyberVein supports AI developers and researchers with distribution of ontologies, models and data, models are trained with, as well as allowing for access to them in exchange for tokens.

1.6 DNA Sequencing

The process of reading nucleotides from a DNA strand in correct order is called DNA sequencing [28]. The fact that DNA encodes enormous amounts of data is nicely illustrated by a printout with the smallest font possible of the human DNA containing 100 books consisting of 1000 pages each[1]. The significance for researchers of DNA sequencing lies in the opportunities to understand diseases in connection with genes from which they emerge. This entails comparison of DNA strands as well. It is still a challenging task to store such large data sets in a distributed system, allowing for parallel processing of the same data set, as well as means for appending amendments to it, in order to keep information up-to-date.

CyberVein's advantage is the ability to share intermediate results which otherwise would require a node to do the very same computation another one already performed - this reduction of redundancy becomes more relevant the more time a computation takes and the longer resources are therefore bound.

2 Prerequisites

The CyberVein platform employs established formalisms and techniques to facilitate tokenisation of data sets in a distributed setting. In the following we will introduce key concepts that enable the CyberVein platform and explain them in detail. First, Section 2.1 focuses on blockchain technol-

[1] https://wellcomecollection.org/articles/visit-kettles-yard-and-wellcome-library/?image=2

ogy, followed by general introduction of MapReduce in Section 2.2. Afterwards, Section 2.3 and Section 2.4 focus on directed acyclic graphs (DAGs) and graph databases.

2.1 Blockchain Technology

The central idea behind blockchains is storing and modifying data in a network of nodes. Data is synchronised among nodes which function as an automated recovery mechanism in case nodes drop out of the network temporarily.

As the name suggests, a blockchain consists of a chronologically ordered chain of blocks. Every block consists of a certain number of validated transactions and each of those blocks links to its predecessor by a hash reference. As a result, changing the content of one block also changes all succeeding blocks and hence breaks the chain. All blocks are stored on and verified by all participating nodes. While the initial Bitcoin blockchain only supported a very limited set of scripting instructions. Competitors such as Ethereum [42] and Tezos [23], offer Turing-complete programming languages on the protocol-layer level that enable so called smart contracts. Smart contracts are, "orchestration- and choreography protocols that facilitate, verify and enact with computing means a negotiated agreement between consenting parties" [12]. Thus, the parties participating in the enactment of a smart contract establish binding agreements and deploy applications using such smart contracts in order to provide blockchain-based applications. Applications and services enabled by smart contracts include the finance sector [32], academic and business authentication and identity solutions [5][10][25] as well as reputation systems [6].

A major advantage of blockchain technology is the maintenance of a consistent state following an agreed upon consensus protocol. To date the blockchain industry is missing sufficiently expressive programming languages that cover process awareness and parallel execution as well as means for storing data in a structured way, let alone in large amounts.

2.2 MapReduce

MapReduce [13] is a framework that simplifies and increases performance of processing large data sets when the algorithm to run on the data set can be expressed as a map and a reduce operator. Both are common terms in functional programming, especially in the family of dialects known as Lisp. A key concept of functional programming is referential transparency of expressions. Since the result of a function only depends on its arguments and not on global variables or any kind of state or context, function calls can be evaluated in parallel—either in sequence on the same main processing unit or on a different node of a cluster or network.

In data centers, large data chunks are naturally stored on several nodes of a cluster. The ideal way of processing a data set is to ensure that all parts of a certain data set reside on the same cluster instead of different machines. This is usually done for tasks where processing one unit or chunk of the data set does not require read access to the rest of the data set, e.g., when counting words in a document or base-pairs of a DNA strand. They can be split up into smaller parts which are then processed in parallel. Counting the frequency of words in a paragraph will always yield the same result, independent of the rest of the document (as is for DNA).

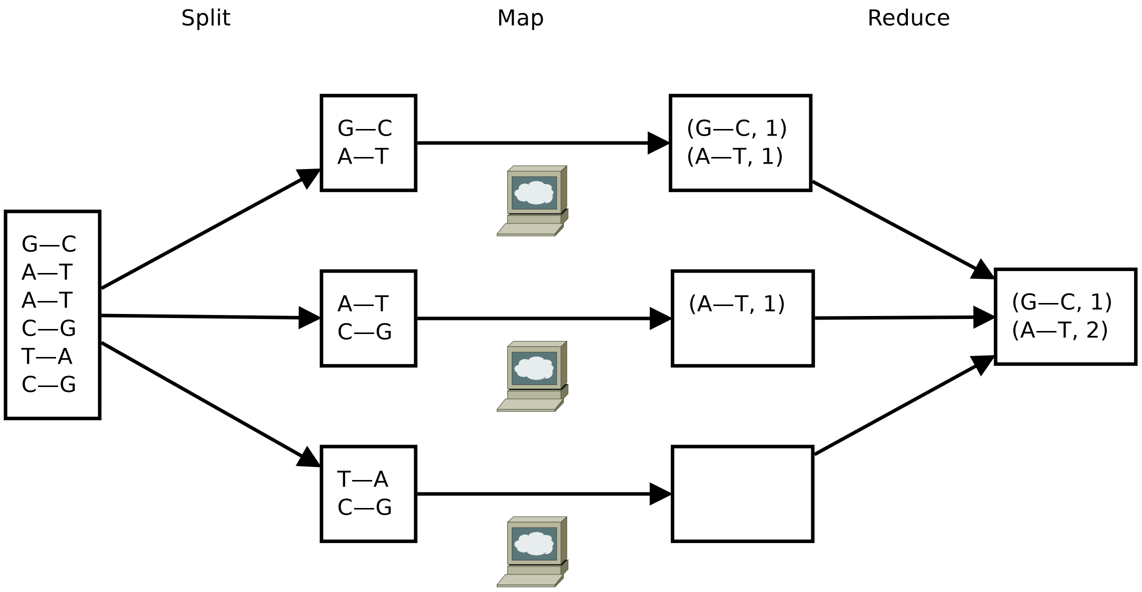


Figure 1: Illustration of a map-redeuce task by counting DNA base-pairs.

Figure 1 illustrates a map-reduce example using the task of counting occurrences of the base-pairs `"A-T"` and `"C-G"` in a DNA strand. In the *split* phase, the data is divided into chunks so that each node has a piece of data to work on. In the *map* phase, a task-specific function is mapped on each record of the chunk a node received. During the *reduce* phase, the results from each node are collected and "summed up". The base pair `"A-T"` occurs at the first and the second node, `"C-G"` only at the first. Both occurrences of `"A-T"` are reduced into one with the number `"2"` indicating the amount of occurrences. In a certain sense, *reduce* can be seen as extraction of cross-cutting concerns.

CyberVein nodes can join their computational power for running algorithms on a data set as distributed system and using the CyberVein's MapReduce functionality allows to either process a data set which is split up into parts where each node processes one, or execute several processes simultaneously on a single data set. Each node runs one of the processes on the whole data set.

2.3 Directed Acyclic Graphs (DAG)

Graphs consist of vertices and edges connecting them. A very simple example of a DAG are Bitcoin-like proof-of-work blockchains: each block (vertex) points at its predecessor by storing the predecessor's hash (edge)[2]. *Directed* and *acyclic* means that an edge between two vertexes can connect the both only in one direction—otherwise the result is a cycle where it is no longer possible to determine which is the more recent block. Figure 2 illustrates the representation of a blockchain via graph system. The chain, e.g. Bitcoin blockchain, has three vertices and two edges. The second block has hash `TT81VZ` and stores—besides transactions and other information—the hash `X2T34H` of the genesis block. This can be interpreted as an edge from the second to the genesis block. The edge from the 3rd block to the genesis block, indicating the 3rd block is also more recent than the genesis block, is usually omitted, since it is intuitively clear from the fact that it is already newer than the 2nd block.

In the context of blockchains, directed acyclic graphs [41] generalise the notion of a chain (all transactions in sequence) while maintaining an ordering of blocks regarding the time line. Figure 3 shows transactions constituting a directed acyclic graph. Although, for example, transaction `PEIL4J` cannot directly be compared to `EEKUU3` (in terms of age; on the time line), since there is no directed edge connecting them, they are still ordered. Regardless of randomly picking a transaction

[2] More precisely, Bitcoin's blockchain is not only a directed acyclic graph (partial order) but also satisfies the stronger condition of a total order.

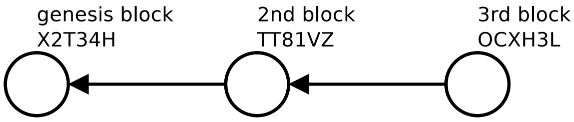


Figure 2: Graph representation of a blockchain.

and following the directed edges, one will always arrive at the genesis transaction. As demonstrated by the totally ordered blocks of Bitcoin, transactions stored in a directed acyclic graph convey a notion of time passing by. Although not all transactions (circles) of Figure 3 can be compared to each other directly (i.e. not being connect with a directed edge) - which would result in a chain like Bitcoin - traversing the graph by following directed edges (arrows), will always lead backwards to the genesis transaction, thus conveying the notion of time (like in Bitcoin).

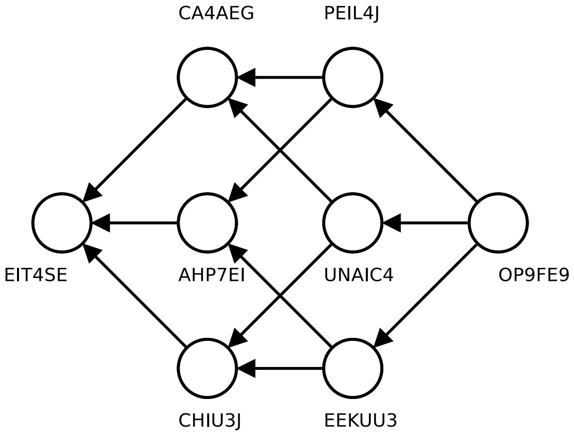


Figure 3: A blockchain-based DAG.

Similar to the IoT blockchain IOTA[37], CyberVein uses also a DAG-based data representation instead of a traditional block-based representation as introduced by Bitcoin[31].

2.4 Graph Databases

Graph databases have experienced wide range adoption recently as an alternative to traditional relational databases that store information in tables. In relational databases, entities are modelled as rows, properties as columns and relationships with foreign keys or separate tables. In comparison, a graph database stores entities as vertices with properties attached to them. Relations, including attached properties, are expressed as edges in the graph. Observe in Figure 4 that the very same information can be captured with both representations. Compared to tables (although holding the same information), domain knowledge is much easier to grasp for the eye when depicted as graph.

Another advantage of graph databases stems from avoiding *join* operations. In order to derive implicit information from a table database, these are often necessary. It consists of making the crossproduct of two tables and applying a join-criterion to filter out irrelevant records. Figure 5 shows how to determine that Changsha and Beijing are indirectly connected via Shanghai, e.g., by a one-way rail track. Intuitively, this is achieved by finding those one-way connections between two

A.orig	**A.dest**
Changsha	Shanghai
Shanghai	Beijing

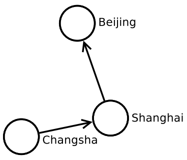


Figure 4: Both representations, table and graph, capture the exact same information.

cities where the destination of the first is the origin of the second. The large table in the bottom right is the result of combining each row from the bottom left with the each row of the top right table (Figure 5).

B.orig	***B.dest***
Changsha	*Shanghai*
Shanghai	*Beijing*

A.orig	**A.dest**
Changsha	Shanghai
Shanghai	Beijing

A.orig	**A.dest**	***B.orig***	***B.dest***
Changsha	Shanghai	*Changsha*	*Shanghai*
Changsha	Shanghai	*Shanghai*	*Beijing*
Shanghai	Beijing	*Changsha*	*Shanghai*
Shanghai	Beijing	*Shanghai*	*Beijing*

Figure 5: The large table is the result of the cross product of the two smaller tables. The join criterion is applied to the larger one in order to find that *Changsha* and *Beijing* are indirectly connected via *Shanghai*.

Observe that the second row (excluding the row with column names) is the only one where the destination of the first one-way connection (`A.dest`), `Shanghai`, coincides with the origin of the second one-way connection (*B.orig*; *Shanghai* as well). Although obvious from an intuitive point of view, still the database system has to make the cross product first, otherwise no rows would exist to which the join-criterion can be applied.

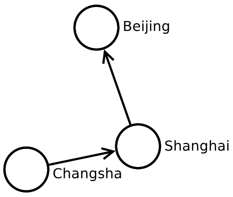


Figure 6: Single graph representation.

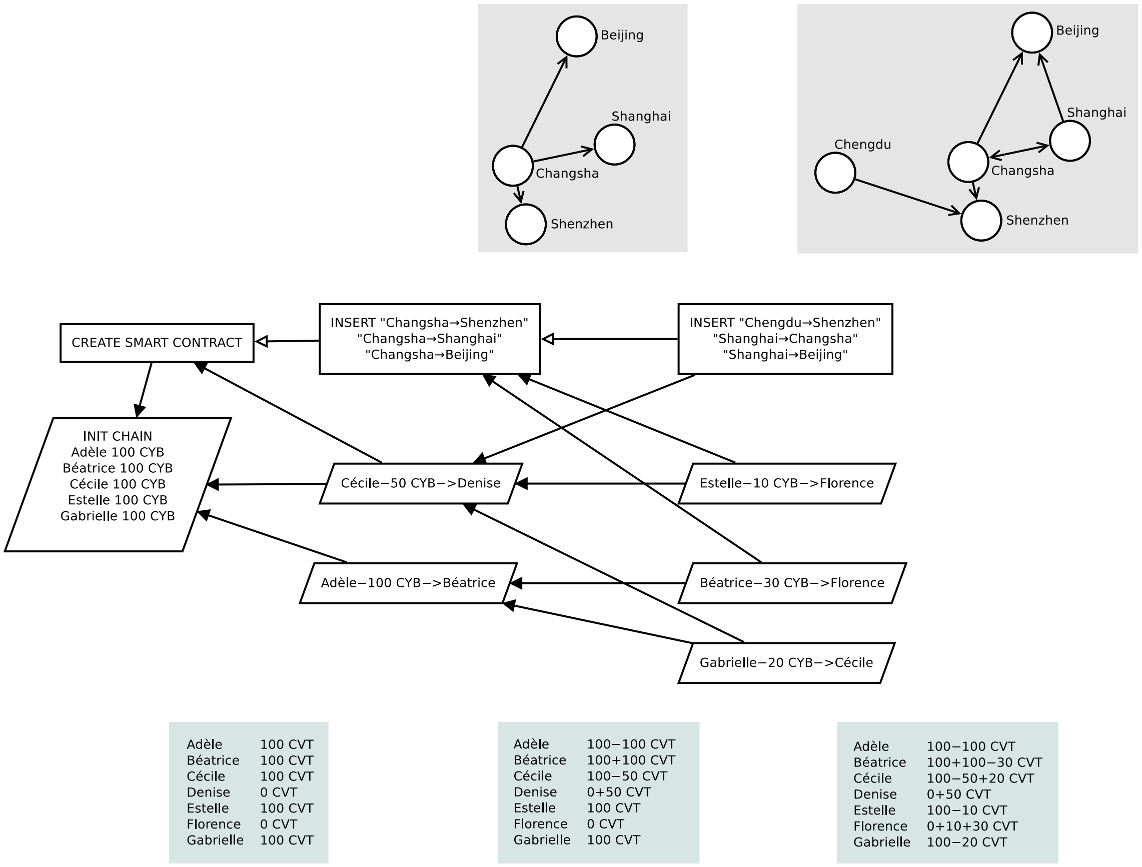
The graph representation in Figure 6, reduces the same information into a graph with only one directed edge from *Changsha* to *Shanghai*, and from there only one leads to *Beijing*. This way the redundancy of sorting out three irrelevant rows can be avoided. Graph databases excel when performing recursive queries or when single entities and their relationships need to be investigated. Table databases in comparison perform much better at investigating entities sharing a certain property—for example finding all co-workers belonging to either one of three departments and having a salary of 100'000 or above.

3 Conceptual Solution

CyberVein is a new blockchain platform that enables storage of a continuous influx of information as structured data in smart contracts. The system contains a value transaction layer to exchange the CyberVein platform token (CVT), as well as a smart contract data layer on top in order to create sidechains where structured data is stored. Transactions of both layers are stored in the same directed acyclic graph. Consensus on the value transaction layer is achieved with a resource-saving proof-of-contribution algorithm.

Figure 7 presents both layers that form the DAG that constitutes the CyberVein blockchain. In conventional blockchains like Bitcoin and Ethereum, transactions are grouped into blocks. As an analogy, in the DAG, each transaction of platform tokens (rhombuses) is mined in a single block—therefore we just refer to them as transactions instead of single-transaction blocks.

In the context of smart contracts the term *transaction* has to be clarified: from a blockchain perspective, smart contract transactions are represented as rectangles in the middle row in Figure 7 and



Adèle	100 CVT
Béatrice	100 CVT
Cécile	100 CVT
Denise	0 CVT
Estelle	100 CVT
Florence	0 CVT
Gabrielle	100 CVT

Adèle	100−100 CVT
Béatrice	100+100 CVT
Cécile	100−50 CVT
Denise	0+50 CVT
Estelle	100 CVT
Florence	0 CVT
Gabrielle	100 CVT

Adèle	100−100 CVT
Béatrice	100+100−30 CVT
Cécile	100−50+20 CVT
Denise	0+50 CVT
Estelle	100−10 CVT
Florence	0+10+30 CVT
Gabrielle	100−20 CVT

Figure 7: The middle row shows an example for CyberVein's blockchain. Rhombuses contain transactions of CyberVein's platform tokens. Their effect on the state of the ledger can be observed in the bottom row. The directed acyclic graph in the middle row also contains smart contract transactions (rectangles) with data set modifications inside them. The effect on the data set represented by the smart contract can be observed in the top row.

record the smart contract's state transitions, including the represented data set. In case of database systems, the term *transaction* refers to an isolated modification of data stored in the database. In the CyberVein system, the smart contract transactions contain these modifications to the data set.

3.1 Value Transaction Layer

This layer facilitates the exchange of CyberVein platform tokens between network participants. These are used to pay for creating smart contracts and invoking a smart contract's functions, which both require other network participants to check and mine the transactions. In proof-of-work blockchains, each block has one successor only. Out of several miners trying to advance the network's state with a new block, only one can succeed—its perception of the ledger's state is adopted by all other nodes. Fundamentally, most transactions do not interfere with each other in the sense of modifying the same entry in the ledger. Therefore they can be mined in parallel. In blockchains with linear ordered blocks, miners waste time on trying to mine a block in case a competing node is quicker, the effort was wasted.

Using a DAG structure instead, allows network participants to verify and append transactions to the ledger in parallel. There is always a delay between sending a transaction and receiving a confirmation regarding the transaction's finality in the network. Compared to linear ordered proof of work chains, effort is not wasted in DAG based chains.

A disadvantage of proof of work is its tendency towards centralisation: less economical powerful miners drop out of the network since the probability to mine a block decreases quadratically with

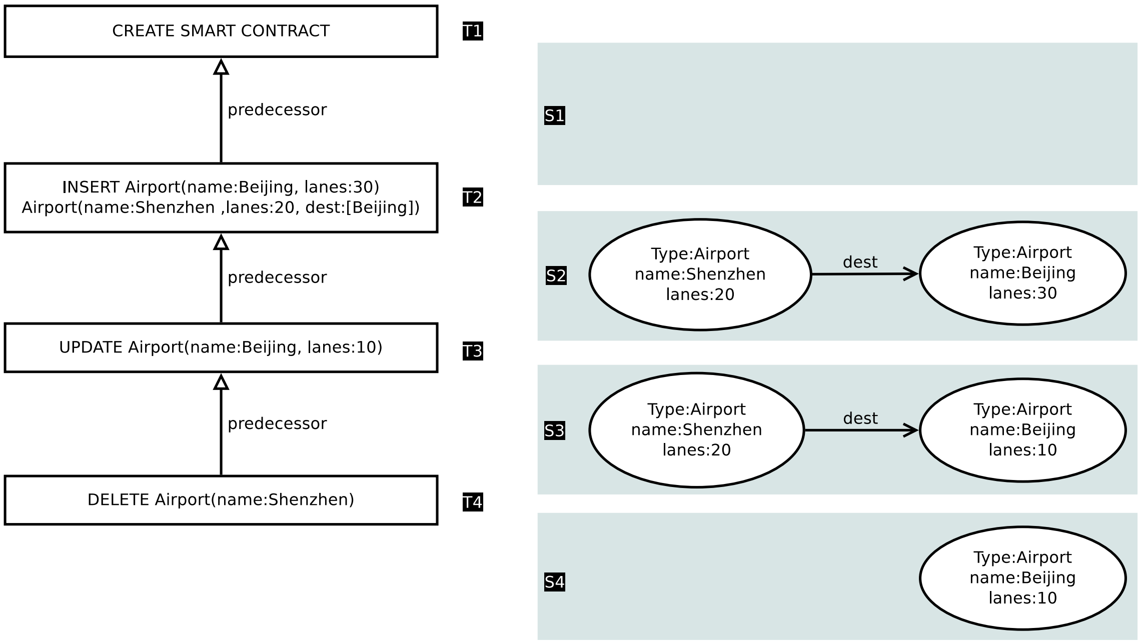


Figure 8: Nodes with permission to access a smart contract's transactions replay these locally in order to obtain a database view of the data set. Transaction *T1* only initialises the smart contract. Therefore nodes, consuming *T1* obtain state *S1* of the smart contract's data set, not containing any data yet at this stage. S4 is obtained after *T2*, *T3* and *T4* are applied as well. Although *T4* removes an entity from the data set, no data is lost. It is always possible to roll back to *S2* or *S3* and investigate that entity (in this example the airport *Shenzhen*).

increasing hash power. Apart from securing the network, the effort taken in a linear proof-of-work chain does not fulfill any purpose. CyberVein takes a new approach towards securing the network by making contributed disk storage the scarce resource miners have to harness, instead of CPU time. The probability of winning the transaction fee depends on the amount of transactions (rhombuses in Figure 7) in disk space a network participant stores.

3.2 Smart Contract and Data Storage Layer

In CyberVein, data is stored in smart contracts. In contrast to Ethereum, smart contracts are only shared among the group of network participants that use and provide the data to it. As a result, larger amounts of data can be stored using robust blockchain consensus mechanisms while not congesting the network with data that is not relevant for most network participants. To be more precise, the transactions belonging to a smart contract are only stored at those nodes that work with the data set stored in the smart contract's transactions - that makes the smart contract with its transactions a sidechain.

A fundamental difference between blockchains and database management systems (DBMSs) is mutability of stored information. To be fraud resistant and tamper proof, data needs to be stored immutably. On the other hand, data is constantly added and manipulated in databases. CyberVein combines both approaches for storing large amounts of structured data. The state of the data set is the sum of all its modifications applied in sequence. These modifications are packed into the transactions belonging to the smart contract that represents the data set.

Figure 8 shows how the state of the smart contract is advanced. In Ethereum, mining a smart contract transaction into a block coincides with executing the called smart contract function on every node of the network. CyberVein differs from that in two ways:

1. Smart contract transactions (rectangles) can be confirmed by *any* network participant, regardless of the `modify` actions stored in it, not only those who access the smart contract's

data set.

2. Smart contract transactions containing data set updates *only* affect the nodes that maintain a local database of the data set (grey boxes in Figure 8) implicitly represented by the smart contract transactions (rectangles). They update their local database by carrying out the modify instructions in the transaction, e.g., `INSERT Airport(name:Beijing, lanes:30)` or `DELETE Airport(name:Shenzhen)`).

The results are sidechains for smart contracts, that are still woven into the overall net of transactions since other transactions can point to them as their parents, regardless of the content.

Storing each modification to a data set is especially suitable for data possessing a temporal dimension like continuous measurement data, provenance records or statistical data for a metropolitan area, only being able to compare data points at different dates allows for the discovery of trends and developments. The idea of storing each data set modification is less suitable for temporary and transitional data naturally accruing in processes as a side product in order to produce a final result.

A primary goal of the CyberVein platform is to tokenise data, data streams and the access to them. Smart contracts in this context are interfaces to data sets. Smart contract functions can be annotated with a value in the token, specific to the smart contract (comparable to ERC-20). Access to the smart contract's data set is only possible through its functions, paying the requested amount of contract specific tokens, as well as the fee in CVT token for reimbursing the network for handling the transaction in the first place.

An important aspect for the data stored in a smart contract is management of permissions. A compound of research institutes might be interested in sharing data with the public, while granting write access only to members of the compound. Which network participant has access to which functionality as well as ownership of the smart contract, is managed by access control lists inside the smart contract.

The necessity for this second layer arises simply from the fact that not all data stored in a data set is relevant for most network participants.

3.3 Data Set Consistency

A key property of database systems is maintaining consistency. For example in an accounting system, the sum of all assets must always equal all liabilities. A transfer from one account to another is reflected by both afterwards. Given the first account is debited but the second one not credited, the accounting system is in an inconsistent state where no valid conclusions can be drawn any more from the stored information, e.g., account balances.

Especially for larger corporations and organisations, it is important that the used storage system supports simultaneous read/write access to records of the database while ensuring consistency of the contained information in the record. A common scenario is tracking the number of goods in stock. Assuming a buyer proceeds to checkout, one sample of the goods needs to be locked in the database to ensure the buyer will actually receive the item after billing information is already provided. On the other hand, the locked sample has to be released in case the buyer cancels the checkout procedure.

CyberVein differs from this conventional approach in several ways:

- As opposed to e-commerce systems, CyberVein's first priority is *not* capturing transitional state of concurrent business processes but being a platform to facilitate sharing and replication of data sets in smart contracts among authorised nodes (permissioned consensus).

- Blockchain systems are rather coarse-grained compared to databases featuring locking primitives like *mutexes*, *semaphores* and *critical sections*. Either a transaction is committed to the ledger (or smart contract sidechain) or not.

- CyberVein distinguishes between data producers (members of a smart contract's permissioned consensus group who append new data) and data consumers who only access and process the data without modifying it which reduces the potential for read/write conflicts between users.

CyberVein's smart contract transactions support atomic data set modification operations. Ethereum supports a similar kind of atomic transaction by including a smart contract function call in the transaction. It is equivalent to the transaction containing the instructions of the called function. Nevertheless, transactions in Ethereum are restricted to the logic already stored in the smart contract. CyberVein, in comparison, provides a lock-free language for smart contract transactions, incorporating primitives for conditionals.

3.4 Proof-of-Contribution

A key challenges in blockchain technology, meaning open and distributed ledger systems, is maintaining the network although generally being open to everyone, including malicious actors. This requires an economic model where malicious behaviour is penalised while actors supporting and maintaining the network are reimbursed.

A common approach is to introduce some kind of an economic barrier for participating in the consensus process. As a result, malicious actors cannot attack the consensus process for free. On the other hand, the reward for participating in the consensus process needs to exceed this economic barrier, otherwise no incentive and motivation for benign participation would exist. Usually, this barrier is a trade of a scarce resource on the network participant's side for platform access (being able to write to the shared ledger/records) and coins/tokens.

In Bitcoin, the scarce resource the potential network/consensus participants have to provide, is computational power [31]. The key idea is that accumulation of computational power - and therefore being able to gain control over the consensus process - is economically infeasible, due to the cost of hardware. Ideally, as a result we are a to prevent a centralisation of power via the consensus process. Nevertheless, the majority of hashing power is controlled by a small number of mining pools nowadays [1]. Participating in the consensus process without economic disadvantage leads to participation in mining pools for consumer-range users. This of course transfers power to the maintainer of the mining pool and its client software. The increase in hashing/computational power of the network makes it harder (i.e. highly unlikely) for single entities to attack the network by employing resources in order to control 51% or more of the network's hashing power. On the other hand, the computational effort is not harnessed for any useful task and therefore wasted.

CyberVein implements a consensus algorithm where the scarce resource potential consensus participants have to provide is disk space. Providing disk space to store value- and smart contract transactions directly benefits the system's purpose of facilitating distributed storage of structured data. In order to be entitled to collect a share of the transaction fees paid by users looking to have their transactions committed to the network, consensus participants have to prove that they store transactions previously confirmed by the network.

In general, what eventually secures the network is not the consumed CPU time, or some other scarce resource spent: it is the nodes in the network verifying transactions and protecting the information stored on-chain against malicious actors by rejecting transactions that try to advance the consensus into an inconsistent state.

4 Summary and Outlook

CyberVein is the first blockchain platform to introduce on-chain database functionality for sharing and replicating structured data among nodes. Transactions are stored in a directed acyclic graph to increase transaction throughput and to be able to put data sets in smart contracts on side-chains, drastically reducing the number of transactions each network participant has to store: nodes only subscribe to side-chains that store data relevant to them. Nodes can join their computational power for running algorithms on a data set as distributed system. CyberVein's MapReduce functionality allows for two use-cases:

- Run an algorithm/process on the data set which is split up into parts where each node processes one.
- Run several algorithms/processes simultaneously on a data set. Each node runs one of the algorithms/processes on the whole data set.

In both cases, CyberVein acts as the coordinator for the nodes (also for discovering and handling incomplete runs). Besides sharing, storing and processing data, CyberVein supports valuing information contained in a smart contract's data set by annotating the smart contract's functions with prices in a ERC20 like token.

***

CyberVein will adapt a philosophy of homogeneity where all parts—technology and business alike—are derived from a common vision for the project in order to gain a flexible and consistent platform, able to positively impact the constantly evolving blockchain ecosystem long-term. The definition of value of data will be investigated beyond the context of blockchain technology in order to successfully contribute to an economy where information becomes the key production factor. Middle-term goals for CyberVein's technology is adoption by labs, universities and organisations as well as creating a community in order to expand to other industries. Further steps include the formalisation of the CyberVein Virtual Machine and the definition of its language called Vein. Technical challenges include refinement of the consensus mechanism, developing a transaction mechanism which despite its non-interactivity allows for deterministic serialisation of transactions simultaneously received and map reduce functionality to effectively make use of data sets being replicated in a distributed computing platform.

References

[1] Bitcoin network hashrate. URL: https://data.bitcoinity.org/bitcoin/hashrate/6m?c=m&g=15&t=a (2018), (Accessed April 30, 2018)

[2] Aguiar, L.M., Pereira, B., Lauret, P., Díaz, F., David, M.: Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intraday solar forecasting. Renewable Energy 97, 599–610 (2016)

[3] Baker, M.: Biotech giant publishes failures to confirm high-profile science. URL: https://www.nature.com/news/biotech-giant-publishes-failures-to-confirm-high-profile-science-1.19269 (2016), (Accessed April 18, 2017)

[4] Ben Fagan-Watson: Big business using trade groups to lobby against EU climate policy. URL: https://www.theguardian.com/sustainable-business/2015/apr/15/big-business-trade-groups-lobby-against-eu-climate-change (2018), (Accessed April 29, 2017)

[5] Bochem, A., Leiding, B., Hogrefe, D.: Unchained Identities: Putting a Price on Sybil Nodes in Mobile Ad hoc Networks. In: Security and Privacy in Communication Networks (SecureComm 2018). Singapore (August 2018)

[6] Calcaterra, C., Kaal, W.A., Vlad, A.: Semada Technical Whitepaper - Blockchain Infrastructure for Measuring Domain Specific Reputation in Autonomous Decentralized and Anonymous Systems. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3125822 (2018), (Accessed April 18, 2018)

[7] Case, M.A., Burwick, H.A., Volpp, K.G., Patel, M.S.: Accuracy of smartphone applications and wearable devices for tracking physical activity data. Jama 313(6), 625–626 (2015)

[8] Chen, Y., Argentinis, J.E., Weber, G.: Ibm watson: how cognitive computing can be applied to big data challenges in life sciences research. Clinical therapeutics 38(4), 688–701 (2016)

[9] Christopher, M.: Logistics & supply chain management. Pearson UK (2016)

[10] Civic Technologies, Inc.: CIVIC - Whitepaper. URL: https://tokensale.civic.com/CivicTokenSaleWhitePaper.pdf (2017), (Accessed April 07, 2018)

[11] Collado, A.D., Chuvieco, E., Camarasa, A.: Satellite remote sensing analysis to monitor desertification processes in the crop-rangeland boundary of argentina. Journal of Arid Environments 52(1), 121–133 (2002)

[12] Dai, P., Mahi, N., Earls, J., Norta, A.: Smart-Contract Value-Transfer Protocols on a Distributed Mobile Application Platform. URL: https://qtum.org/uploads/files/a2772efe4dc8ed1100319c6480195fb1.pdf (2017), (Accessed April 18, 2018)

[13] Dean, J., Ghemawat, S.: Mapreduce: simplified data processing on large clusters. Communications of the ACM 51(1), 107–113 (2008)

[14] Delmas, M.: Research: Who's Lobbying Congress on Climate Change. URL: https://hbr.org/2016/10/research-whos-lobbying-congress-on-climate-change (2016), (Accessed April 28, 2017)

[15] Fecher, B., Friesike, S., Hebing, M., Linek, S., Sauermann, A.: A Reputation Economy: Results from an Empirical Survey on Academic Data Sharing (2015)

[16] Gandomi, A., Haider, M.: Beyond the hype: Big data concepts, methods, and analytics. International Journal of Information Management 35(2), 137–144 (2015)

[17] Giannakis, M., Louis, M.: A multi-agent based system with big data processing for enhanced supply chain agility. Journal of Enterprise Information Management 29(5), 706–727 (2016)

[18] Gunion, J.F.: The Higgs Hunter's Guide. CRC Press (2018)

[19] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al.: Deep speech: Scaling up end-to-end speech recognition. arXiv preprint arXiv:1412.5567 (2014)

[20] Hassabis, D.: Artificial intelligence: Chess match of the century. Nature 544(7651), 413 (2017)

[21] ISO/TC 268 Sustainable cities and communities: Sustainable development of communities – Indicators for city services and quality of life (ISO 37120:2014. URL: https://www.iso.org/standard/62436.html (2014), (Accessed April 22, 2017)

[22] John Walker, S.: Big data: A revolution that will transform how we live, work, and think (2014)

[23] L. M. Goodman: Tezos - A Self-Amending Crypto-Ledger (White paper). URL: https://www.tezos.com/static/papers/white_paper.pdf (2014), (Accessed April 27, 2018)

[24] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature 521(7553), 436 (2015)

[25] Leiding, B., Cap, C.H., Mundt, T., Rashidibajgan, S.: Authcoin: Validation and Authentication in Decentralized Networks. In: The 10th Mediterranean Conference on Information Systems - MCIS 2016. Cyprus, CY (September 2016)

[26] Leiding, B., Memarmoshrefi, P., Hogrefe, D.: Self-Managed and Blockchain-Based Vehicular Ad-Hoc Networks. In: Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct. pp. 137–140. ACM (2016)

[27] Lord, N.: The Third Party Data Breach Problem. URL: https://digitalguardian.com/blog/third-party-data-breach-problem (2017), (Accessed April 23, 2018)

[28] Mardis, E.R.: Dna sequencing technologies: 2006–2016. Nature protocols 12(2), 213 (2017)

[29] Marr, B.: How Blockchain Will Transform The Supply Chain And Logistics Industry. URL: https://www.forbes.com/sites/bernardmarr/2018/03/23/how-blockchain-will-transform-the-supply-chain-and-logistics-industry/#335ee6a85fec (2018), (Accessed April 23, 2018)

[30] van der Meulen, R.: Gartner Says 8.4 Billion Connected ”Things” Will Be in Use in 2017, Up 31 Percent From 2016. URL: https://www.gartner.com/newsroom/id/3598917 (2017), (Accessed April 17, 2018)

[31] Nakamoto, S.: Bitcoin: A Peer-to-Peer Electronic Cash System. URL: https://bitcoin.org/bitcoin.pdf (2008), (Accessed April 30, 2018)

[32] Nguyen, Q.K.: Blockchain - A Financial Technology for Future Sustainable Development. In: Green Technology and Sustainable Development (GTSD), International Conference on. pp. 51–54. IEEE (2016)

[33] Norall, S.: 3PL vs 4PL: What are these PLs, Anyway? Layers of Logistics Explained. URL: http://cerasis.com/2013/08/08/3pl-vs-4pl/ (2013), (Accessed April 23, 2018)

[34] Nordrum, A.: Popular Internet of Things Forecast of 50 Billion Devices by 2020 Is Outdated. URL: https://spectrum.ieee.org/tech-talk/telecom/internet/popular-internet-of-things-forecast-of-50-billion-devices-by-2020-is-outdated (2016), (Accessed April 17, 2018)

[35] Nuijten, M.B., Hartgerink, C.H., van Assen, M.A., Epskamp, S., Wicherts, J.M.: The prevalence of statistical reporting errors in psychology (1985–2013). Behavior research methods 48(4), 1205–1226 (2016)

[36] Petersen, A.M., Fortunato, S., Pan, R.K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H.E., Pammolli, F.: Reputation and impact in academic careers. Proceedings of the National Academy of Sciences 111(43), 15316–15321 (2014)

[37] Popov, S.: The Tangle - Version 1.4.2. URL: https://iota.org/IOTA_Whitepaper.pdf (2018), (Accessed April 22, 2018)

[38] Russell, S.J., Norvig, P.: Artificial intelligence: a modern approach. Malaysia; Pearson Education Limited, (2016)

[39] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al.: Mastering chess and shogi by self-play with a general reinforcement learning algorithm. arXiv preprint arXiv:1712.01815 (2017)

[40] TOP500.org: Top 500 Supercomputers - Listing. URL: https://www.top500.org/lists/2017/11/ (2017), (Accessed April 30, 2017)

[41] Weisstein, E.W.: Acyclic Digraph. URL: http://mathworld.wolfram.com/AcyclicDigraph.html (2018), (Accessed April 30, 2018)

[42] Wood, G.: Ethereum: A Secure Decrentralized Generalised Transaction Ledger. URL: http://gavwood.com/paper.pdf (2014), (Accessed April 02, 2018)

[43] Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. IEEE transactions on knowledge and data engineering 26(1), 97–107 (2014)