SoChainDB: A Database for Storing and Retrieving Blockchain-Powered Social Network Data

Hoang H. Nguyen
ehoang@l3s.de
L3S Research Center,
Leibniz Universität Hannover
Germany

Dmytro Bozhkov
bozhkov@l3s.de
L3S Research Center,
Leibniz Universität Hannover
Germany

Zahra Ahmadi
ahmadi@l3s.de
L3S Research Center,
Leibniz Universität Hannover
Germany

Nhat-Minh Nguyen
nguyenminh180798@gmail.com
Sunshine Tech Ho Chi Minh
Ho Chi Minh City, Vietnam

Thanh-Nam Doan
me@tndoan.com
Independent Researcher
Atlanta, Georgia, USA

ABSTRACT

Social networks have become an inseparable part of human activities. Most existing social networks follow a centralized system model, which despite storing valuable information of users, arise many critical concerns such as content ownership and over-commercialization. Recently, decentralized social networks, built primarily on blockchain technology, have been proposed as a substitution to eliminate these concerns. Since decentralized architectures are mature enough to be on par with the centralized ones, decentralized social networks are becoming more and more popular. Decentralized social networks can offer both common options like writing posts and comments and more advanced options such as reward systems and voting mechanisms. They provide rich ecosystems for the influencers to interact with their followers and other users via staking systems based on cryptocurrency tokens. The vast and valuable data of the decentralized social networks open several new directions for the research community to extend human behavior knowledge. However, accessing and collecting data from these social networks is not easy because it requires strong blockchain knowledge, which is not the main focus of computer science and social science researchers. Hence, our work proposes the SoChainDB framework that facilitates obtaining data from these new social networks. To show the capacity and strength of SoChainDB, we crawl and publish Hive data - one of the largest blockchain-based social networks. We conduct extensive analyses to understand the insight of Hive data and discuss some interesting applications, e.g., game, non-fungible tokens market built upon Hive. It is worth mentioning that our framework is well-adaptable to other blockchain social networks with minimal modification. SoChainDB is publicly accessible at *http://sochaindb.com* and the dataset is available under the CC BY-SA 4.0 license.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531735

CCS CONCEPTS

• **Information systems → Data management systems**; **Information retrieval**; **RESTful web services**; • **Computing methodologies → Distributed computing methodologies**.

KEYWORDS

datasets, social networks, blockchain, decentralized social networks, decentralized applications, database, network analysis

ACM Reference Format:
Hoang H. Nguyen, Dmytro Bozhkov, Zahra Ahmadi, Nhat-Minh Nguyen, and Thanh-Nam Doan. 2022. SoChainDB: A Database for Storing and Retrieving Blockchain-Powered Social Network Data. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3477495.3531735

1 INTRODUCTION

Social networks provide many useful services for their end-users; therefore, they are a part of billions of users' lives worldwide nowadays. For example, Facebook had around 2.85 billion monthly active users at the end of March 2021 [16], and Twitter has nearly 300 million active users [35]. Such a large userbase creates a rich and colossal dataset of various aspects of human activities. However, unfortunately, most social network services are deployed upon a centralized architecture. In other words, each social network is under the umbrella of a particular organization or company. Despite having many advantages, a centralized architecture still contains several fundamental disadvantages:

(1) The content ownership is not in the hands of its creators. Although users generate content through their interactions in social networks, the content is hosted by the service providers. Hence, "who is the actual owner?" is still an open question. The data leakage risk makes the problem worse.
(2) Internet censorship is another thread of centralized architectures. Service providers are under pressure from other organizations to remove or delete posts or comments displeasing those organizations. Therefore, they are ineffective tools to protect the voice of their users.
(3) Due to their large userbase, social networks are being exploited for commercialization by marketing companies and advertising agencies. Advertising contents are increasingly

integrated into many popular social media platforms without permission from users. Such activities could harm the engagement of users' experience, and have negative impact on users' behavior.

Due to these drawbacks, distributed social network architecture has been proposed as a reasonable substitution. The technology behind it is blockchain which ensures the completeness of data by leveraging cryptography. Since its proposal, social networks powered by blockchain have been evolving gradually, and they are now ready to serve millions of users. Decentralized social networks provide many benefits: First, their functionality is on par with the conventional social network platforms. For instance, standard features such as following users, posting articles, writing comments are all available in blockchain-based social networks. Second, they inherit the transparency from blockchain technology. Since all data is stored via blockchain, it is publicly accessible, and decentralization makes blockchain social networks impossible to manipulate. Third, users of blockchain social networks are rewarded for their activities. Such a system encourages users to engage more in social networks and benefits their activities. Finally, we can leverage blockchain technology of decentralized social networks to build applications such as games or trading platforms to help the existing network userbase. These potentials in decentralized social networks offer the research community many novel directions to understand human behaviors through valuable and massive data generated by variety of user interactions on the system. However, collecting data from decentralized social networks has several challenges: First, blockchain knowledge is required, which is not negligible and could create a high barrier for scientists from other fields e.g. social science who are not knowledgeable in cryptography. Second, the demand for a computational resource to synchronize complete data is usually high and expensive. For example, we required a server with more than 100GB RAM to synchronize one full node of the Hive network. Third, a cleaning process is a must since blockchain can be used for multiple purposes, not only social networks. Therefore, we propose SoChainDB, a framework for crawling data from decentralized social networks, and publish Hive's data - one of the largest decentralized social networks built on blockchain technology. The contribution of our work can be summarized as follows:

- We first propose SoChainDB, a generalized database framework and publicly available pipeline for extracting data from blockchain-based decentralized social networks.
- We publish the entire dataset of one of the largest blockchain-based social networks called Hive. It is available to download via multiple methods such as public APIs service and compressed archive files.
- We provide several unique use-cases of blockchain-based social networks that could be potential future directions for the research community to explore.

The remaining of this paper is organized as follows: Section 2 surveys the related literature, Section 3 gives an overview of blockchain-based, in general, and Hive, in specific, social networks. Section 4 describes the dataset collection pipeline, and Section 5 presents the analysis of three use cases of SoChainDB. Finally, Section 6 wraps up our study and discusses the future direction.

Ethics Declaration: Data stored in the Hive blockchain cannot be manipulated by any individuals or organizations, and there is no restriction in accessing data. Therefore, the Hive blockchain data, including comments, votes, posts, and all other blockchain-type transactions, are considered public data. Accordingly, no permission is required to collect, store, analyze and publish the data. The data stored in our SochainDB is marginally different from the stored data in Hive due to our noise filtering preprocessing step.

2 RELATED WORKS

The karate club dataset of Zachary [37] is probably the first public social network dataset. Over the years, many more public datasets have been proposed for social networks studies: Reddit dataset [4], Telegram dataset [5], Wikipedia dataset [10], Gowalla social network dataset [11], Youtube dataset [36], Yelp dataset [31], ArXiv dataset [12], eBay dataset [27], and the Kaggle network dataset [9]. McAuley and Leskovec [25] also published a dataset containing user connections in Facebook, Twitter, Google+ social networks.

Recently, Li *et al.* [22, 23] released a dataset of operation in the blockchain-based social network Steem, called *SteemOps*. We provide in-depth analysis to show the excellence of our work over these existing ones:

- *Data Completeness*: These works only provide a subset of the entire network. For example, Li *et al.* [23] released data with only operations of the Steem network. We argue that without other kinds of data such as friend networks and post content, it restricts research questions that help us understand the whole picture of blockchain-based social networks. Our system is guaranteed to provide the complete data of social networks.
- *Data Readiness and Accessibility*: The only way to use data in the current literature [22, 23] is to download the archival files that contain all information. Such a method has two problems: First, it does not allow users to investigate the network. For example, suppose researchers want to study an event in a specific period. They need to download and extract the whole archival file, then filter data according to the desired period. Second, archival files can produce computational problems. Since the archival files of the whole network are enormous, it is easy to exceed the computational resource. To overcome such obstacles, SochainDB provides two friendly options for users, especially data scientists and researchers, to query and download the data, including SQL-like interfaces and APIs.
- *Data (Sub-)Instantaneity*: Archival files are usually not in-sync with the original social networks. Notably, the latest SteemOps dataset [23] was released on Dec 1st, 2019. It is not possible to solve emerging problems like the COVID-19 outbreak and the record price of Bitcoin with obsolete data. For this reason, the data in SochainDB is guaranteed to be near-real-time with the actual data of Hive blockchain.

3 PRELIMINARY

3.1 Overview of Hive Blockchain

After a hard-fork named v0.23.0 from Steem network on March 20, 2020, which is related to some conflicts between Steem core

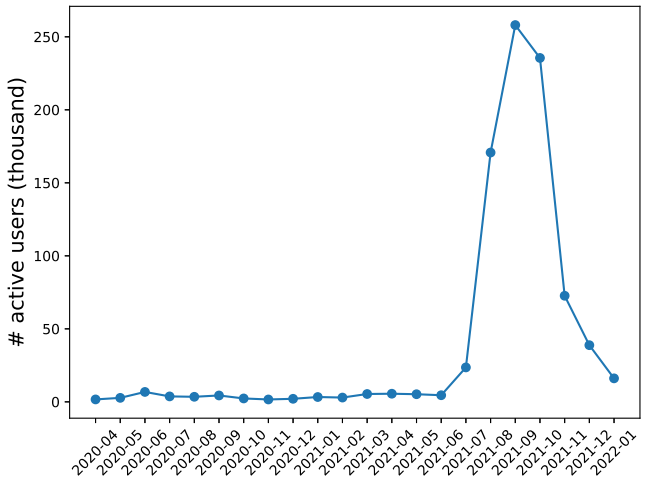


Figure 1: Growth of active users on Hive over time. From April 2020 to May 2021, the monthly growth of active users is stable, around 1,600 to 6,700 users per month. Still, since June 2021, its monthly growth has increased significantly because of the peak price of Hive tokens during this period [13].

communities and the new board of Steemit Inc. [3], Hive became an entirely community-based blockchain. In more than a year, the growth of active users per month in Hive blockchain has remarkably increased from approximately 1,600 active users in April 2020 to more than 250,000 in August 2021 (see Figure 1)[1]. This is partly due to the transformation from the Steem communities to the Hive network, besides the fact that the decentralized social network concepts have become more and more widespread over time. In the current version of the SoChainDB framework, we only focus on supporting and integrating the data extracted from the Hive blockchain to ensure the overall best performance. Moreover, Hive and Steem share a nearly similar architecture; therefore, our approach could easily and quickly adapt to the Steem network with minimal modification in the subsequent versions. Figure 2 shows the overview architecture of Hive blockchain. Here, we only discuss the core components of Hive blockchain used for generating a decentralized social network.

3.2 Preliminary of Blockchain-related Social Networks

The rapid growth of blockchain technology has attracted more and more companies and organizations working in areas related to the social network to support and integrate this technology into their platforms. Although several firms [1, 21, 26] claim that their social networks have built-in blockchain technology, most only incorporate a small part of this cutting-edge technology. For instance, one of the most straightforward approaches is often chosen to only use cryptocurrency tokens as loyalty points or a type of integrated-in currency for the users to trade in their systems through Smart Contract platforms in some of the blockchain available in the market, such as Ethereum [15], EOS.IO [7], and Binance Smart Chain [6]. The approach has the advantage of low cost and does not require advanced knowledge of blockchain technology. However, users' data is still entirely in the hands of organizations operating and developing such social networks. Therefore, these social platforms, in reality, still follow centralized models, and users do not fully control their data. At the same time, this is an essential point of a decentralized system like blockchain is aiming.

On the other hand, besides the mentioned social networks, few other platforms still focus on building and providing a completely decentralized social network, which requires a solid technical background in blockchain technology and much effort to persuade and attract the users from traditional social networks. Specifically, the most notables are the Steem network [30] and its successor, Hive blockchain [20]. The primary difference between these two blockchain-powered social networks and the rest is their completely decentralized architectures. Mainly, the core characteristics of blockchain technology, including immutability, transparency, no double spent, append-only, non-repudiation, and no single point failure, are the crucial targets in designing these social networks. These characteristics are evident through the Delegated Proof-of-Stake (DPoS) consensus mechanism using a voting system for choosing the block producers called "witnesses" per cycle and the fully decentralized designed ledger.

In addition, *Block.one*, the organization behind the popular network EOS.IO, introduced Voice being a promising decentralized social network based on EOS.IO blockchain in 2019 [8]. The announced platform Voice inherits the advantages mentioned on Steem and Hive and also integrates *Smart Contract* system, a notable feature of EOS.IO. However, at the moment, in January 2022, the new social network is still under the beta version. Therefore, we will integrate this network in our SoChainDB framework after its public release.

Consensus Mechanism: As a successor to Steem, Hive applies the Delegated Proof-of-Stake (DPoS) consensus [20], which Daniel Larimer invented in 2014 as an alternative to the Proof-Of-Work consensus algorithm widely used by several famous blockchains, e.g., Bitcoin and Ethereum. The first implementation of DPoS was on a decentralized platform exchanging cryptocurrency named Bitshares in 2015 [32], then in Steem [30], ARK [2], Lisk [24], and most recently in the platforms of TRON [34], EOS.IO [7], and Hive [20]. DPoS consensus encourages blockchain users to vote and elect delegates to validate the next block as a popular evolution of the Proof-of-Stake concept. Regular users could vote on delegates through the consensus mechanism by staking their tokens into a pool and assigning those to a specific delegate. On the other hand, the delegates are responsible for achieving consensus to generate and validate new blocks. Usually, the collected rewards of the delegates are proportionally shared with their respective voters when contributing to the blockchain.

The proponents of the DPoS consensus mechanism believe that it is a better democratic approach for a more comprehensive and diverse group of people participating in the selection process of the next block validator. Moreover, the DPoS election system is based on the earned reputation of the delegates and not the entire wealth. Therefore, if an elected node misbehaves or does not obtain the required performance, it will be quickly suspended and substituted

[1]The users are classified as "active" if they publish at least one post, comment, or vote, even if they do not have any actions later.

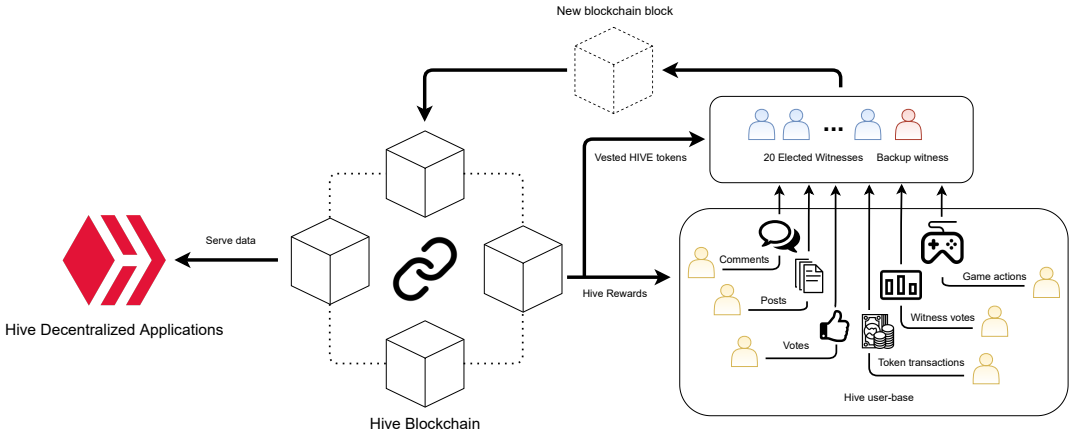


Figure 2: Overview architecture of Hive blockchain-based social network.

by another. Additionally, the limited number of validators (e.g., in Figure 2, the Hive network has twenty actual elected witnesses and one backup witness) allows the blockchain network to reach consensus more quickly. As a result, DPoS-powered blockchains are more scalable and can process more transactions per second (TPS) than Proof-of-Work and traditional Proof-of-Stake.

Hive - Social Network: Since Hive supports all basic functionalities of traditional social networks, users can create, edit, comment, and share posts. The posts can contain text, links, hashtags, mentions, and metadata such as timestamp, author, edit the information. Note that multimedia data, including images or videos, are not stored directly in the Hive blockchain due to the block size limitations and the blockchain's overall performance, especially when a new block is created and broadcasted throughout the network in seconds. Moreover, other users can view and interact with the post by reply, comment, and reblog. These ensure users' consistent experience and similarity to regular social networks.

Hive - Blockchain: The Hive underlying blockchain architecture allows easy storage and retrieval of immutable chains of large amounts of data and information. It also provides an efficient transaction platform in only three seconds without any fee. Transaction confirmation time and fees are usually among the most important challenges of promoting a blockchain's development and adaptability of use. For example, the Bitcoin network takes an average of ten minutes to validate a new block with transaction fees that tab to 60 USD at a price in April 2021 [29]. Besides, when Hive witnesses generate a new block, it includes all verified transactions or operations that users perform. These operations could be classified into four primary groups [20, 30]: (i) post and vote, (ii) witness election, (iii) followers/followings, and (iv) cryptocurrency transfer.

Hive - Tokens and Rewards System: Similar to the miners of Bitcoin and Ethereum, the Hive witnesses receive Hive cryptocurrency tokens rewarded by Hive blockchain when generating and validating new blocks. The Hive tokens can power up a Hive account for more substantial voting power and increased curation rewards, more resource credits to make transactions on Hive blockchain, and more stake in Hive governance to assign and vote witnesses and projects. Also, the Hive platform provides another

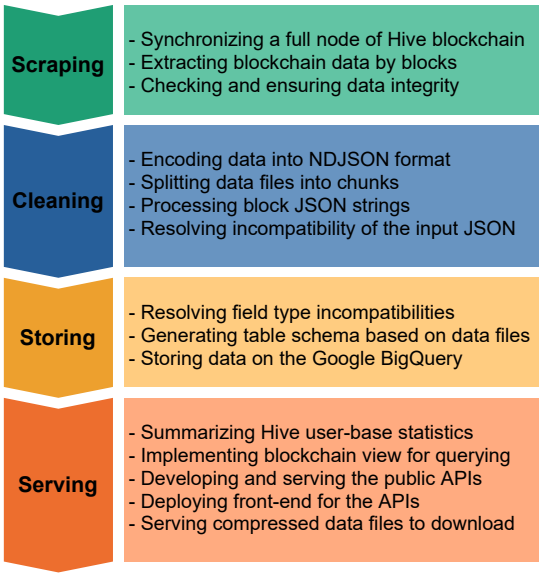


Figure 3: SoChainDB general pipeline.

unique reward system based on an upvote and downvote mechanism. It is integrated into the blockchain core using Hive tokens, and Hive Backed Dollars (HDB) tokens. Those authors who write engaging and trending content can receive Hive tokens or HDB tokens from other users on the network. Moreover, some Hive-based back-ends such as *PeakD* [28] or *Hive.blog* [19] can rank a post based on users' interaction and the number of Hive tokens staked. The higher the rank, the more likely the post would appear on these decentralized web applications' front page or trending tabs.

4 DATASET COLLECTIONS & API SERVICE

4.1 Pipeline

Figure 3 illustrates the pipeline of SoChainDB, including the following four steps:

- *Scraping:* To obtain the entire Hive blockchain dataset, we should synchronize a Hive full-node. Setting up a full node often requires advanced blockchain knowledge of custom configuration to select the suitable blockchain plugins and index the data in a sufficient time with reasonable computational resources. These settings are also relevant to the peer-to-peer protocol that synchronizes the data block. Thus, if a regular researcher or a data scientist wants to access and exploit information from these types of decentralized social networks, these are significant obstacles. Moreover, ensuring data integrity is always the highest priority of our framework when extracting from each block on the Hive distributed ledger. Hence, all scraped data are re-checked with public peer nodes for completeness, accuracy, and consistency.
- *Cleaning:* Our system encodes the data into the NDJSON-formatted files, in which every line has JSON format, and then splits them into chunks as an intermediate step for simultaneous preprocessing. This technical step is necessary because the size of the uncompressed blockchain dataset is almost one terabyte. Moreover, the Hive blockchain provides several dynamic JSON fields that allow users to push data in custom fields encoded as a regular string. Thus, the custom JSON strings should be decoded, and all field incompatibility issues should be refined and resolved before storing in a cloud service.
- *Storing:* In this step, table schemas based on the data chunks are automatically generated with the types of each block field being homogeneous before uploading them to a suitable cloud platform specialized in big data. In the current version, we use Google BigQuery as the central platform to store the decentralized social network data because of its high adaptability and compatibility with multiple varieties of data, as well as the high performance on querying big data and scalability [33]. However, our system architecture allows us to migrate to other platforms with similar functionalities to Google BigQuery flexibly with little change.
- *Serving:* We have implemented several blockchain views to optimize the query time and secure our APIs from anomaly users' behaviors while crawling data. As significant and core features, our public APIs are listed and served through our front-end at **http://sochaindb.com**. The detail of these APIs is described in Section 4.2. Likewise, we summarize blockchain operation statistics and present some prominent analyses about the Hive decentralized social network in Sections 5.2.1 and 5.2.2. We provide an HTTP service to download the compressed archive files at **http://sochaindb.l3s.uni-hannover.de**.

4.2 SoChainDB's Public APIs and Homepage

APIs service: SoChainDB provides a RESTful API service built on Falcon [17] to query our clouded data on Google BigQuery through some end-points. This highly optimized framework has significant features, such as asynchronous I/O support and simple API modeling. Falcon also showed an outstanding performance via intensive experiments on benchmarks and comparison with various other Python web API frameworks in several realistic scenarios in 2018 [18]. It assists us in accelerating the incoming requests to access our database in parallel versus sequential ways. Our APIs are generally designed to process big datasets with thousands of requests per second. The API service is deployed at http://sochaindb.com/hive-api/v1.0.0/, and its source code is publicly accessible at our Github repository [2]. Since the hard-fork of Hive from Steem happened in late March 2020 [3], our APIs in the early version could only support the Hive blockchain data from March 27, 2020 to December 6, 2021. We schedule to update the database every month and add the Steem blockchain data in the subsequent versions. All our RESTful APIs use the GET methods divided into three groups to meet the basic requirements of datasets that suit social network researchers and data scientists:

- *Blocks*: could be used to crawl the entire blocks containing all of the transactions from the Hive blockchain data. This "blocks" API allows users to collect complete data of each block in the public ledger, including a large amount of information about various operations types.
- *Posts*: could be employed to crawl posts data that we filtered from the blocks transactions. In general, the collected posts are the transactions containing operation type as *comment_options_operation* having a title field.
- *Comments*: could be used to get comments that we filtered from the blocks transactions. The comments transactions have the same post-operation type *comment_options_operation* with an empty string in title.

We also provide APIs specialized in collecting data for statistical purposes with a 10,000 default size. However, users can easily modify the size by changing the *size* parameter before requests. For example, through the GET request, we can:

(1) Crawl a list of users having the top amount of posts: http://sochaindb.com/hive-api/v1.0.0/top_posts?size=1000
(2) Crawl a list of users having the top amount of comments: http://sochaindb.com/hive-api/v1.0.0/top_comments?size=1000
(3) Crawl a list of contents for top posts and comments: http://sochaindb.com/hive-api/v1.0.0/top_words?size=1000

In Table 2, we summarize the list of API parameters used in the APIs service and their compatibility for each provided API.

SoChainDB Homepage: We built a website to show the demos and update more information. Figure 4 is a screenshot of the SoChainDB homepage. The top left corner of our homepage links to our Github repository. The navigation bar on top of the website provides several valuable resources to explore our website. For instance, the "API Docs" menu shows a detailed document of our provided APIs with various examples for ease of use. The rest of our homepage visualizes some notable statistics of our dataset, e.g., word cloud of top comments and posts, top active users of Hive networks according to the number of posts and comments.

5 USE CASES

In this section, we present some potential SoChainDB use cases through three different analyses: (i) Hive social network; (ii) Splinterlands - a card game stored in Hive; and (iii) NFTShowroom - an

[2] https://github.com/SOCHAINDB/hive-db

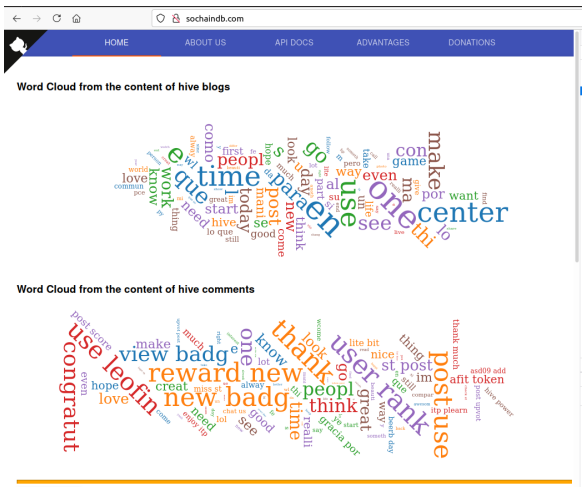


Figure 4: SoChainDB homepage.

NFT marketplace built on top of Hive. The SQL queries [3] incorporated to extract all data described in this section from SoChainDB, and some sample data of our APIs [4] can be found in our Github repository. We first introduce the overview of the Hive ecosystem and then present an in-depth analysis of the use cases.

5.1 Hive Ecosystem Overview

Hive is first used for its social network platform but with Hive Engine, a side-chain layer enabling smart contracts to work on its network, Hive also allows developers to build various decentralized applications on top of its blockchain. There are several applications built on the Hive ecosystem. Some famous ones are: (i) *Game*: Splinterlands, a multiplayer magic card game, and Rabona, a soccer management game; (ii) *Social*: LeoFinance, a crypto traders community, and Actifit, a community for users to share their workouts; (iii) *Non-Fungible Tokens Markets*: NFTMart and NFTShowroom as notable marketplaces to purchase digital assets whose owners can have proof of ownership; (iv) *DeFi*: The platforms to raise funds for cryptocurrency for any project. The most famous decentralized application is DLease; and (v) *Video*: Vimm and 3Speak are two most well-known Hive blockchain-based video-sharing services.

5.2 Analysis of Hive Social Network

5.2.1 Overall Analysis. This section analyzes various aspects of the Hive social network to depict Hive users' rich and massive data. Table 1 shows the overview statistics of the social network. The table shows that Hive social network obtains a sizable adoption. For instance, the total number of active users is more than 760,000, and around 1,200 new users join the network every day. These users generate more than three million posts in total and more than five thousand posts per day. To further illustrate the evolution of the network, we plot the growth of active users over time in Figure 1. The figure shows that the growth of active users has significantly increased by approximately 15,500% in over a year. It is an important

[3] https://github.com/SOCHAINDB/hive-db/blob/master/assets/queries.md
[4] https://github.com/SOCHAINDB/hive-db/tree/master/sample_data

	Total Count	Avg. new per days
# created users	766,080	1,246
# posts	3,574,862	5,813
# comments	8,999,321	14,633
# comment edits	259,191	421
# upvotes	162,242,767	263,809
# downvotes	1,182,265	1,922
# communities	2,527	4

Table 1: Hive social network statistics until January 31, 2022.

signal to show that the Hive social network is an active platform that receives more and more users' attention. We further plot the Hive users' activities over time in Figure 5 through the number of posts, comments and communities.

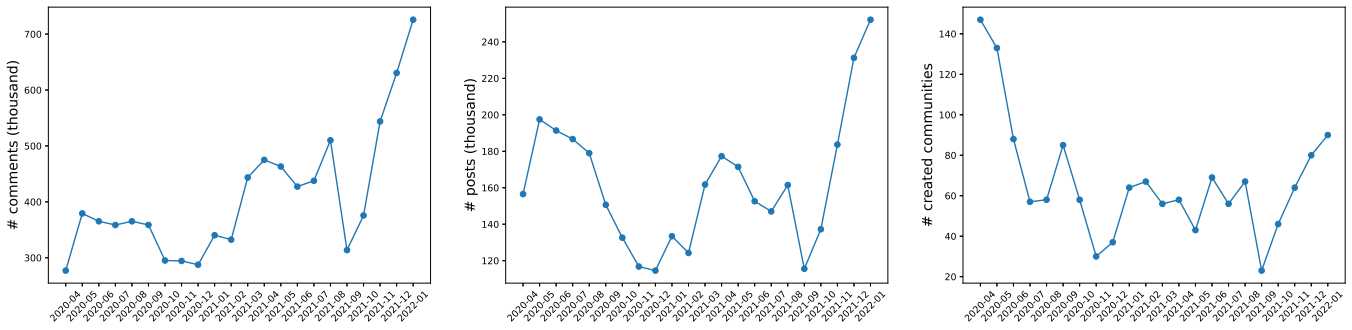
The figure depicts the regular activities of Hive users every month. Specifically, these users' posts and comments continuously increased from August 2021 and reached 240K and 700K respectively in January 2022. It shows another evidence that Hive is steadily becoming a platform for users to share their daily life habits.

To form a network, each user has an option to follow others. The essential feature of social networks and the follower/following network should follow a power-law distribution [14]. Figure 6 shows the follower/following distribution of users in the Hive network.

The figure states that Hive social network's follower/following relationship follows a power-law distribution, and it is similar to the centralized networks such as Facebook and Twitter. This pattern is also observed in the distribution of top active users by posts and comments in Figure 7.

Reward system is a unique feature of the Hive decentralized social network. Users could receive and claim their rewards in the Hive platform by the "Hive Backed Dollars (HDB)" tokens and the staked Hive tokens called "Hive Power." The former is related to user content, e.g., posts, comments, and the latter is paid directly to boost users' popularity. While liquid Hive tokens could convert to HBD, Hive Power is only based on the number of Hive tokens users staked in the platform. Figure 8 displays the value of HBD and Hive Power used by each user account over time. The figure shows an interesting phenomenon: From April 2020 to June 2021, the value of HBD per account increases over time. However, from June 2021 to October 2021, this value decreases significantly, and such a sudden drop also happens with the value of Hive power per account. This is expected since the growth of active users in this period is increased (see Figure 1) due to the Hive cryptocurrency price surge during this time [13]. The figure also shows a complex pattern of average reward in the Hive network and suggests further studies to understand how reward can affect users' behaviors in the decentralized social network. Such research can open a new direction to improve existing centralized social networks.

5.2.2 Network Analysis. One of the outstanding advantages of a blockchain-based social network is information transparency. For instance, building networks of subscribers with the communities or authors with readers or followers with followings can be exploited instantly from the public data of the blockchain. In contrast, in



(a) Number of comments (b) Number of posts (c) Number of communities

Figure 5: The Hive dynamics from April 2020 to January 2022.

Parameter	Description	Default	Accepted Values	APIs			
				blocks	posts	comments	statistics
size	Limit the results size of a request. A data sample might be large, especially the block samples. Users can set size for reducing runtime.	25	Integer	✓	✓	✓	✓
fields	Get fields in the schema. Not all fields are useful, and it depends on individuals' purposes. Users can add a list of fields for reducing runtime.	"*"	String; List of strings separated by comma	✓	✓	✓	
witnesses	Filter data by a "witness" or a list of "witnesses". It is sometimes essential information for analyzing.	None	String; List of strings separated by comma	✓	✓	✓	
ids	Filter data by the identified blocks IDs.	None	String; List of strings separated by comma	✓	✓	✓	
block_ids	Filter data by the blocks hash, which is similar to IDs, however, this is used to reference each block in the database.	None	String; List of strings separated by comma	✓	✓	✓	
operations	Filter by the operation types of the transactions in the blocks.	None	String; List of strings separated by comma	✓			
after	Filter data after a specified time. The first available time in our database is at 16:40:09 UTC on 27th March 2020 for the current version.	None	UTC format or timestamp	✓	✓	✓	
before	Filter data before a specified time. The last available time in our database is at 23:59:57 UTC on January 31st, 2022 for the current version.	None	UTC format or timestamp	✓	✓	✓	
authors	Filter by the authors. If users are interested in some posts or comments, they can add a list of authors to search for more actions.	None	String; List of strings separated by comma		✓	✓	
permlinks	Filter by "permlink" being a partition of posts or comments' URL on Hive social network. Users can add a list of "permlinks" for reducing runtime.	None	String; List of strings separated by comma		✓	✓	
post_permlinks	Filter the comments in the posts having the "permlinks."	None	String; List of strings separated by comma			✓	
words	Filter the posts or comments which contain the specified input words. This could help users catch some social network trends by searching the hot trending words.	None	String; List of strings separated by comma		✓	✓	
tags	Filter the posts which contain the specified hashtags. This might help users search the posts more accurately than the words parameter.	None	String; List of strings separated by comma		✓		

Table 2: SoChainDB API parameters. The details of parameters *fields* and *operations* can be found in the two respective links: https://github.com/SOCHAINDB/hive-db/blob/master/assets/fields.md and https://github.com/SOCHAINDB/hive-db/blob/master/assets/summary.org/#operation-types.

traditional social networks, researchers usually require the approval of the organizations or firms that own such information.

Figure 9 illustrates a directed network with 527 nodes and 933 links of various active communities with their new subscribers on the Hive blockchain social network on May 2, 2021. A more intensified red color represents the Pagerank influence score of each node. Accordingly, the most influential node in the network is *#hive-168042*, a community named Planetauto that is specialized in providing automotive content such as car guides, car reviews, events, and games for cars. Similarly, Table 3 shows the top-three influential communities every month from January to December in 2021 using the networks sharing approximate graph structures in Figure 9 but more extensive and complex. Likewise, there are three most influential communities in the Hive decentralized network in this period, including *LeoFinance(#hive-167922)*, one of the largest crypto and finance content communities, *GEMS(#hive-148441)*, a community with a wide range of topics from lifestyle, cooking, and food hobby to history and philosophy in many different languages, and *Splinterlands(#hive-13323)*, a community specialized in Splinterlands being a digital collectible card game based on Hive blockchain. Interestingly, *Aquatic Sentinels (#hive-154473)*, a new and only-26-subscribers community specialized in sharing the beauty, diversity,

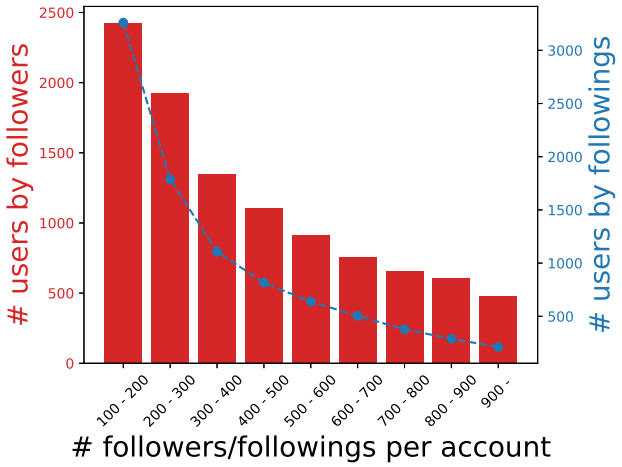


Figure 6: Number of Hive users based on the number of followers/followings per account.

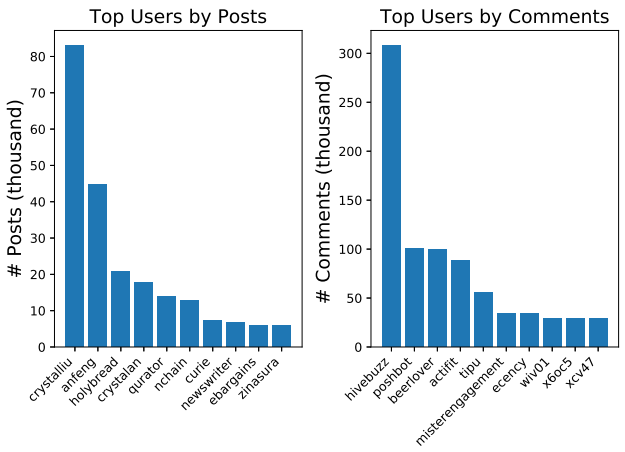


Figure 7: Top 10 users by posts and comments.

and science of the aquatic and marine ecosystems, reached third place in May 2021. The reason originates from another environment-related famous community with over 3,500 users called *Amazing Nature(#hive-127788)*, which has subscribed to the *Aquatic Sentinels* community at this time.

5.2.3 Comparison with Available Hive Statistics Analysis. There are minor differences between the statistics reported in Figures 1 and 5 and the daily and weekly reports of two users *arcange* and *penguinpablo* in *PeakD* [28] and *Hive.blog* [19]. However, we present monthly statistics in this paper, while the existing statistics reports include daily and weekly statistics. This may cause data mismatch misconception, but our investigations do not indicate significant differences.

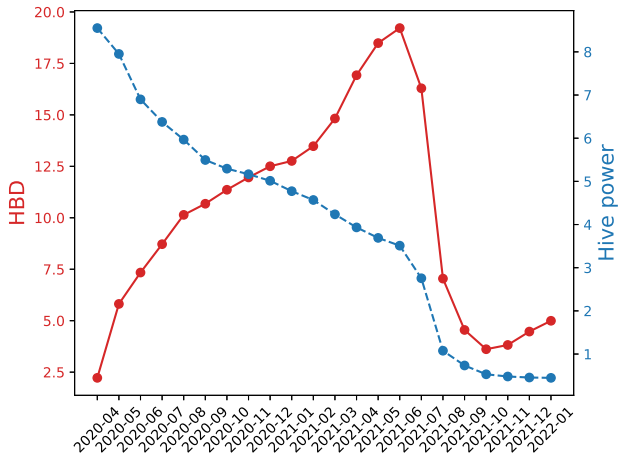


Figure 8: Average reward claimed per account in Hive social network from April 2020 to January 2022.

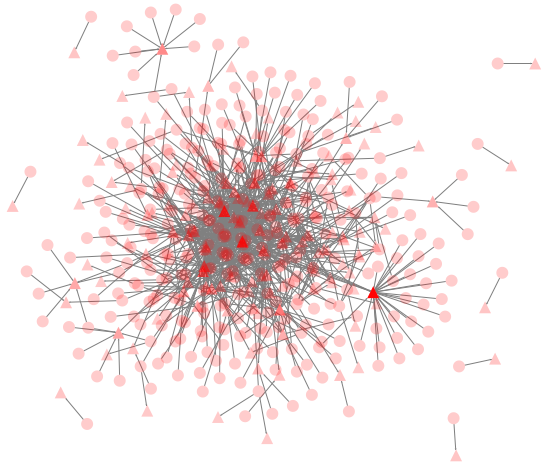


Figure 9: Network of active communities on Hive blockchain and their new subscribers on May 2, 2021. The circle nodes represent the subscribers, and the triangle nodes represent the communities. The intensity of the red color illustrates the influence of the node in the network.

5.3 Splinterlands - A Hive-based decentralized card game

Splinterlands is a collectible card game that leverages the power of Hive blockchain in storing all of its information. This is the principal technical difference of Splinterlands compared to the other blockchain-based games. For example, Gods Unchained, another collectible card game built upon Ethereum, stores only the players' assets and purchases information in blockchain while the rest, e.g., the battle result, is stored on the game publisher servers. Table 4 shows Splinterlands's overall statistics, which clearly indicates its popularity.

Networks (By Month)	Network Statistics		Top Influential Communities		
			Rank	*Hive ID*	*Name*
January	*# Nodes*	3590	**1st**	**hive-180164**	**Hive Book Club**
	# Edges	10461	2nd	hive-167922	LeoFinance
	Avg. Degree	2.9139	3rd	hive-196037	DTube
February	*# Nodes*	3690	**1st**	**hive-167922**	**LeoFinance**
	# Edges	11220	2nd	hive-196037	DTube
	Avg. Degree	3.0407	3rd	hive-148441	GEMS
March	*# Nodes*	4993	**1st**	**hive-167922**	**LeoFinance**
	# Edges	18560	2nd	hive-145666	Photo Lovers
	Avg. Degree	3.7172	3rd	hive-148441	GEMS
April	*# Nodes*	6329	**1st**	**hive-148441**	**GEMS**
	# Edges	27134	2nd	hive-174578	OCD
	Avg. Degree	4.2872	3rd	hive-167922	LeoFinance
May	*# Nodes*	6065	**1st**	**hive-148441**	**GEMS**
	# Edges	25535	2nd	hive-174578	OCD
	Avg. Degree	4.2102	3rd	hive-154473	Aquatic Sentinels
June	*# Nodes*	4775	**1st**	**hive-130560**	**Hive Diy**
	# Edges	16724	2nd	hive-131619	Blockchain Gaming
	Avg. Degree	3.5024	3rd	hive-148441	GEMS
July	*# Nodes*	4919	**1st**	**hive-110011**	**Aliento**
	# Edges	18330	2nd	hive-148441	GEMS
	Avg. Degree	3.7264	3rd	hive-174578	OCD
August	*# Nodes*	5521	**1st**	**hive-13323**	**Splinterlands**
	# Edges	20400	2nd	hive-148441	GEMS
	Avg. Degree	3.6950	3rd	hive-104151	Beyond Horizon
September	*# Nodes*	4171	**1st**	**hive-13323**	**Splinterlands**
	# Edges	12964	2nd	hive-148441	GEMS
	Avg. Degree	3.1081	3rd	hive-174578	OCD
October	*# Nodes*	4773	**1st**	**hive-181450**	**Education & Training**
	# Edges	15176	2nd	hive-13323	Splinterlands
	Avg. Degree	3.1796	3rd	hive-184127	Regional Press
November	*# Nodes*	6951	**1st**	**hive-13323**	**Splinterlands**
	# Edges	23605	2nd	hive-167922	LeoFinance
	Avg. Degree	3.3959	3rd	hive-148441	GEMS
December	*# Nodes*	8776	**1st**	**hive-181450**	**Education & Training**
	# Edges	31594	2nd	hive-184127	Regional Press
	Avg. Degree	3.6000	3rd	hive-173286	Gods On Chain

Table 3: Top influential communities based on networks of active communities and their new subscribers on Hive blockchain in 2021.

Gameplay: After signing up for the game, each new player can select numerous cards to start battles. Each card's properties fall into four categories: (i) Rarities determine how rare the card is. There are four levels of rarities: common, rare, epic, and legendary. (ii) Each card has seven stats: Mana cost, Speed, Armor, Health and Attach including Melee, Ranged, and Magic. (iii) Fire, Earth, Water, Life, Death, Dragon, and Neutral define the faction of this card. (iv) Each card has more than 46 abilities to increase the fun and randomness of battles. Before a battle between two players, each player is provided with a fixed amount of mana, and each player chooses the same number of cards to organize on the battlefield. The result of the battle is determined by position, strength, cards' ability, and some randomness injected by the system. There are three types of battle: ranked, practiced, and friendly matches. The last two do not affect the ranking of players. The players' cards can be traded with other ones.

Data Analysis: Each blockchain transaction records an action that happened in the game. We can cluster these actions into the following four categories. Table 4 represents some notable activities for each category:

Category	Feature	Count
Overview Statistic	# of active users	380,476
	# of daily games	9,240,084
	# of cards	283+
Account-related Actions	# Claim Reward	58,248,195
	# Upgrade Account	664,180
	# Add Wallet	517,058
Battle Actions	# Match Finding	186,016,801
	# Match Starting	2,477,844
	# Surrender	1,685,134
Asset Actions	# Burn Cards	323,701
	# Lock Assets	137,777
Purchase Actions	# Sell Cards	6,387,080
	# Purchase Record	52,786

Table 4: Splinterlands statistics until January 31, 2022.

(1) *Account-related actions*: Some example operators in this category are rewarded with more than 36 million transactions, while upgrade account operator has more than 626K transactions on the Hive blockchain, and adding wallet is nearly 480K operators.
(2) *Battle actions*: It contains activities related to a battle, e.g., players of a battle, each player's order of card deck in the match, the battle result, and the battle type. Some examples include more than 178 million match-finding transactions, where users have played nearly 2.5 million matches.
(3) *Asset actions*: It is the card information that each player has. So, for example, players can destroy cards, which they do not want to use. According to Hive transactions, there are 314 thousand actions recorded on the system. Lock asset is also done more than 92K times.
(4) *Purchase actions*: Since the game allows its players to buy, sell, or transfer their assets, all transactions are stored in the blockchain to avoid manipulation, even from the game publisher side. As stated in Table 4, around 4.5 million transactions are related to the users' card selling activities. It expresses the users' high trading activities in this game.

5.4 NFTShowroom

NFTShowroom is a marketplace for artists selling their digital art. The platform associates each digital art with a unique Non-Fungible Token (NFT). Since Hive blockchain is not designed as a decentralized computational system as Ethereum or EOS.IO, a smart contracts side-chain layer called Hive Engine is used to issue token SWAP.HIVE of NFTShowroom. Then, the ownership of the artwork can be verified employing the Hive Engine smart contracts, and the primary layer of the Hive blockchain is leveraged for the verification in the case of NFTShowroom. Moreover, the purchase history of the digital art is trackable via the transactions created on the Hive Engine nodes before automatically sending them to the Hive main chain. When digital art is transferred, sold, or bought, the

transactions are recorded in the Hive blockchain. This information is publicly transparent and could be verified by other users.

Till end of January 2022, NFTShowroom consists of more than 11,866 artworks sold. On average, ten artworks are purchased through the system every day. Based on our dataset, the total number of NFTShowroom tokens minted is 45,961, and such number is increasing over time because of its rapid development.

6 CONCLUSION AND DISCUSSION

In this paper, we presented SoChainDB, a framework to crawl blockchain-based social networks. Its robust and general architecture can handle various kinds of blockchain systems. Along with our system, we provide the public dataset of Hive - one of the largest blockchain social networks. We also discussed and released the data of Splinterlands, a collectible decentralized card game, and NFTShowroom, a platform for purchasing the ownership of digital arts, both built upon Hive blockchain technology. All data presented in this paper is ready and accessible via our website through a RESTful API service or archival data files. As future work, several research directions on information retrieval might fit this blockchain-powered social network database:

- *Massive Scale Social Network Analytics:* With more than 100 GB of post-processed data, SoChainDB allows for large-scale combined analysis of social networks' various aspects or topics. Thus, along with common social networking factors, extensive research can exploit the unique blockchain characteristics such as users' motivation to contribute highly-rated content through the built-in crypto-tokens and rewards mechanisms to explore the impact of articles. Hence, the assessments could be more comprehensive than similar approaches on the regular social network data.
- *Cross-domain Behavior Analytics:* Since we can obtain the entire Hive data, we could use the data to answer different basic questions related to the behavior of users across services. For example, game players could use social network platforms to publicize their achievements and build friendships with other players. Such activities allow us to understand users' behaviors in various domains and build a more accurate recommender system that offers helpful information.
- *Impact of Reward on User Engagement:* The reward system is a unique feature of blockchain-based social networks. Understanding the causality between earning and user activities in social networks can open a direction to redesign existing social networks toward offering a better user experience.

ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

REFERENCES

[1] Allabout.me Tokens Ltd. 2017. all.me Whitepaper. https://allmestatic.com/mepaytoken/all-me_whitepaper.pdf
[2] ARK Ecosystem, SCIC. 2019. ARK Ecosystem Whitepaper.
[3] Paddy Baker. 2020. Steem Hard Fork Confiscates $6.3M, Community Immediately Takes It Back. (2020). https://www.coindesk.com/steem-hard-fork-hive
[4] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 830–839.
[5] Jason Baumgartner, Savvas Zannettou, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Telegram Dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 840–847.
[6] Binance. 2020. Binance Smart Chain: A Parallel Binance Chain to Enable Smart Contracts. https://github.com/binance-chain/whitepaper/blob/master/WHITEPAPER.md
[7] Block.one. 2018. EOS.IO Technical White Paper v2. https://github.com/EOSIO/Documentation/blob/master/TechnicalWhitePaper.md
[8] Block.one. 2019. Voice: The Road to Beta. https://b1.com/news/voice-the-road-to-beta/
[9] Casper Solheim Bojer and Jens Peder Meldgaard. 2021. Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting* 37, 2 (2021), 587–603.
[10] Andrea Capocci, Vito DP Servedio, Francesca Colaiori, Luciana S Buriol, Debora Donato, Stefano Leonardi, and Guido Caldarelli. 2006. Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Physical review E* 74, 3 (2006), 036116.
[11] Eunjoon Cho, Seth A Myers, and Jure Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1082–1090.
[12] Colin B Clement, Matthew Bierbaum, Kevin P O'Keeffe, and Alexander A Alemi. 2019. On the Use of ArXiv as a Dataset. *arXiv preprint arXiv:1905.00075* (2019).
[13] Coin Market Cap. 2021. Hive Price. https://coinmarketcap.com/currencies/hive-blockchain/
[14] Gábor Csányi and Balázs Szendrői. 2004. Structure of a large social network. *Physical Review E* 69, 3 (2004), 036131.
[15] Ethereum Foundation. 2014. Ethereum Whitepaper. https://github.com/ethereum/wiki/wiki/White-Paper
[16] Facebook. 2021. Facebook press release. https://investor.fb.com/investor-news/
[17] Falcon Framework. 2021. Falcon Framework. https://falconframework.org/
[18] Fotis Gimian. 2018. Choosing a Fast Python API Framework. https://fgimian.github.io/blog/2018/05/17/choosing-a-fast-python-api-framework/
[19] HiveBlog. 2021. HiveBlog. https://hive.blog
[20] Hive.io. 2020. Hive: Fast. Scalable. Powerful. The Blockchain for Web 3.0. https://hive.io/whitepaper.pdf
[21] Indorse Pte. Ltd. 2020. Indorse 2.0. https://indorse-staging-bucket.s3.amazonaws.com/Indorse+2.0+Light+Paper.pdf
[22] Chao Li and Balaji Palanisamy. 2019. Incentivized blockchain-based social media platforms: A case study of steemit. In *Proceedings of the 10th ACM Conference on Web Science*. 145–154.
[23] Chao Li, Balaji Palanisamy, Runhua Xu, Jinlai Xu, and Jingzhe Wang. 2021. SteemOps: Extracting and Analyzing Key Operations in Steemit Blockchain-based Social Media Platform. *arXiv preprint arXiv:2102.00177* (2021).
[24] Lisk Foundation. 2021. Lisk Consensus Algorithm. https://lisk.com/documentation/lisk-sdk/protocol/consensus-algorithm.html
[25] Julian J McAuley and Jure Leskovec. 2012. Learning to discover social circles in ego networks.. In *NIPS*, Vol. 2012. Citeseer, 548–56.
[26] Minds Inc. 2021. Minds Whitepaper v2. https://cdn-assets.minds.com/front/dist/browser/en/assets/documents/Minds-Whitepaper-v2.pdf
[27] Mehryar Mohri and Andres Munoz Medina. 2014. Learning theory and algorithms for revenue optimization in second price auctions with reserve. In *International Conference on Machine Learning*. PMLR, 262–270.
[28] PeakD. 2021. PeakD. https://peakd.com
[29] Jamie Redman. 2021. Bitcoin Fees Tap $60 per Transaction, Users Say Fees Restrict Adoption, Others 'Embrace' the BTC Fee Pump. (2021). https://news.bitcoin.com/bitcoin-fees-tap-60-per-transaction-users-say-fees-restrict-adoption-others-embrace-the-btc-fee-pump/
[30] Steemit Inc. 2018. Steem: An incentivized, blockchain-based, public content platform. https://steem.com/steem-whitepaper.pdf
[31] Abu Saleh Md Tayeen, Abderrahmen Mtibaa, and Satyajayant Misra. 2019. Location, location, location! quantifying the true impact of location on business reviews using a Yelp dataset. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. 1081–1088.
[32] The BitShares Organization. 2018. The Bitshares Blockchain. https://whitepaper.io/document/388/bitshares-whitepaper
[33] Jordan Tigani and Siddartha Naidu. 2014. *Google BigQuery Analytics*. John Wiley & Sons.
[34] TRON Foundation. 2018. TRON: Advanced Decentralized Blockchain Platform. https://tron.network/static/doc/white_paper_v_2_0.pdf
[35] Twitter. 2021. Twitter About. https://about.twitter.com
[36] Jaewon Yang and Jure Leskovec. 2015. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* 42, 1 (2015), 181–213.
[37] Wayne W Zachary. 1977. An information flow model for conflict and fission in small groups. *Journal of anthropological research* (1977), 452–473.