

Web3 social platforms: modeling, mining and evolution

Cheick T. Ba¹, Alessia Galdeman¹, Manuel Dileo¹, Christian Quadri¹,
Matteo Zignani^{1,*} and Sabrina Gaito¹

¹Department of Computer Science “Giovanni degli Antoni”, University of Milan, via Celoria 18, Milan, 20133, Italy

Abstract

Web3, one of the arising paradigms which may rule the future Web, is also representing a source of big data stored in the underlying blockchains. Many different research fields are benefiting from these large collections of temporal and heterogeneous data, which capture different aspects of the interactions among people and between people and Web3 platforms. Specifically, since each piece of information is validated and timestamped, Web3 platforms are becoming an invaluable source for understanding the dynamics of these techno-social systems at a high temporal resolution. In this contribution, we focused on the analysis of the evolution of the networked structure of Web3 social networks through the lens of discrete choice models, and on the changes in the structure of the relationships after a shocking event has occurred on the platform - namely a hard-fork in the supporting blockchain. To support large-scale analysis, we represent Web3 platform data as temporal multigraphs manageable by modern graph database management systems. The main findings, which represent a summary of our effort in mining data from Web3 platforms, highlight some interesting aspects: i) when applied to Web3 social networks, discrete choice models allow us to decompose the evolution of social networks into different growing mechanisms, which are quite stable during the observation period; and ii) in a stratified context, such as Web3 platforms, interactions resulting from economic actions, such as transfers or loans of crypto-tokens, are as important as social relationships to predict how users will behave during a shocking event. These are a few examples of how Web3 social platforms may represent a challenging playground for a more in-depth understanding of the users’ behaviors when social and economic interactions are strictly intertwined.

Keywords

Web3 social networks, social network evolution, link creation dynamics, customer migration

1. Introduction

In the last years, the actual structure of Web 2.0 has been questioned by novel paradigms which are trying to reduce the over-centralization around a few big platforms and tech companies. One of the ideas gaining momentum is Web3, i.e. the design of platforms and software systems built on blockchain technologies to promote a decentralized Web. In fact, we are witnessing the birth of decentralized counterparts of Twitter or Reddit, embodied by Hive, Mind, or Steemit;

ITADATA’22: The 1st Italian Conference on Big Data and Data Science, September 20–21, 2022, Milan, Italy


*Corresponding author.

✉ cheick.ba@unim.it (C. T. Ba); alessia.galdeman@unimi.it (A. Galdeman); manuel.dileo@unimi.it (M. Dileo); christian.quadri@unimi.it (C. Quadri); matteo.zignani@unimi.it (M. Zignani); sabrina.gaito@unimi.it (S. Gaito)

🆔 0000-0002-4035-7464 (C. T. Ba); 0000-0003-3286-4666 (A. Galdeman); 0000-0002-3608-8142 (C. Quadri); 0000-0002-4808-4106 (M. Zignani); 0000-0003-3779-2809 (S. Gaito)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

but also services that are specific to the Web3 world, such as Decentralized Finance (DeFi), Decentralized Autonomous Organizations (DAOs), and non-fungible token (NFT), a financial asset, linked to data stored in the blockchain, that can be traded. Although the idea of Web3 is at the heart of a heated debate between enthusiasts and skeptics, the platforms following this paradigm offer a great opportunity to researchers in different fields thanks to the huge volume of high-resolution data stored in the supporting blockchains. Indeed, a broad set of data about these techno-social systems can be easily accessible: by the nature of blockchains, data are publicly available, validated, and affordable by interfacing with the API blockchain. Moreover, data from Web3 platforms offer two advantages: *i*) each piece of information is timestamped since each blockchain block has a validation timestamp; and *ii*) each block reported multi-faceted interactions - social, economic, financial - among people and between people and platform. So, these data sources have all the features to face tasks and issues related to modern techno-social networks and to support detailed and in-depth analysis of users' traits. Specifically, here we focus on a few issues - which summarize our effort in the recent years to mine Web3 platforms - related to the growth of Web3 social networks from the perspective of link creation mechanisms and the effects of shocking events.

Understanding and mining how the networks behind large techno-social systems grow and react to exceptional events are fundamental elements for the comprehension of the main processes driving the evolution of such systems and for the identification of specific patterns of growth which are consequences of the platform design. In this regard, in the past years, many models, mechanisms, and measures describing the network growth from a link formation perspective have been proposed, however, most of these approaches rely on the assumption that the growth is guided by a single parameterized mechanism. But, current techno-social networks are the result of different and heterogeneous behaviors where different choices and mechanisms occur [1], an observation which is further emphasized by Web3 platforms, where social and economic motivations mix together and determine the choice of users in both common and unusual situations as user migration resulting from conflicts within the social network. For these reasons, in this contribution, we adopt a modeling approach based on temporal multidigraphs which allows us to represent the heterogeneity of the interactions expressed in Web3 social platforms and a few machine learning methods able to decouple the mechanisms - social and economical - driving the choice of creating new relationships or migrating to a new blockchain in case of splitting event of the network determined by a hard fork of the blockchain. As a case study, we focus on the Web3 social platforms supported by the Steem blockchain, whose main social platform is Steemit, and by Hive, the blockchain born from a hard fork - splitting event - of the Steem blockchain. Besides a few advances in the methodological aspects concerning the application of discrete choice models to stratified social networks¹ and the prediction of whether or not users migrate towards other platforms, the findings resulting from the analysis have highlighted a few main aspects of the Web3 social platforms:

- by applying discrete choice models to the lifetime of Steemit, we observe that the different mechanisms driving the choice of establishing new connections are quite stable along the observation period, where the attitude towards reciprocating links and making

¹Stratified network is one of the terms indicating multi(di)graphs, where vertices are connected by one or more links, each of a specific type.

- connections with users with many common relationships are the leading factors;
- it is possible to predict, with reasonably good performances, whether or not a user will migrate to a new Web3 platform born by a fork. Specifically, these outcomes have been reached even with the only information on the network structure, without including textual or external data such as the trend of cryptocurrencies. Moreover, in a stratified context with different types of connection, interactions resulting from economic actions, such as transfers or loans of crypto-tokens, are as important as social relationships to predict how users will behave during a shocking event, such as a hard fork.

The paper is organized as follows. Section 2 provides a brief introduction of Web3 social platforms and an overview of the related studies. In Section 3 we describe how we model and manage data from Web3 social platforms. Sections 4 and 5 report the main findings of the growth of the Web3 platforms by applying discrete choice models and a discussion about predictability of user migration after a shocking event, such as a hard fork of the underlying blockchain.

2. Backgrounds and related works

Different services and platforms lie under the umbrella of the Web3 paradigm, but they all have the usage of blockchain technology as a common factor. Hereby we introduce the reader to a specific type of Web3 platform: Web3 social platforms or networks, broadly speaking online social networks or blogging platforms that rely on a blockchain to validate and persist all the interactions and actions established by their users.

Web3 social platforms. By Web3 social platform, we denote a web application which *i*) offers a set of “social actions” - following, commenting and voting - facilitating online interactions among accounts; and *ii*) whose core functions are ground in an underlying blockchain that guarantees the persistence and the validity of the operations. One of the most interesting consequences of this architecture is a strong link between economical aspects and online social behaviors, in fact, most of the current Web3 social platforms implement: *i*) a token ecosystem based on blockchain technology for promoting high-quality content and users, and validating social and economic operations; and *ii*) a rewarding system for distributing the wealth of the platform. In particular, the rewarding system defines the set of rules and mechanisms regulating the distribution of tokens among the users who actively participate in the platform activities. In most of these platforms, rewards are assigned to accounts that publish content - posts or comments - and accounts that promote content through upvoting, downvoting or sharing. Specifically, content promotion is based on a stake-based voting system, where the voter decides how much of its economic power - the amount of gained crypto-tokens - to put behind a vote.

In the landscape of Web3 social platforms, most of the research studies have been focused on Steemit. Launched in 2016, Steemit has been one of the most widespread Web3 social platforms and it is considered a pioneer for the Web3 ecosystem since it has introduced the seminal concepts of rewarding system and proof-of-stake consensus algorithm for block validation. In detail, the platform is hosted on a blockchain called Steem, and implements three different tokens: STEEM, Steem Dollar - (SBD), and Steem Power - SP, where the last is on the basis of the internal rewarding system and the first two tokens are tradable on exchange markets. Steemit

has gathered the interest of researchers for its characteristics and has been dissected in many aspects. For example, a few studies have focused on the features of different types of social networks resulting from diverse interactions or specific subsets of accounts ([2, 3]). In particular, Chonan [4] and Kim *et al.* [5] have focused on the structure of Steemit “follow” network and its characteristics. Also, Guidi *et al.* [6] have delved into a study of the follower–following graph, and have studied other operations in Steemit [7]. Aside from social relationships, they have also focused on block producers (witnesses) and highlighted their social impact on the platform [8]. As for economic aspects, Ciriello *et al.* [9] and Thelwall *et al.* [10] have analyzed the relationship between rewards and content, while Li *et al.* [11] have analyzed the rewarding system in Steemit from a network perspective. Additional information provided by the underlying blockchain has been exploited in other studies; for instance, users’ content has been also used for facing text mining [12] and bot detection [13] tasks. Finally, a few recent works have also taken into account temporal information to investigate network dynamics. For instance, Jia *et al.* [14] focus on the diffusion of content, while in our previous works we discussed the interplay between cryptocurrency price and the link creation process [15], the impact of user migration on the social networks [16] and the role of groups in this phenomenon [17], and the bursty dynamics of the link creation process [18].

Hard fork and user migration. Web3 social platforms may offer data about a phenomenon which is peculiar to blockchain-based systems. Indeed, in these systems situations where miners/validators change the consensus protocol may happen, leading to what is known as blockchain fork. Specifically, two types of fork may occur: *i)* soft forks, where changes retro compatible with the previous consensus protocol are introduced so that new blocks are added to the same chain; and *ii)* hard forks, where miners do not consider as valid the blocks validated with the new protocol so that two different branches are created if validators do not reach a consensus on which protocol to use. In the latter case, the members of the original branch may opt to migrate to the platform based on the new branch, leading to a phenomenon denoted as user migration. User migration is a “universal” process spanning both centralized and decentralized online social media but is not fully understood yet, especially in the Web3 world. Most of the studies are based on centralized social platforms. For instance, Kumar *et al.* [19] have analyzed user migration patterns, by matching user accounts through external data. Newell *et al.* have conducted an analysis of user activity during a cross-platform migration through surveys to understand the motivations behind migration. Other works have focused on users migrating across communities on the same platform, showing non-random migration patterns in Facebook groups [20]. A more in-depth analysis has been conducted [21] on Reddit, where user migration across COVID-19 subreddits has been analyzed through different time scales. All previous studies are based on data collected from centralized social platforms, however only one work [16] has focused on user migration in Web3 social platforms, as a consequence of a hard fork. Also in this case, Steemit represents a reference point since it experienced a hard fork as a reaction to a hostile takeover of the Steem blockchain, leading to the new branch Hive.

3. Modeling and managing Web3 social platform data

Web3 social platforms make available to their users a rich set of operations to support different kinds of interaction, namely *interaction actions*. From a network perspective, interaction actions - comments, likes, reacting and following - result in different types of relationships connecting accounts/users. Interaction actions in Web3 social platforms have two further important features inherited from blockchain: *i)* each interaction action has a timestamp corresponding to the time the block containing the action has been validated and recorded into the chain, and *ii)* interaction actions are not merely social, rather also related to economic or financial operations, such as borrowing or transferring tokens or assets between accounts. The latter aspect makes Web3 social platforms complex techno-social systems with different intertwined layers of actions.

Formally, interaction actions are represented as a set of tuples $I = \{(u, v, t, r)\}$ where u and v are accounts, who interact through an action of type r , validated by the blockchain at time t . Given the temporal information associated to each tuple in I , we build two different kinds of sequence of directed multigraphs [22], which describe the evolution of the platform: *i)* a sequence of incremental directed multidigraphs, and *ii)* a sequence of differential snapshots, where each snapshot captures the network generated by the interaction actions occurring only in a given time window. To cope with issues treated in this work, here we detail the incremental approach. Specifically, we consider an evolving edge-labeled multidigraph \mathcal{G} represented by a sequence $\langle G_1, \dots, G_T \rangle$ where each $G_t = (V_t, E_t, R, w_t)$ is a weighted edge-labeled multidigraph, and T is the maximum timestamp in I . Through multidigraphs we include in the model different types of relations expressed by r . Each multidigraph of the sequence is defined by the following elements:

- V_t : the set of users u which belong to at least one interaction action $(u, v, t_i, r) \in I$ such that $t_i \leq t$;
- E_t : the set of triple (u, v, r) with $u, v \in V_t$ and $r \in R$, denoting a specific action type taking value on the set R of actions made available by the blockchain;
- $w_t : E_t \rightarrow \mathbb{R}$: a weighting function which returns the number of interaction actions of type r involving u and v occurring before or at the timestamp t .

Finally, it is worth noting that the above definition does not take into account the meaning of operations, especially complementary operations, such as “follow” and “unfollow” where the latter determines a removal of the link created by the former operation. This way actions can only increase the state of a multidigraph, while semantic constraints can be introduced in a successive phase of the analysis.

3.1. FlowChains: a platform for gathering and managing Web3 social data

Handling data produced by Web3 social platforms requires an effort not only in modeling and representing temporal and heterogeneous data but also in designing and developing scalable analytics platforms which have to gather data from different blockchains and manage large-scale multidigraphs. To this aim and to support our analysis of the temporal aspects of Web3 social platforms, we have developed FlowChains, an analytics platform for Web3 social and trading data, whose architecture is depicted in Figure 1. The core of the architecture is the

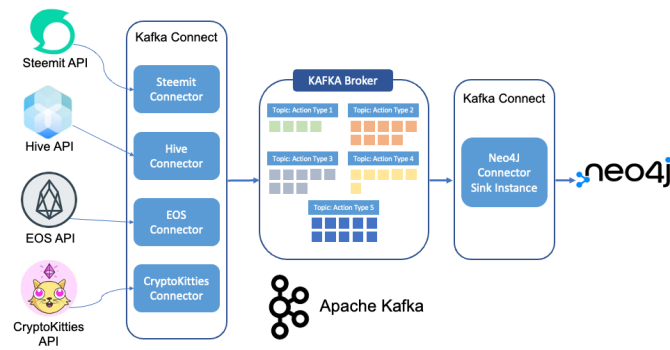


Figure 1: The FlowChains platform. Through the Kafka Connectors in the Kafka Connect module, the Kafka Broker handles different data sources linked to diverse blockchains, and, at the same time, sends the interaction actions to the temporal multidigraph implemented by the graph database Neo4J.

event streaming platform Apache Kafka which collects data through a set of Kafka Connectors. Each Kafka Connector handles a specific stream of blocks gathered through the APIs released by Web3 platforms and applications. Then, syntactically different operations coming from different blockchains which express the same interaction action are grouped in the same Kafka topic. This aggregation step facilitates the alignment among the data produced by the rich set of blockchains. Finally, the interaction actions belonging to a specific topic are transformed into a temporal annotated link in a graph database management system. In FlowChains the graph database is Neo4J, the leading solution for handling and managing large-scale multidigraph. The choice of Neo4J has been dictated by the fact that it natively supports multiedges between two vertices, complex attributes can be assigned both to vertices and links, and it follows the schema-free paradigm so allowing a certain level of flexibility in the data model.

3.2. Web3 social platform datasets: Steemit and Hive

Through the FlowChains platform, we were able to collect a large dataset covering the lifespan of two important blockchains Steem and Hive, at the basis of the two Web3 social platforms: Steemit and Hive blog. Specifically, the blockchain Hive has originated by a hard fork of Steem on the 20th of March 2020, after a 51% attack. Although they are two different blockchains both Steem and Hive have released more than 50 operations in common. Among them, we are interested only in operations generating interaction actions, of which a subset has been reported in Figure 1b. These actions have been further grouped into two main categories: *i)* financial and *ii)* social operations. Financial operations are for rewards and token management, and asset and share transfer; whereas social operations correspond to posting, rating, voting, sharing and following. From the modeling perspective, this aggregation corresponds to defining R as $\{social, financial\}$.

The details about blocks and operations for both platforms have been gathered through official public APIs. Moreover, due to the implementation of the hard fork, data between the two blockchains are identical up to the fork event. From there, Hive and Steem have recorded different data, as they have become two separated entities. So, we collected operations from

the very first block on Steem blockchain, produced on 24th March 2016, up to January 2021. For Hive, we start from the first block after the fork (20/03/2020), and up to January 2021. Overall, from Steem, we extracted 993,641,075 operations related to social interaction actions and 72,370,926 operations related to financial actions; from Hive we have a total of 206,224,132 social actions and 4,041,060 financial actions. It is worth noting that both blockchains record the operations with a three-second granularity.

4. Mining the growth of Web3 platforms

The understanding of how online social networks grow and evolve is a central issue in temporal network analysis, however, traditional approaches have failed to catch the complexity of such phenomenon, especially when online social networks are the result of a mix of growth mechanisms based on different aspects, from social to economic or financial. Recently, a few approaches based on discrete choice models [1] have shown promising results as the possibility of decoupling the different mechanisms which drive the formation of social networks. The fundamental idea behind these approaches is that we can think of the formation of a directed link (u, v) as a choice made by u to connect with v , where the set of possible choices for u is the set of all other nodes. For further details about discrete choice models applied to social network formation we refer the reader to Overgoor *et al.*'s seminal work [1]. Essentially, the underlying inference problem is an estimate of the parameter θ^T which characterizes a random utility function a node u uses to choose the target node v , i.e. $U_{u,v} = \theta^T x_v + \epsilon_{u,v}$, where x_v is a vector of features for the node v and ϵ models the noise in choosing among the alternatives. In our analysis we apply a conditional logit model with a linear utility function so that the likelihood function is convex with respect to the parameters θ : θ can be inferred by efficiently maximizing the likelihood by using gradient-based optimization methods. In the context of conditional logit choice models, the parameters θ represent the importance of a specific feature in x_v for the choices made by u ; while the features summarize the different mechanisms which may act during the formation of the network. Here we leverage the above model and its interpretation by applying it to different snapshots of the Steemit lifespan so to highlight if the choice model is stable. Moreover, we take into account features based on social and financial actions so to find the main aspects driving the evolution of Web3 social platforms.

Experimental setting. To investigate the stability of the choice model and the utility function parameters, first we identify a set of dates - 2017-03-15, 2017-04-15, 2017-05-15, 2017-06-15, 2017-07-15, 2017-08-15, 2017-09-15 - which correspond to specific stages of the growth of Steemit, as shown in [15]. Then, we proceed by transforming the temporal links of type “follow” into choice data: given an interval $[t_s, t_e]$, we extract G_{t_s} from the sequence of multidigraphs \mathcal{G}^S . Finally, from the set of “follow” links created in $[t_s, t_e]$, we extract n elements which corresponds to the choices (u, v) . On the other hand, alternative choices have to be selected. Indeed we have to estimate the utility assigned by node u to other alternatives, to approximate the likelihood function. To speed up the computation, a negative sampling of the alternatives is applied, i.e. for each positive choice (u, v) , we randomly select s nodes not connected to u in $[t_s, t_e]$. Once, the choice data have been selected, we compute for each pair (u, v) the feature vector x_v on the basis of the information provided by G_{t_s} and infer the parameters θ .

Table 1

Parameters inferred for the conditional logit model. Each column represents a stage of the evolution of Steemit.

	(1)	(2)	(3)	(4)	(5)	(6)
	2017-03-15	2017-04-15	2017-05-15	2017-06-15	2017-07-15	2017-08-15
log(DS)	0.699***	0.597***	0.512***	0.473***	0.374***	0.455***
HDS	-1.871***	-1.354***	-1.338***	-1.238***	-1.687***	-1.975***
RS	4.814***	5.239***	5.995***	5.697***	5.429***	5.003***
FoFS	1.128***	1.402***	1.167***	1.223***	1.631***	1.710***
log(CNS)	0.372***	0.533***	0.669***	0.609***	0.795***	0.528***
IF	0.359***	0.258***	0.590***	0.393***	0.222***	0.342***
Train accuracy	0.62	0.624	0.608	0.515	0.507	0.45
Test accuracy	0.635	0.622	0.612	0.505	0.509	0.448

Note:

* p<0.1; ** p<0.05; *** p<0.01

Social and financial features. In the discrete choice model approach, the selection of the right features is crucial, since features capture the different mechanisms and options a user takes into account to make a choice. Here we list the features extracted from the network made by links belonging to the category *social* and the network made by links belonging to the category *financial*:

1. **In-degree - social (DS)**: in-degree of the node in the social network;
2. **Has degree (HDS)**: boolean flag pointing out if $DS > 0$;
3. **Hops (HS)**: length of the shortest path between u and v in the social network;
4. **Reciprocity (RS)**: boolean flag indicating whether $(v, u) \in G_{t_s}$;
5. **Common neighbors - social(CNS)**: number of in-neighbors in common between u and v in the social network;
6. **Friends of friends (FoFS)**: boolean flag indicating if CNS is different from 0;
7. **In financial (IF)**: boolean flag indicating whether v is in the financial network;
8. **In-degree - financial(DF)**: number of transactions received by node v ;
9. **Transactions (TF)**: number of transactions exchanged between u and v
10. **Common neighbors - financial (CNF)**: number of neighbors in common between u and v in the financial network.

Results. In Table 1 we report the parameters of the conditional logit choice model which has achieved the best performance in terms of accuracy, i.e. the choice model whose random utility function is $U_{i,j} = \theta_1 \log(DS) + \theta_2 HDS + \theta_3 RS + \theta_4 FoFS + \theta_5 \log(CNS) + \theta_6 IF + \epsilon_{i,j}$. Each column reports the weights - θ - users have assigned to a particular feature when they chose to connect to other nodes during a specific stage of the growth of the Steemit social network. From the analysis of the parameters and their trends we observe that:

- the weights of the different features are quite stable in each of the time windows. This might be a possible indicator of the stability of the users' behavior when choosing to establish new relationships. It is also interesting to observe that the preferential attachment mechanism, captured by the feature $\log(DS)$, has lost its importance as the network evolved;

- reciprocating links (RS) and making connections with users with at least one common relationship ($FoFS$) have got the highest impact on the growth of the Steemit social network. Specifically, while the choice of reciprocating links has been strong and stable throughout the observation period, the mechanism based on common neighbors has strengthened its importance as the network evolved;
- in the choice of link formation the financial features are marginal and almost irrelevant. In fact, most of the financial features are not even present in the formulation of the utility function, and the only financial feature IF got values close to zero during the whole observation period.

To sum up, through discrete choice models we are able to disentangle the complexity behind the evolution of a Web3 social platform as Steemit, and identify which are the main growth mechanisms that are leading the evolution of Web3 social platforms.

5. User migration across platforms: when a shocking event happens

In the previous section we dealt with a certain level of stability of the mechanisms leading the evolution of Web3 platforms. In this section, we drastically change our setting by focusing on the effects of a shocking event such as a hard fork of the blockchain supporting a Web3 social platform - Steem in our case. Specifically, we are interested in the user migration consequent to the hard fork. In the light of the data representation described in Section 3, modeling user migration is quite straightforward: both the original blockchain - Steem - and the new branch - Hive - are described by two distinct evolution multidigraphs: \mathcal{G}^S and \mathcal{G}^H , respectively, with a common ancestor representing the multidigraph at fork time t_F . Given these definitions, we assign to each account a state. Specifically, a user u migrates - *migrant* - from platform S to H after a fork, if after t_F s/he does at least one action on H ; while a user u remains on the original platform - *resident* - if s/he keeps performing actions on the platform S and no actions on H after the fork event. We also introduce a third category - *inactive users*, i.e. people who are inactive or have abandoned both platforms. So, for each user, we are interested in the predictability of the choice to adopt a new platform given some user's characteristics or activities. In particular, we ask whether early signals indicating that s/he will move to a new platform exist. We cope with these research questions by casting this issue into a machine learning task, specifically a node classification task.

Definition: User migration prediction task. Given the graph G_{t_F} and considering the successive timestamps t' , where $t' > t_F$, we define the user migration prediction task as the prediction of a node migration in one of the successive time steps.

Indeed, the main goal is classifying a user as migrant or resident as a function of several features related to the structure of the temporal multidigraph describing the platform; whereas the assumption is that user features, at the network structure level, could be predictive of a future user migration. It is worth noting that features are extracted by exploiting both financial and social links.

Table 2

User migration prediction. Metrics (Weighted F1, Accuracy, Precision, Recall, AUC) are the average over a 5-fold cross-validation.

	F1 Weighted	Accuracy	Precision	Recall	AUC
Random Forest	0.722	0.737	0.790	0.870	0.626
Logistic Regression	0.622	0.601	0.837	0.559	0.635
Linear SVM	0.632	0.612	0.820	0.596	0.624
Gradient Boosting	0.706	0.707	0.797	0.801	0.628
MLP_128_64_32	0.683	0.679	0.795	0.756	0.615
MLP_100	0.689	0.686	0.791	0.773	0.613

Features and labels. As features we selected the most common node-level features used in many network-based prediction tasks. Such features encode information about a node and its neighborhood. Specifically, for each user in G_{t_F} , we compute in-degree and out-degree, weighted in-degree, Pagerank, neighborhood average degree, and local clustering coefficient. Alongside the structural information, we also include information on the status of nodes in the neighborhood: *i)* the percentage of inactive neighbors, i.e. the number of neighbors whose status is inactive at time t_F , divided by the total number of neighbors, and *ii)* the percentage of resident neighbors, i.e. the number of neighbors whose status is resident at time t divided by the total number of neighbors. In addition to the structural features, we compute a set of features related to the activity of the users before the hard fork, mostly inspired by the similar task of churn prediction: *i)* the number of active/inactive days in the three-month period before the fork, *ii)* the average number of daily actions in the three-month period before the fork, *iii)* the lifetime in the original blockchain, *iv)* the distance - in days - between the first and the last action in the three-month period before the fork, and *v)* the average length of the activity sessions². These features have been extracted from the financial and social layers, i.e. subgraphs of a multidigraph where links have a specific label, social or financial actions in this case.

Experimental setting. The experimental context is given by the migration from Steem to Hive. Hence, we rely on the Steem evolving multidigraph \mathcal{G}^S , and its financial and social layers. More precisely, we select the snapshot at fork time, $t_F = 2020/03/20$, at 2:00 PM. Then, we obtain the labels migrant and resident by inspecting the sequences of multidigraph \mathcal{G}^S and \mathcal{G}^H after the fork time. The two classes are imbalanced. For instance, social layer, where there is a more severe imbalance, residents are 3/4x more than migrants (66.9 %, 33.1%); vice-versa in the financial layer there are more migrants (56.1%) than residents (43.9%). In this case we deal with the sample imbalance by oversampling through the SMOTE method.

We perform the performance evaluation in a 5-fold cross-validation setting. For each fold, we apply oversampling on the training portion of the fold. Then, we train a model, compute a set of evaluation metrics and we average the performances over the five folds. To compare the performances of the different learning algorithms, we compute the main evaluation metrics for classification tasks: weighted F1, accuracy, precision, recall and AUC. For the classification task, we rely on standard machine learning methods: Logistic Regression, Random Forest,

²An activity session corresponds to a bursty train of actions as described in [18]. In this case, we use different thresholds to identify bursty trains.

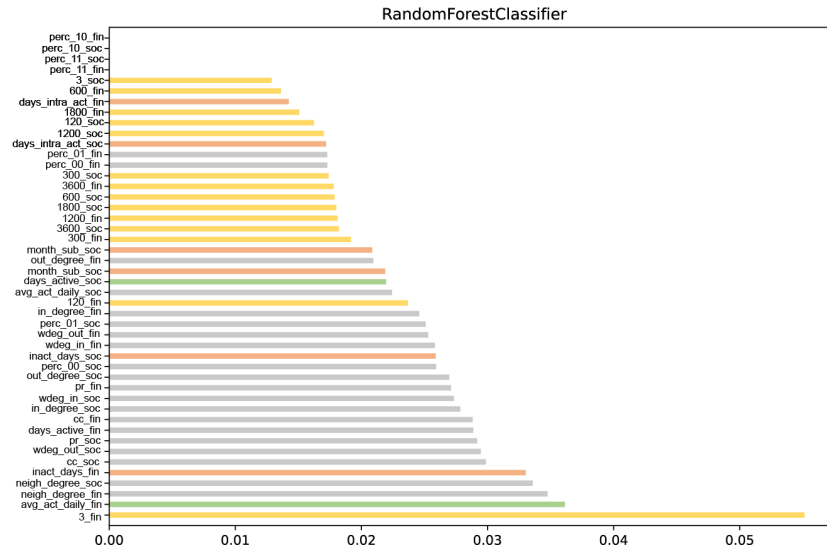


Figure 2: Feature importance for the best performing model, i.e. Random Forest, on the user migration prediction task. Importance values are based on the mean accumulation of the impurity decrease within each tree of the Random Forest.

Support Vector Machine with linear kernel, a Gradient Boosting classifier, and two multilayer perceptrons - MLP - with 100-unit hidden layer and three hidden layers (128 units - 64 units - 32 units), respectively.

Results. In Table 2 we report the results for the user migration prediction task. In this task, we are combining structural and activity-related features from both the social and financial layers, fully leveraging both the evolving graphs. This latter aspect is resulted crucial for improving the performances of all the models w.r.t. settings where features are taken from a single layer only - financial or social - have been considered. By comparing the model, as expected, the ensemble methods - Random Forest and Gradient Boosting - have obtained the best performances, which in general are quite good for the task. In addition, we observe that SVM and Logistic Regression have benefited from the addition of features coming from both layers. In short, the results suggest that in a stratified context with different types of links, interactions and features resulting from financial actions should be used together with social-based features to enhance the predictability of users in the case of a user migration.

Finally, we also performed a feature importance analysis to highlight the most predictive features to identify the early signals of willingness to migrate. The features ordered by their importance are depicted Figure 2. We observe that the most important features have been extracted from both social and financial layers and concern both structural and activity-related aspects. In fact, the average length of the session in the financial layer, and the neighbor degree in both social and financial ones are among the most important features. To sum up, this analysis on feature importance confirms the importance of taking into account information derived from both types of interaction actions when we tackle the user migration prediction task in Web3 social platforms.

References

- [1] J. Overgoor, A. Benson, J. Ugander, Choosing to grow a graph: Modeling network formation as discrete choice, in: *The World Wide Web Conference, WWW '19*, Association for Computing Machinery, New York, NY, USA, 2019, p. 1409–1420. URL: <https://doi.org/10.1145/3308558.3313662>.
- [2] R. Ciriello, R. Beck, J. Thatcher, The paradoxical effects of blockchain technology on social networking practices, in: *Proceedings of the Thirty Ninth International Conference on Information Systems, AIS*, 2018.
- [3] A. Kiayias, B. Livshits, A. M. Mosteiro, O. Litos, A puff of steem: Security analysis of decentralized content curation, *ArXiv abs/1810.01719* (2019).
- [4] U. W. Chohan, The concept and criticisms of steemit, *CBRI Working Papers: Notes on the 21st Century*, Available at SSRN: <http://dx.doi.org/10.2139/ssrn.3129410>, 2018.
- [5] M. S. Kim, J. Y. Chung, Sustainable growth and token economy design: The case of steemit, *Sustainability* 11 (2019) 167.
- [6] B. Guidi, A. Michienzi, L. Ricci, A graph-based socioeconomic analysis of steemit, *IEEE Transactions on Computational Social Systems* PP (2020) 1–12. doi:10.1109/TCSS.2020.3042745.
- [7] B. Guidi, A. Michienzi, L. Ricci, Steem blockchain: Mining the inner structure of the graph, *IEEE Access* 8 (2020). doi:10.1109/ACCESS.2020.3038550.
- [8] B. Guidi, A. Michienzi, L. Ricci, Analysis of witnesses in the steem blockchain, *Mobile Networks and Applications* (2021) 1–12.
- [9] R. Zhang, J. Park, R. Ciriello, The differential effects of cryptocurrency incentives in blockchain social networks, 2019.
- [10] M. Thelwall, Can social news websites pay for content and curation? the steemit cryptocurrency model, *Journal of Information Science* 44 (2018) 736 – 751.
- [11] C. Li, B. Palanisamy, Incentivized blockchain-based social media platforms: A case study of steemit, in: *Proceedings of the 10th ACM Conference on Web Science*, 2019, pp. 145–154.
- [12] K. Kapanova, B. Guidi, A. Michienzi, K. Koidl, Evaluating posts on the steemit blockchain: Analysis on topics based on textual cues, in: *Proceedings of the 6th EAI International Conference on Smart Objects and Technologies for Social Good, EAI*, 2020.
- [13] T.-H. Kim, H. min Shin, H. Hwang, S. Jeong, Posting bot detection on blockchain-based social media platform using machine learning techniques, *ArXiv abs/2008.12471* (2020).
- [14] P. Jia, C. Yin, Research on the characteristics of community network information transmission in blockchain environment, in: *IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, volume 1, IEEE, New York, NY, 2019, pp. 2296–2300.
- [15] C. T. Ba, M. Zignani, S. Gaito, The role of cryptocurrency in the dynamics of blockchain-based social networks: the case of steemit, *PloS One* (2022).
- [16] C. T. Ba, A. Michienzi, B. Guidi, M. Zignani, L. Ricci, S. Gaito, Fork-based user migration in blockchain online social media, in: *Proceedings of the 14th ACM conference on web science*, 2022.
- [17] C. T. Ba, M. Zignani, S. Gaito, The role of groups in a user migration across blockchain-based online social media, in: *2022 IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events (PerCom Workshops)*, IEEE, 2022, pp. 291–296.
- [18] C. T. Ba, M. Zignani, S. Gaito, Social and rewarding microscopical dynamics in blockchain-based online social networks, in: *Proceedings of the Conference on Information Technology for Social Good*, 2021, pp. 127–132.
- [19] S. Kumar, R. Zafarani, H. Liu, Understanding user migration patterns in social media, in: *AAAI*, 2011.
- [20] M. Senaweera, R. Dissanayake, N. Chamindi, A. Shyamalal, C. Elvitigala, S. Horawalavithana, P. Wijesekara, K. Gunawardana, M. I. E. Wickramasinghe, C. Keppitiyagama, A weighted network analysis of user migrations in a social network, 2018 *18th International Conference on Advances in ICT for Emerging Regions (ICTer)* (2018) 357–362.
- [21] C. Davies, J. R. Ashford, L. Espinosa-Anke, A. D. Preece, L. D. Turner, R. M. Whitaker, M. Srivatsa, D. H. Felmlee, Multi-scale user migration on reddit, in: *Workshop on Cyber Social Threats at the 15th International AAAI Conference on Web and Social Media (ICWSM 2021)*, AAAI, 2021.
- [22] P. Holme, J. Saramäki, Temporal networks, *Physics Reports* 519 (2012) 97–125. Temporal Networks.