Account Clustering in the Polkadot Network: Heuristic, Experiments, and Insights.

Maurantonio Caprolu* and Roberto Di Pietro*, *IEEE Fellow*

*Division of Information and Computing Technology, College of Science and Engineering
Hamad Bin Khalifa University, Qatar Foundation, Doha, Qatar

***Abstract*—This paper investigates, for the first time, user account clustering in the Polkadot network, one of the most innovative account-based altcoins in the market. To achieve this goal, we levereged the "deposit address reuse" heuristic on the Polkadot relay chain. In detail, we propose a novel deposit address detection methodology, combined with a general clustering strategy. To show the viability of our approach, we present a case study involving Binance and Kraken, the two major exchanges active in the Polkadot network. The analysis extends over a sensitive time window—starting from Polkadot genesis (May 2020) up to block 12,532,600 (October 2022). Thanks to the proposed methodology, we clustered more than 145,440 accounts belonging to exchanges, and more than 25,000 user accounts, representing around 25% of all the Binance/Kraken on-chain customers. The general applicability of our technique, the preliminary achieved results—showing both the viability and the value provided by our approach—, and the research hints discussed in the paper, also pave the way for further research in the field.**

***Index Terms*—Account Clustering, Cryptocurrency, Blockchain, Polkadot, Crypto Exchange**

I. Introduction

Cryptocurrency exchanges play an essential role in the new digital financial ecosystem by allowing anyone to buy, sell, and trade digital assets. On the one hand, being offered outside traditional economic systems, their services are cheap and characterized by a highly satisfactory user-experience. For these reasons, in recent years, they have received considerable attention, bringing even small investors closer to the world of finance. On the other hand, however, these services lack a legal framework that protects investors from market manipulation, unethical behaviors, and other frauds.

One of the major risks for cryptocurrency users is loss of privacy. Many users are, in fact, attracted to the world of cryptocurrencies by the high level of privacy provided, which usually ranges from pseudo-anonymity to more robust anonymization and unlinkability features. However, several deanonymization methodologies have been proposed to violate user privacy. These techniques can link an on-chain address to a real identity through different approaches, such as network analysis [1], artificial intelligence [2], and graph learning [3], to name a few. An immediate countermeasure to these attacks is to use multiple addresses for the same wallet. However, this additional level of privacy can be compromised by using different blockchain forensic methodologies [4], such as address clustering.

In this paper, for the first time in the literature, we investigate the applicability of address clustering in the Polkadot network and assess the resulting impact on user privacy. In particular, we investigated how Polkadot users lose privacy by interacting with crypto exchanges. Indeed, while the deposit of crypto-coins from the blockchain to the exchange platform is usually implemented in an efficient manner, it also opens up the flank to privacy attacks. To evaluate the impact on privacy of account clustering in the Polkadot network, we first design a novel mechanism to detect deposit addresses, i.e., bridge accounts used by exchanges to receive funds (on-chain) from clients. Then, we propose a case study involving the two major crypto exchanges active in the Polkadot network: Binance and Kraken. In particular, we applied the "deposit address reuse" heuristic over the two cited exchange networks, acquiring several insights into the interactions between exchanges and their customers. Although relatively new, Polkadot is gaining attention not only in the market but also in the Academia. After a seminal paper investigated its internals and highlighted its weaknesses [5], other recent works have begun to study Polkadot from different perspectives [6], [7].

Contributions. Our main contributions can be summarized as follows:

- We designed a new mechanism to detect deposit addresses that does not rely on forwarding times; this latter feature was proven unreliable by our experiments.
- We present a case study based on an extensive experimental campaign that involves two major crypto exchanges operating in the Polkadot network, Binance and Kraken, to assess the impact of account clustering on their customers' privacy.
- We clustered more than 145,440 accounts that belong to exchanges, and more than 25,000 user accounts, that is about the 25% of all the user addresses that interacted with the two exchanges in the observed period.
- We presented several insights on the two analyzed cryptocurrency exchanges and how they interact with their customers. In particular, we observed longer-than-usual forwarding times in the case of Binance, which questions the methodologies previously used in the literature to detect deposit addresses.

978-8-3503-1019-1/23/$31.00 ©2023 IEEE

II. Crypto Exchanges and Deposit Addresses

Centralize cryptocurrency exchanges work with "custodial trading"; customers are required to send and hold their assets in the exchange's wallet. Then, all trading takes place on the exchange platform without direct on-chain transactions. The user only appears in the public cryptocurrency ledger when sending crypto-coins to the exchange, and when withdrawing assets from the exchange's platform to on-chain addresses. In this paper, we are interested in the first interaction: when users send assets to the exchange platform. This operation consists of many steps: (i) the user, through the exchange platform, initiates a new deposit operation; (ii) the exchange typically creates what is called a "deposit address", which never appeared in the blockchain before; (iii) the user makes an on-chain transaction from a personal account to the deposit address; and, (iv) the exchanges forward all the assets received in the deposit address to its main address, and credits the same amount in the user account (inside the exchange platform).

The possibility of clustering user accounts through the reuse of deposit addresses is a known fact in the research community. However, despite being proven effective in the Ethereum blockchain [8], this technique is still underutilized. The logic of this heuristic is straightforward. Deposit addresses are created per customer. Once created, the deposit address remains the same for the same user (for simplicity and not to burden the cryptocurrency infrastructure). As a result, multiple addresses sending funds to the same deposit address will most likely be controlled by the same entity.

III. Methodology

Our methodology is divided into three consecutive steps: (i) data collection; (ii) deposit addresses detection; and, (iii) user account clustering.

Data Collection In the first step of our methodology we retrieve the transaction data from the Polkadot ledger. To this end, we set up a Polkadot full node in archive mode to download and query the entire ledger. After fully syncing the blockchain ledger on our full node, we query and parse the blocks data into a MySQL database, from the genesis to block 12,532,600 (mined on October 2022). To query our Polkadot full node, we have used two open-source Python libraries: `substrate-interface` to interact with the ledger and `scalecodec` for decoding transaction data, encoded in SCALE (Simple Concatenated Aggregate Little-Endian) format. Since we are only interested in user-initiated fund transfers, we have collected only extrinsics having the attributes reported in Listing 1.

```
module_id = Balances; and
call_id = Transfer or transfer_keep_alive or
transfer_all; and
signed = True; and
Success = 1
```

Listing 1. Attributes that identify an extrinsic as a user-initiated funds transfer successfully validated and stored in the ledger.

In this way, we stored all the extrinsics related to successful transfers of DOTs between users, i.e., the transaction amount was drawn from the sender's account and successfully deposited into the recipient's account. Consequently, we can easily retrieve all the extrinsics involving one or multiple addresses by querying the MySQL database. For the sake of simplicity, we refer to extrinsics as "transactions" in the rest of the paper.

Deposit Addresses Detection To detect deposit addresses among all the accounts in a cryptocurrency network, we leverage the deposit operations discussed previously in Section II. In particular, starting from a given account address $\mathcal{X}$, known to be owned by an exchange, we select a set of accounts suspected of being a deposit address, i.e., all the addresses that sent funds to $\mathcal{X}$ and never received any funds from $\mathcal{X}$. Subsequently, we refine this list by checking for the existence of a forwarding mechanism. For every account $\mathcal{Y}$ identified previously, we first retrieve all the transactions involving it, i.e., having $\mathcal{Y}$ either as a sender or receiver. Then, we sort those transactions by date. Finally, for every transaction $\mathcal{T}$, where $\mathcal{Y}$ is the sender and $\mathcal{X}$ is the receiver, we check if it exists one or more transactions that happen before $\mathcal{T}$, where $\mathcal{Y}$ is the receiver, with value (or the sum of the values) equal to the value of $\mathcal{T}$.

Address Clustering The address clustering procedure starts by iterating over the deposit addresses identified in the previous step. For every considered deposit address, the goal is to check if it has been reused. If this is the case, we can group in the same cluster all the different user accounts that paid the same deposit address. To this end, for every deposit address $\mathcal{Y}$ identified previously, we retrieved the list P of payers, i.e., user accounts that sent DOTs to $\mathcal{Y}$. Then, for every user account u in P, if u is included in another cluster P^1, we merge P and P^1 in the same cluster.

Possible Pitfalls The main objective of this study is to evaluate the loss of privacy that users suffer while interacting with exchanges. For this reason, we are interested in clustering users' accounts only. However, we have no guarantees that all deposit addresses identified with our methodology have been used by regular users. In fact, it is not uncommon for exchanges to interact with each other to buy and sell cryptocurrencies. For this reason, in order to ensure the correctness of our results and the reliability of our evaluation, we separated the deposit addresses that have received DOTs from an address that belongs to an exchange. Then, we clustered them separately to discover other accounts that, for the first time, can be labeled as owned by an exchange.

IV. Results

Deposit Address Detection For our experiments, we consider the account addresses of two popular exchanges operating in the Polkadot network, Binance and Kraken, reported in Listing 2. Our goal is to evaluate the impact of the deposit address reuse heuristic, also highlighting the differences between two distinct crypto exchange user communities. The results of our deposit addresses detection experiments are reported in Table I. The "overall payers" field reports the

TABLE I
RESULTS OF THE DEPOSIT ADDRESSES DETECTION FOR BINANCE AND KRAKEN ACCOUNTS.

	Binance	Kraken
Overall Payers	197,223	122,734
Overall Deposit Addresses	197,011	122,683
Used by other Exchanges	126,918	100,532
Used by Regular Users	70,093	22,151

total number of (unique) addresses that transferred funds to the main Binance/Kraken account.

```
Binance:
(1) 1exaAg2VJRQbyUBAeXcktChCAqjVP9TUxF3zo23R2T6EGdE
(2) 1qnJN7FViy3HZaxZK9tGAA71zxHSBeUweirKqCaox4t8GT7

Kraken:
(1) 12xtAYsRUrmbniiWQqJtECiBQrMn8AypQcXhnQAc6RB6XkLW
```

Listing 2. The addresses of accounts considered in our study as owned by cryptocurrency exchanges.

The "overall deposit addresses", instead, details the total number of (unique) accounts identified as deposit addresses. The other two fields, "used by other exchanges" and "used by regular users" detail the type of customer that used the deposit address, i.e., how many of them received funds from an address known to be owned by another exchange or not.

In the case of Binance, 99.8% of paying accounts were identified as deposit addresses. However, $\approx$ 64% of them have been used by other exchanges.Kraken, in turn, has fewer customers than Binance, but numbers are similar: 99.9% of all the payers are deposit addresses, 82% of which have been used by other exchanges. On the one hand, this results are not surprising. In fact, financial operations between different exchanges are very common in every cryptocurrency network. This is because every exchange needs to buy/sell coins on a regular basis to meet the daily needs of customers' trading. On the other hand, however, it is unexpected that such a large share of the overall deposit accounts is dedicated for trading between exchanges.

What highlighted, also showed another usage for the proposed methodology: to audit or signal anomalous transactions between crypto exchanges or entities. Indeed, our methodology would have likely highlighted an anomalous flow of transactions between FTX and Almenda Research [9].To summarize, from this experiment we can observe that $\approx$ 35% of the (unique) accounts that sent DOTs to Binance are regular users' deposit addresses. In the case of Kraken, instead, this percentage drops to around 18%.

After discovering the deposit addresses, we investigated the forwarding mechanism adopted by the two exchanges in terms of timing. To this end, we computed the time difference between the receipt of the DOTs (sent by the customer) to the deposit address, and the forwarding to the main address of the exchange (operated by the exchange itself). The result of our investigation on Binance and Kraken forwarding behavior is reported in figures 1 and 2, respectivelyFor this experiment, we

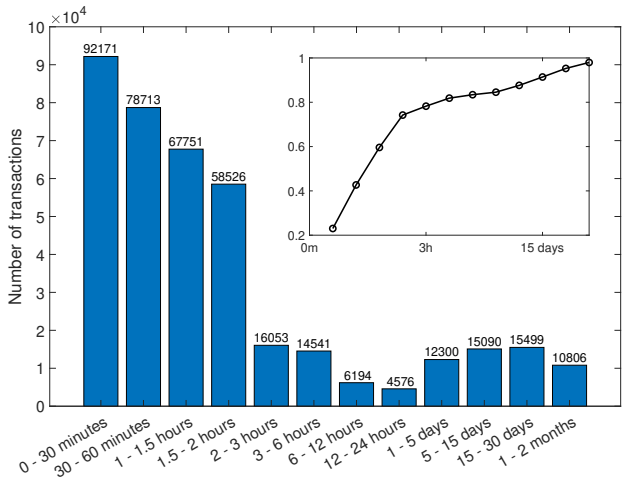

Fig. 1. Binance deposit addresses forwarding time. Each bin includes the number of times a forwarding happened within a particular time interval. For example, 19,164 transactions forwarded DOTs to the Binance main account between 30 minutes and 1 hour after being received at the deposit address. Inset figure: same data cumulative and normalized. Each point represents the probability that the forward occurred within the corresponding time.

considered all the discovered deposit addresses, i.e., used by either an exchange or a regular user, which amount to $400,384$ for Binance and to $326,512$ for Kraken.

Figure 1 shows that Binance forwarded approximately 80% of the funds to its main address within 2 hours of receipt. However, a considerable amount of forwarding transactions occurred several hours (and even months) after the funding transaction was received in the deposit account. This behavior is probably due to a fee minimization strategy implemented by Binance. Figure 2, instead, shows that Kraken adopts a completely different strategy. The 84% of the forwarding transactions happened within 5 minutes, the 97% within 10 minutes, while all the remaining transactions where settled within 24 hours.

User Account Clustering As discussed in Section III, we differentiate from deposit addresses used by regular users and exchanges to ensure the correctness of our evaluation. For this reason, before clustering all the discovered deposit addresses, we started from the deposit addresses labeled as "Used by other exchanges" in Table I, and we put in the same cluster all the addresses that sent DOTs to them. During the Binance experiment, we clustered 61,419 addresses. During the Kraken experiment, instead, we discovered 85,334 exchange addresses. Overall, we have clustered 145,440 unique accounts as belonging to exchanges. These addresses, for the first time associated with an exchange, could play a crucial role in future work aimed at investigating crypto exchanges.

After removing deposit addresses not related to regular users, we clustered the remaining deposit addresses. The results of these experiments are reported in Table II. In the case of Binance, the 70,093 discovered deposit accounts have received funds by 81,392 different user accounts. Around the 12% of the deposit accounts have been reused by Binance

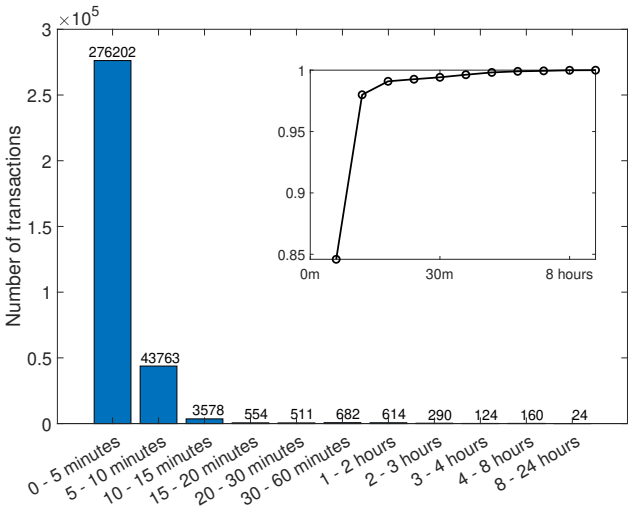

Fig. 2. Kraken deposit addresses forwarding time. Each bin includes the number of times a forwarding happened within a particular time interval. Inset figure: same data cumulative and normalized. Each point represents the probability that the forward occurred within the corresponding time.

TABLE II
RESULTS OF THE USER ACCOUNT CLUSTERING EXPERIMENT FOR BINANCE AND KRAKEN ACCOUNTS.

	Binance	**Kraken**
Deposit Addresses	70,093	22,151
User addresses involved	81,392	24,761
Deposit Addresses reused	8,621	1,680
User addresses clustered	22,970	3,700
Number of Clusters	8,355	1,516

customers. This behavior allowed us to cluster almost 23,000 different user addresses in 8,355 different clusters. In the case of Kraken, instead, the 22,151 discovered deposit accounts have received funds by 24,761 different user accounts. The 7.6% of the deposit addresses have been reused by Kraken customers, allowing us to cluster 3,700 user accounts into 1,516 different clusters.

Key Findings Our experiments give us several insights into how cryptocurrency exchanges operate in the Polkadot network. In particular, we can compare Binance and Kraken from different perspectives, allowing us to evaluate the impact of the exchange behavior on user privacy. First, from the data presented in Section IV, we can see that Binance has more on-chain customers than Kraken. In addition, more than one third of them are regular users. On the contrary, the vast majority of Kraken's on-chain customers seems to be other exchanges. Then, from figures 1 and 2, we can observe two different strategies for handling client funds during cryptocurrency withdrawals. Binance tends to group multiple incoming transactions on deposit accounts, likely to save on transaction fees. Kraken, on the contrary, forward all the received funds within few hours.

V. CONCLUSION

This paper presents, to the best of our knowledge, the first investigation on user account clustering in the Polkadot network. To achieve the cited goal, after observing that most of the existing heuristics cannot be applied to account-based architectures, we leverage the "deposit address reuse" heuristic combining it with a novel deposit address detection methodology.

We present a case study to assess the impact of the proposed technique on the privacy of Binance and Kraken customers. Our extensive experimental campaign started with the Polkadot genesis (May 2020) and ended up with block 12,532,600 (October 2022). We were able to cluster more than 145K addresses that belong to exchanges and around 25% of all the Binance/Kraken customers. Moreover, we extracted several other insights into the exchange behavior that calls for further investigations in this field. For instance: the two different strategies in the management of deposit addresses exhibited by the two considered exchanges may lead to incomplete results in existing clustering methodologies; the considerable number of inter-exchange transactions; and, the low deposit addresses reuse rate—but still sufficient to cluster a large share of user addresses.

ACKNOWLEDGMENTS

This publication was partially supported by the Qatar National Research Fund (QNRF), a member of The Qatar Foundation, through the awards [NPRP-S-11-0109-180242] and [NPRP11C-1229-170007]. The information and views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the QNRF.

REFERENCES

[1] A. Biryukov and S. Tikhomirov, "Deanonymization and linkability of cryptocurrency transactions based on network analysis," in *2019 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2019.

[2] F. Béres, I. A. Seres, A. A. Benczúr, and M. Quintyne-Collins, "Blockchain is watching you: Profiling and deanonymizing ethereum users," in *2021 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS)*. IEEE, 2021, pp. 69–78.

[3] A. Gaihre, S. Pandey, and H. Liu, "Deanonymizing cryptocurrency with graph learning: The promises and challenges," in *2019 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2019, pp. 1–3.

[4] M. Caprolu, M. Pontecorvi, M. Signorini, C. Segarra, and R. Di Pietro, "Analysis and patterns of unknown transactions in bitcoin," in *2021 IEEE International Conference on Blockchain (Blockchain)*, 2021, pp. 170–179.

[5] H. Abbas, M. Caprolu, and R. Di Pietro, "Analysis of polkadot: Architecture, internals, and contradictions," in *2022 IEEE International Conference on Blockchain (Blockchain)*, 2022, pp. 61–70.

[6] D. Morháč, V. Valaštín, and K. Koštál, "Sharing fungible assets across polkadot paraverse," in *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 2022, pp. 1–7.

[7] D. Haryadi, A. R. Hakim, D. M. U. Atmaja, and S. N. Yutia, "Implementation of support vector regression for polkadot cryptocurrency price prediction," *JOIV: International Journal on Informatics Visualization*, vol. 6, no. 1-2, pp. 201–207, 2022.

[8] V. Friedhelm, "Address clustering heuristics for ethereum," in *24th International Conference on Financial Cryptography and Data Security: Malaysia, February 2020*. Berlin, Heidelberg: Springer-Verlag, 2020.

[9] Y. Li Khoo, L. Choe, D. Chia, S. Leow, and N. Polk, "Blockchain Analysis: The Collapse of Alameda and FTX," https://www.nansen.ai/research/blockchain-analysis-the-collapse-of-alameda-and-ftx, Nansen Pte. Ltd., Online Report. Accessed: 2022-12-16.