CoinMagic: A Differential Privacy Framework for Ring Signature Schemes

Wangze Ni †, Han Wu *, Peng Cheng *, Lei Chen †, Xuemin Lin #, Lei Chen ‡, Xin Lai ‡, Xiao Zhang ‡
†*The Hong Kong University of Science and Technology, Hong Kong, China*
{wniab, leichen}@cse.ust.hk
**East China Normal University, Shanghai, China*
han.wu@stu.ecnu.edu.cn, pcheng@sei.ecnu.edu.cn
#*The University of New South Wales, Australia*
lxue@cse.unsw.edu.au
‡ *Shenzhen Onething Technologies Co., Ltd., Shenzhen, China*
{leichen, laixin, zhangxiao}@onething.net

ABSTRACT

By allowing users to obscure their transactions via including "mixins" (chaff coins), ring signature schemes have been widely used to protect a sender's identify of a transaction in privacy-preserving blockchain systems, like Monero and Bytecoin. However, recent works point out that the existing ring signature scheme is vulnerable to the "chain-reaction" analysis (i.e., the spent coin in a given ring signature can be deduced through elimination). Especially, when the diversity of mixins is low, the spent coin will have a high risk to be detected. To overcome the weakness, the ring signature should be consisted of a set of mixins with high diversity and produce observations having "similar" distributions for any two coins. In this paper, we propose a notion, namely ϵ-coin-indistinguishability (ϵ-CI), to formally define the "similar" distribution guaranteed through differential privacy schema. Then, we formally define the CI-aware mixins selection problem with disjoint-superset constraint (CIA-MS-DS), which aims to find a mixin set that has maximal diversity and satisfies the constraints of ϵ-CI and the budget. In CIA-MS-DS, each ring signature is either disjoint with or the superset of its preceding ring signatures. We prove that CIA-MS-DS is NP-hard and thus intractable. To solve the CIA-MS-DS problem, we propose two approximation algorithms, namely the Progressive Algorithm and the Game Theoretic Algorithm, with theoretic guarantees. Through extensive experiments on both real data sets and synthetic data sets, we demonstrate the efficiency and the effectiveness of our approaches.

PVLDB Reference Format:
Wangze Ni, Han Wu, Peng Cheng et al. CoinMagic: A Differential Privacy Framework for Ring Signature Schemes. *PVLDB*, 11 (3): xxxx-yyyy, 2017.
DOI: https://doi.org/xxx.xxxx/xxx.xxxx

1. INTRODUCTION

Recently, with the success of cryptocurrencies (e.g., Bitcoin [1], Ethereum [2]) and blockchain products (e.g., Thunderchain [3]),

Proceedings of the VLDB Endowment, Vol. 11, No. 3

DOI: https://doi.org/xxx.xxxx/xxx.xxxx

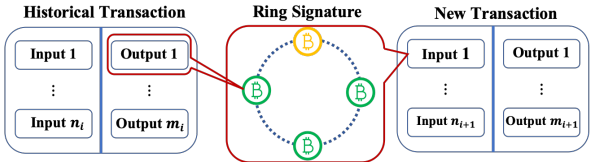
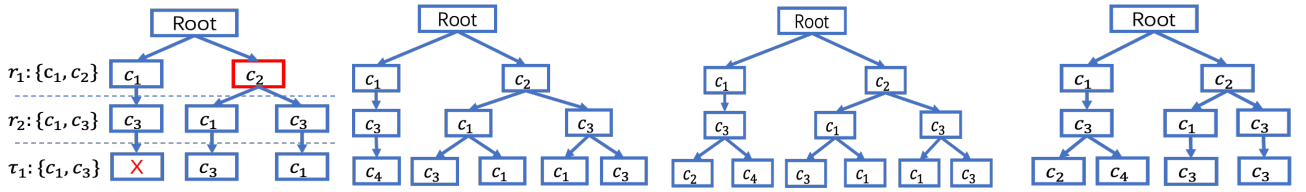


Figure 1: The Unspent Transaction Output (UTXO) model

blockchain technologies have attracted much attention from both academia (e.g., database community [4] [5]) and industry (e.g., supply chain management [6], healthcare [7] and bank [8]). In general, blockchain is a secure data structure maintained by untrusted peers in a decentralized P2P network. It has many valuable features such as transparency, provenance, fault tolerance, and authenticity.

However, transparency property of blockchain can initiate privacy problem, that is, in many real-world applications where users want to keep the transaction information by themselves. For instance, in a trading system, a trader may hope to conceal the information of her/his trade partner who received (sent) money from (to) her/him. Furthermore, the transaction information can be easily linked to a variety of other information that an individual usually wishes to protect. For example, in a blockchain-based ride-hailing system, like MVL [9], there are massive users' location data. By collecting and processing transaction data, it is possible to infer the user's personal information such as addresses of home [10].

To protect privacy, researchers have proposed some privacy definitions [11] [12] and privacy-preserving methods [12] [13] [14]. In [11], T. Okamoto and K. Ohta introduced that "privacy" must be one criterion of ideal electronic cash, which means "the relationship between the user and his purchases must be untraceable by anyone". Van Saberhagen proposed that "Untraceability" must be satisfied for a fully anonymous electronic cash model [12], which refers to "for each incoming transaction all possible senders are equiprobable". To satisfy the untraceability, *ring signature* (RS) schemes were widely implemented in famous privacy-preserving blockchain systems, like Monero [15] and Bytecoin [16]. Users can utilize RS schemes to obscure their transactions by including chaff coins, called "**mixins**", along with the coins they will spend. As shown in Figure 1, in the UTXO model blockchains, there are multiple inputs and outputs in a transaction. Each input is represented by a RS, and each output is a coin. Each RS contains a coin which is spent in the transaction, marked in yellow color, and many other coins as mixins marked in green color. Each coin is a historical transaction's output. The diversity of a RS's mixins is measured by the number of historical transactions outputting them. Since the RS scheme's efficiency (generation and verification time) affects its



(a) The permutation with τ_1 (b) The permutation with τ_2 (c) The permutation with τ_3 (d) The permutation with τ_4

Figure 2: Motivation Example.

verification time and transaction fee, researchers are motivated to improve its efficiency from $\mathbb{O}(n)$ in [12] to $\mathbb{O}(\log n)$ in [14], where n is the number of mixins in each RS.

In contrast to extensive researches in the RS scheme's efficiency, the researches of its effectiveness on preserving privacy remain rather scarce. Currently, the effectiveness of a RS is roughly estimated as the number of mixins, like many k-anonymity methods [17] [18], and the current RS scheme randomly picks mixins [15]. Current RS schema has two pivotal shortcomings: **a)** it is vulnerable to "chain-reaction" analysis [19]; **b)** it lacks of consideration on the diversity of mixins [20].

For the first weakness of current RS schema, through the "chain-reaction" analysis, it is possible to infer spent coins in RSs by leveraging the traffic flow and eliminating the mixins which must have been spent in other RSs. A RS r can be the significant affected by the related RSs which have overlapped coins with r. We illustrate the "chain-reaction" analysis in Example 1. For simplicity, in the remaining paper, we consider a ring signature as the union of mixins and the spent coin.

Example 1 ("Chain-Reaction" Analysis)**.** *There are four RS, $r_1 = \{c_1, c_2, c_3\}$, $r_2 = \{c_1, c_2\}$, $r_3 = \{c_1, c_2\}$, and $r_4 = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$ where $c_1 \sim c_7$ are seven coins. Then, for c_1 and c_2, r_2 produces **observations** to witnesses with high similarly level, since the probability of each coin being spent in r_2 is the same. The observation of a coin in a RS is the probability of the coin being spent in the RS. But for c_2 and c_3, r_1 produces observations to witnesses with low similarly level, since it is deduced that c_3 is the spent coin in r_1. Since r_2 and r_3 only contain c_1 and c_2, c_1 or c_2 is either spent in r_2 or r_3 (although their detailed matches are unknown) and c_3 must be the spent coin in r_1. Besides, r_4 has three useless mixins (i.e., c_1, c_2, c_3), which increases r_4's transaction fee.*

As the second weakness of current RS schema, it does not consider the impact of the diversity of mixins [20]. The diversity of a RS's mixins is measured by the number of historical transactions outputting them. When the diversity of mixins is low, the ring signature's effectiveness is low. In Example 1, if $c_5 \sim c_7$ are outputted by the same historical transaction t ($c_5 \sim c_7$ were spent in t) and c_4 is the spent coin in r_4, then the owner of the historical transaction t, who is not the user that spends coin c_4, can deduce that c_4 is the spent coin in r_4 (as $c_1 \sim c_3$ and $c_5 \sim c_7$ are already detected).

To avoid the "chain-reaction" analysis, we need to consider the related RSs' impact and make RSs produce observations with "similar" distribution for any two coins. To overcome the second defect, we need to find a mixin set with high diversity. Since the transaction fee is proportional to the number of mixins, the users usually want to restrain the number of mixins within limited budget. Thus, we need to pick a set of mixins with high diversity and effectively resist "chain-reaction" analysis based attacks under the constraint of the budget. To tackle this problem, two challenges need to be addressed: (1) for any two coins, how to measure the "similarity level" of the observations; and (2) how to pick a desired set of mixins to maximize their diversity under the constraint of budget.

In this paper, we propose a novel differential privacy concept, namely ϵ-coin-indistinguishability (ϵ-CI) to measure a RS's effectiveness. A RS's sender has ϵ-privacy if any two coins in the RS can produce observations with "similar" distributions, where the "level of similarity" depends on ϵ. The smaller ϵ is, the higher the privacy is. In the sequel, we illustrate the problem in a motivation example.

Example 2 (The Coin-Indistinguishability-Aware Mixins Selection Problem)**.** *Suppose there are four coins (c_1, c_2, c_3, c_4) and two RSs, $r_1 = \{c_1, \underline{c_2}\}$ and $r_2 = \{\underline{c_1}, c_3\}$. The spent coins in RSs are underlined. Among four coins, c_1 and c_4 are the same historical transaction's outputs while c_2 and c_3 are outputs of another two historical transactions. The budget is 3. Assume the required CI is very relaxed and only requires that the spent coin in a RS cannot be inferred. Now we want to generate a RS to spend c_3. As shown in Figure 2, the permutation of possible spent coins under given RSs can be presented with a permutation tree, where the nodes in each level indicate the possible spent coins for the corresponding RSs.*

The first solution is using RS $\tau_1 = \{c_1, c_3\}$. As shown in Fig. 2(a), although τ_1 cannot be inferred, it makes witnesses easily deduce that c_2 is the spent coin in r_1. For simplify, in the rest of this paper, we will only present the valid possible nodes of the permutation tree and ignore the corresponding RSs.

The second solution is using RS $\tau_2 = \{c_1, c_3, c_4\}$. As shown in Fig. 2(b), the CIs of τ_2, r_1, and r_2 are preserved. However, its diversity is not large. Since c_1 and c_4 are the outputs of the same historical transaction, τ_3's diversity is only 2.

The third solution is using RS $\tau_3 = \{c_1, c_2, c_3, c_4\}$. As shown in Fig. 2(c), the CIs of τ_3, r_1, and r_2 are preserved. Its diversity is large, which is 3. However, τ_3 does not meet the budget constraint, since its cardinality is 4, which is larger than the budget.

A good solution is using RS $\tau_4 = \{c_2, c_3, c_4\}$. As shown in Fig. 2(d), the CIs of τ_4, r_1, and r_2 are preserved. Besides, its diversity is large and it meets the budget constraint.

In this paper, we first formally define the coin-indistinguishability-aware mixins selection with disjoint-superset constraint (CIA-MS-DS) problem, which aims to find a mixin set that meets the required CI constraint, as well as the budget constraint, and has a maximal diversity. Moreover, each RS is either disjoint with or the superset of its preceding RSs. We prove the CIA-MS-DS problem is NP-hard through a reduction from the 01-knapsack problem [21] and we propose two approximation algorithms with theoretic guarantees. The Progressive Algorithm gradually narrows down the candidate mixin set one constraint by one constraint. The Game theoretic Algorithm models the problem as a game and let each mixin computes the right to be selected in the new RS.

To our knowledge, this is the first study to apply differential-privacy schema on Blockchain systems and formally estimate a RS's effectiveness. Our work provides new insights into how to pick mixins in RS schemes. Specifically, we make the following contributions:

- We define the notion of ϵ-coin-indistinguishability by applying differential-privacy schema on Blockchain systems in Section 2.
- We formally define the coin-indistinguishability-aware mixins selection with disjoint-superset constraint (CIA-MS-DS) problem and give the proof of its hardness in Section 3.
- We propose two approximation algorithms, the Progressive Algorithm and the Game Theoretic Algorithm, with theoretic guarantees for the CIA-MS-DS problem in Section 5 and Section 6 respectively.

- We conduct extensive experiments on both real and synthetic data sets and show efficiency as well as the effectiveness of our proposed solutions in Section 7.

Besides, we propose a framework, CoinMagic, in Section 4, discuss the related work in Section 8 and conclude in Section 9.

2. COIN-INDISTINGUISHABILITY

In this section, we formalize the concept of coin-indistinguishability. As aforementioned, coin-indistinguishability is utilized to guarantee that it is hard to distinguish the spent coin in a RS. Our proposal is based on a generalization of differential privacy [22]. Our notion and technique abstract from the side information of the adversary, such as prior probabilistic knowledge about a RS's mixins. The advantages of the independence from the prior are that: first, the mechanism is designed for any prior probabilistic. Second, and even more important, it is also applicable when we do not have the information of the prior probabilistic of mixins [10]. Because RSs can hide the spent coins, we even do not know whether the selected mixins have been spent or not.

2.1 The Related RS Set and Mixin Universe

As shown in Example 2, a RS's effectiveness is impacted by some related RSs. We first formally define the mixin universe and the related RS set. Suppose c_τ is the coin that a user wants to spend in a new RS.

Definition 1. (Mixin Universe) The mixin universe $\mathbb{C}_n = \{c_1, c_2, \cdots, c_n\}$ is a set of coins that can be picked up as mixins in a new RS. A coin c_i is the output of a transaction t_i.

In blockchain systems, each coin is a transaction's output. In a transaction, there may be more than one output. In Example 2, the mixin universe is $\mathbb{C} = \{c_1, c_2, \cdots, c_4\}$.

Definition 2. (Related Ring Signature Set) For a RS $r_f = \mathbb{C}_x \cup \{c_\tau\}$, the related RS set $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}$ is a set of RSs with earlier spending timestamps than r_f which contain the coin c_τ or any coins in $\mathbb{C}_x$.

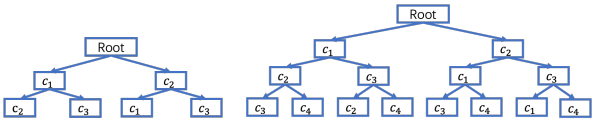
For instance, in Example 2, the related RS set is $\{r_1, r_2\}$. Since RSs in $\mathbb{R}_f$ may contain common coins with the new RS r_f, they may impact the effectiveness of r_f on privacy reserving. For a set of RSs $\mathbb{R}_f$, let $I_f(r_i)$ indicate the position of RS r_i in the the ascendingly ordered list of $\mathbb{R}_f$ sorted according to their spending timestamps.

2.2 Probabilistic Model of MIXINS

Since we do not know whether the mixins are spent or not, we introduce a probabilistic model here. Probabilities come into place in two ways. First, the adversary may have side information about the coins' expense, (e.g., knowing that some coins contained in a RS are not the spent coins since the adversary is the owner of these coins [20]). The adversary's side information can be modeled by a prior distribution $\pi(c, r)$ indicating the probability of coin c being spent in RS r. Second, RS r spending coin c is also a probabilistic event. Since the spent coin is obscured by mixins, any coin in the RS r is likely to be the spent coin.

Since RSs may not be disjoint and each coin can only be spent in a RS, given a set of RSs, there may be more than one possible *spent coin permutation* over the given RS set. Here, we formally define a spent coin permutation as follow.

Definition 3 (Spent Coin Permutation)**.** Given a related RS set, $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}$, a spent coin permutation $\mathbb{P}$ is an ordered list of coins $[sc_1, sc_2, \cdots, sc_m]$, where sc_i is a spent coin of a RS $r_j \in \mathbb{R}_f$ whose $I_f(j)$ is i and $\forall sc_h, sc_g \in \mathbb{P}$, $sc_h \neq sc_g$.



(a) $\mathbb{T}_2$ (b) $\mathbb{T}_3$

Figure 3: Example 3.

As shown in Fig. 2, we represent these permutations by a tree structure for easier understanding.

Definition 4 (Spent Coin Permutation Tree)**.** Given a related RS set, $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}$, we can generate a spent coin permutation tree $\mathbb{T}_f$. Except for the root node, each node in $\mathbb{T}_f$ is presented by $\kappa = \langle d_\kappa, \mathbb{P}_\kappa, sc_\kappa \rangle$, where d_κ is the node's depth in $\mathbb{T}_f$ (root's depth is 0), $\mathbb{P}_\kappa$ is a spent coin permutation $[sc_1^\kappa, sc_2^\kappa, \cdots, sc_{d_\kappa - 1}^\kappa]$, where each sc_i^κ is the spent coin of a RS $r_j \in \mathbb{R}_f$ whose $I_f(j) = i$, and sc_κ is the spent coin in this node (i.e., $sc_\kappa = sc_{d_\kappa}^\kappa$). For node κ, $sc_\kappa \notin \mathbb{P}_\kappa$. For any two nodes, κ and κ', $\mathbb{P}_\kappa \cup sc_\kappa \neq \mathbb{P}_{\kappa'} \cup sc_{\kappa'}$, and if $\mathbb{P}_{\kappa'} = \mathbb{P}_\kappa \cup sc_\kappa$, κ is the father node of κ'.

We can estimate the probability of each coin-and-signature pair by constructing its spent coin permutation tree. We denote $\mathbb{N}_{i,j}$ as the set of nodes in $\mathbb{T}_i$ whose depth is $I_i(j)$. Let $\mathbb{N}_{i,j}^{k,t}$ be the set of nodes in $\mathbb{T}_i$ whose depth is $I_i(j)$ and k^{th} element in $\mathbb{P}_\kappa$ is coin c_t.

We can calculate the probability of coin c_t being spent in the RS r_k when the permutation tree is $\mathbb{T}_i$ by:

$$Pr_i(c_t, r_k) = \frac{|\mathbb{N}_{i,i}^{k,t}|}{|\mathbb{N}_{i,i}|} \tag{1}$$

Thus, the probability of c_t having been spent in $\mathbb{R}_i$ can be calculated as:

$$Pr_i(c_t) = \frac{\sum_{k=1}^{i} |\mathbb{N}_{i,i}^{k,t}|}{|\mathbb{N}_{i,i}|} \tag{2}$$

Example 3. *There are three RSs,* $r_1 = \{c_1, c_2\}$, $r_2 = \{c_1, c_2, c_3\}$, $r_3 = \{c_1, c_2, c_3, c_4\}$. *The Figure 3(a) shows the permutation tree* $\mathbb{T}_2$ *and Figure 3(b) shows the permutation tree* $\mathbb{T}_3$. *Then,* $|\mathbb{N}_{2,2}| = 4$, $|\mathbb{N}_{3,3}| = 8$, $|\mathbb{N}_{2,2}^{2,3}| = 2$, $Pr_2(c_3, r_2) = \frac{1}{2}$, *and* $Pr_3(c_3, r_2) = \frac{1}{2}$.

2.3 Coin-Indistinguishability

We propose a formal definition of the coin-indistingushability to restrict the information leakage from observations. In other words, any adversary cannot obtain extra information from RSs w.r.t. the pre-defined privacy reserving level.

Definition 5. (ϵ-Coin-Indistinguishability (ϵ-CI)) Given a related RS set, $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}$, a RS $r_k \in \mathbb{R}_f$ satisfies ϵ-coin-indistinguishability (ϵ-CI) if for any two coins c_i, c_j which are both contained in the RS r_k (i.e., $c_i, c_j \in r_k$):

$$Pr_f(r_k|c_i) \leq e^\epsilon \cdot Pr_f(r_k|c_j)$$

$Pr_f(r_k|c_i) = \frac{Pr_f(r_k, c_i)}{Pr_f(c_i)}$ is the probability that when c_i is spent, it is spent in r_k. For instance, in Example 3, r_3 satisfies $\frac{\ln 8}{3}$-CI. If r_k satisfies ϵ-CI, for any $\bar{\epsilon} \geq \epsilon$, r_k also satisfy $\bar{\epsilon}$-CI.

2.4 Bounded Information Leakage

We prove a characterization that quantifies over all priors to explain how the prior affects the privacy guarantees.

Theorem 2.1. *An upper bound of the posterior distribution of coin* c_i *being spent in a RS* $r_k \in \mathbb{R}_f$ *which satisfies* ϵ*-CI can be obtained by* $Pr_f c_i|r_k) \leq e^\epsilon \cdot \frac{\pi(c_i, r_k)}{\sum_{c_j \in r_k} \pi(c_j, r_k)}$, *where* $\pi(c_i, r_k)$ *is the prior distribution modeled by the adversary's side information, indicating the probability of coin* c_i *being spent in the RS* r_k.

Proof. We can calculate the posterior distribution of coin c_i being spent in r_k by $Pr_f(c_i|r_k) = \frac{\pi(c_i,r_k)\cdot Pr_f(r_k|c_i)}{\sum_{c_j\in r_k}\pi(c_j,r_k)\cdot Pr_f(r_k|c_j)}$. Since r_k satisfies ϵ-CI, we have $Pr_f(c_i|r_k) = \frac{\pi(c_i,r_k)\cdot Pr_f(r_k|c_i)}{\sum_{c_j\in r_k}\pi(c_j,r_k)\cdot Pr_f(r_k|c_j)}$ $\leq \frac{\pi(c_i,r_k)}{e^{-\epsilon}\sum_{c_j\in r_k}\pi(c_j,r_k)} = e^{\epsilon}\cdot\frac{\pi(c_i,r_k)}{\sum_{c_j\in r_k}\pi(c_j,r_k)}$. □

The upper bound of posterior probability implies that no matter what prior information the adversary has, ϵ-CI constrains the multiplicative distance between the posterior distribution and prior distribution within e^{ϵ}, and thus limits the posterior information gain of the adversary.

3. PROBLEM DEFINITION

In this section, we first formally introduce some notions and then formulate the CI-aware mixins selection problem with disjoint-superset constraint (CIA-MS-DS). Besides, we give the proof of properties of the CIA-MS-DS problem, which can help to solve the problem more efficiently. Furthermore, we give the proof of the NP-hardness of the CIA-MS-DS problem.

3.1 Preliminaries

We first define the ϵ-CI-keeping RS as follow:

Definition 6. (A ϵ-CI-Keeping Ring Signature (ϵ-CIK-RS)) Given a related RS set $\mathbb{R}_f$, a RS r_f is a ϵ-CI-keeping RS if $\forall i \leq f, \forall c, c' \in r_i, Pr_f(r_i|c) \leq Pr_f(r_i|c')\cdot e^{\epsilon}$.

As aforementioned, subsequent RSs have impacts on the effectiveness of the previous RSs. To preserve the effectiveness existing RSs, new RSs should be ϵ-CIK-RS.

When each RS r_i in a related RS set $\mathbb{R}_f$ is disjoint with any RS r_j in $\mathbb{R}_i$ which is not the subset of r_i, the related RS set $\mathbb{R}_f$ is a **disjoint-superset related RS set**. It is practical for the real world applications since it does not divulge the users' privacy. Besides, it is helpful for calculating the conditional probabilities $Pr_f(r_k|c_i)$ for calculating the level of CI. We will introduce and give the proof os some properties of a disjoint-superset related RS set shortly in Subsection 3.3.

Definition 7. (Disjoint-Superset Related Ring Signature Set). For a related RS set, $\mathbb{R}_f$, it is a disjoint-superset related RS set if for any two RSs, they are disjoint or one RS is the superset of another RS.

For instance, in Example 3 $\{r_1, r_2, r_3\}$ is a disjoint-superset related ring signature set. However, in Example 2 $\{r_1, r_2\}$ is not a disjoint-superset related ring signature set, since $r_1 \cap r_2 = c_1$.

By Definition 7, given a disjoint-superset RS set $\mathbb{R}_f$, there are some special RSs in $\mathbb{R}_f$, namely *super ring signature*, whose subsequent RSs are all disjoint with it. We formally define this kind of RSs as follows.

Definition 8. (Super Ring Signature) Given a related RS set $\mathbb{R}_f$, a RS $r_i \in \mathbb{R}_f$ is a super RS if for any $r_j \in \mathbb{R}_f$ whose $I_f(r_i) < I_f(r_j)$, $r_i \nsubseteq r_j$.

By Definition 7, if a RS r_i is a super RS, $\forall I_f(j) \geq I_f(i)$, $r_i \cap r_j = \emptyset$. For example, there are four RSs, $r_1 = \{c_1, c_2\}$, $r_2 = \{c_1, c_2, c_3\}$, $r_3 = \{c_1, c_2, c_3\}$, and $r_4 = \{c_4, c_5\}$. Assume r_4 is the RS whose timestamp is the latest and $I_4(r_1) < I_4(r_2) < I_4(r_3) < I_4(r_4)$. Then, r_3 and r_4 are super RSs. Since $I_4(r_1) < I_4(r_3)$ and $r_1 \subset r_3$, r_1 is not a super RS. Similarly, r_2 is also not a super RS.

To more easily propose the problem and discuss its properties, we introduce some notations here. Given a related RS set $\mathbb{R}_f$, we denote $\mathbb{SRS}_f$ as the set of super RS, $\mathbb{SRS}_f = \{srs_1, srs_2, \cdots, srs_n\}$. For each super RS, srs_i, we define its diversity, $dive_i$, as the number of historical transactions outputting the coins in srs_i. We denote ns_i as the number of RSs in the related RS set $\mathbb{R}_f$ which are subsets of srs_i. We denote the degree d_i of srs_i as the number of coins in srs_i minus ns_i (i.e., $d_i = |srs_i| - ns_i$). Besides, we define the maximal coin spent probability, pr^i_{max}, of srs_i as the maximal value of the probability that a coin in srs_i has been spent in the related RS set $\mathbb{R}_f$, i.e., $pr^i_{max} = \max\{Pr_f(c)|c \in srs_i\}$, where $Pr_f(c)$ is defined as the probability that the coin c having been spent in $\mathbb{R}_f$ in Equation 2. Similarly, we define the minimal coin spent probability, pr^i_{min}, of srs_i as the minimal value of the probability that a coin in srs_i has been spent in the related RS set $\mathbb{R}_f$, i.e., $pr^i_{max} = \min\{Pr_f(c)|c \in srs_i\}$.

Among the mixin universe, there may be some coins that are not contained in any RSs. In Example 2, when we try to generate a RS to spend c_3, c_4 has not been contained in any RSs.

Definition 9. (Fresh Coin Set) Given a related RS, $\mathbb{R}$, and a mixin universe $\mathbb{C}$, a fresh coin set $\mathbb{F} = \{fc_1, fc_2, \cdots, fc_n\}$ is a set of coins in $\mathbb{C}$ that have not been contained in any RSs in $\mathbb{R}$.

3.2 The CI-aware Mixins Selection with Disjoint-superset Constraint Problem

In this subsection, we formally define the CIA-MS-DS problem.

Definition 10. (The CI-aware mixins selection with disjoint-superset constraint (CIA-MS-DS) problem) Given a super RS set $\mathbb{SRS}$, a fresh coin set $\mathbb{F}$, the coin c_τ that will be spent, a required ϵ, and a budget B, a user wants to pick up a set of mixins, combining c_τ, to generate the new RS r_τ, such that its diversity $dive_{r_\tau} = |\{t_i|c_i \in r_\tau\}|$ is maximized and the following constraints are satisfied:

- **DS constraint** r_τ is composed of some super RSs in $\mathbb{SRS}$ and some fresh coins in $\mathbb{F}$;
- **Budget constraint** the number of coins in r_τ does not exceed the budget; and
- **ϵ-CIK constraint** r_τ is a ϵ-CIK-RS.

Since the transaction fee of a RS is proportional to the number of coins in it, to limit the transaction fee of the new RS r_τ, the number of coins in r_τ should meet the budget constraint, i.e., $|\{c|c \in r_\tau\}| \leq B$. Besides, since the new RS r_τ should protect effectiveness of the existing RSs, it should be a ϵ-CIK-RS. Besides, when the new RS r_τ is composed of some super RSs in $\mathbb{SRS}$ and some fresh coins in $\mathbb{F}$, r_τ is disjoint with any RS in $\mathbb{R}_f$ which is not the subset of r_τ. This can help to quickly verify if the new RS r_τ is a ϵ-CIK-RS by the properties of the CIA-MS-DS problem, which will be introduced shortly in the next subsection. In addition, it can help other users to quickly generate the new RSs, which makes the problem practical for real world applications.

3.3 Properties of the CIA-MS-DS Problem

When a related RS set is a disjoint-superset related RS set, there are some important properties, which can help to calculate the attributes of each super RS (Theorem 3.1, 3.2 and 3.3) and reduce the time complexity of verifying if the generated RS r_τ is a ϵ-CIK-RS (Theorem 3.4, 3.5 and 3.6).

We first prove in Theorem 3.1 that when a related RS set is a disjoint-superset related RS set, the probability of coin c being spent in the RS r will keep stable.

Theorem 3.1. *Given a disjoint-superset related RS set,* $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}, \forall I_f(i) \in [2, m], \forall I_f(j) < I_f(i), \forall c_t \in r_j, Pr_i(r_j, c_t) = Pr_h(r_j, c_t)$, *where* $I_f(h) = I_f(i) - 1$.

Proof. By Definition 4 and 7, $\forall i \in [1,m]$, for each node κ in $\mathbb{N}_{i,i}$, we can partition $\mathbb{P}_\kappa$ as two sets $\mathbb{P}^1_\kappa$ and $\mathbb{P}^2_\kappa$, where $\mathbb{P}^1_\kappa \subseteq r_i$, $\mathbb{P}^2_\kappa \cap r_i = \emptyset$, and $\mathbb{P}^1_\kappa \cup \mathbb{P}^2_\kappa = \mathbb{P}_\kappa$. Thus, $|\mathbb{P}^1_\kappa| = ns_i$. Since each RS spent an unspent coin, $\forall i \in [1,m]$, $|r_i| > ns_i$. Thus, by Definition 4, $\forall I_f(i) \in [2,m], \forall I_f(j) < I_f(i)$, $\mathbb{N}_{i,j} = \mathbb{N}_{h,j}$ and for each node $\kappa \in \mathbb{N}_{i,h}$, it has d_i children. Thus, $\forall I_f(i) \in [2,m], \forall I_f(j) < I_f(i), \forall c_t \in r_j, \mathbb{N}^{j,t}_{i,h} = \mathbb{N}^{j,t}_{h,h}$.Thus, by Equation 1, $\forall I_f(i) \in [2,m]$, $\forall I_f(j) < I_f(i), \forall c \in r_j$, $Pr_i(r_j, c_t) = \frac{|\mathbb{N}^{j,t}_{i,i}|}{|\mathbb{N}_{i,i}|} = \frac{|\mathbb{N}^{j,t}_{h,h}| \cdot d_i}{|\mathbb{N}_{h,h}| \cdot d_i} = \frac{|\mathbb{N}^{j,t}_{h,h}|}{|\mathbb{N}_{h,h}|} = Pr_h(r_j, c_t)$. □

Next, we prove in Theorem 3.2 that when a related RS set is a disjoint-superset related RS set, the probability of a coin c_t being spent in the latest RS can be calculated iteratively.

Theorem 3.2. *Given a disjoint-superset related RS set $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}$, $\forall c_t \in r_f$, we have $Pr_f(r_f, c_t) = \frac{1-Pr_h(c_t)}{d_f}$, where $I_f(h) = I_f(f) - 1$.*

Proof. As proved in Theorem 3.1, each node in $\mathbb{N}_{f,h}$ has d_f children and $\forall c_k \in r_h, \mathbb{N}^{h,k}_{f,h} = \mathbb{N}^{h,k}_{h,h}$. Therefore, according to Equation 1, $\forall c_t \in r_f$, we have $Pr_f(r_f, c_t) = \frac{|\mathbb{N}^{f,t}_{f,f}|}{|\mathbb{N}_{f,f}|} = \frac{|\mathbb{N}_{f,h}| - \sum_{j=1}^{h} |\mathbb{N}^{j,t}_{f,h}|}{|\mathbb{N}_{f,h}| \cdot (d_f)}$ $= \frac{\frac{|\mathbb{N}_{f,h}|}{|\mathbb{N}_{f,h}|} - \frac{\sum_{j=1}^{h} |\mathbb{N}^{j,t}_{f,h}|}{|\mathbb{N}_{f,h}|}}{d_f} = \frac{1-Pr_h(c_t)}{d_f}$. □

Then, we show that when a related RS set is a disjoint-superset related RS set, the probability of a coin c having been spent in $\mathbb{R}_f$ can be calculated iteratively.

Theorem 3.3. *Given a disjoint-superset related RS set, $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}$, $\forall i \in [1,m], \forall c$, $Pr_i(c) = \begin{cases} Pr_j(c) & c \notin r_i \\ Pr_j(c) + \frac{1-Pr_j(c)}{d_i} & c \in r_i \end{cases}$, where $I_f(j) = I_f(i) - 1$, $I_f(g) = 0$, and $Pr_g(c) = 0$.*

Proof. As proved, $\forall i \in [2,m], \forall h < i, \forall c_t \in r_h$, $\mathbb{N}^{h,t}_{i,j} = \mathbb{N}^{h,t}_{j,j}$ and each node in $\mathbb{N}_{i,j}$ has d_i children. Thus, $\forall i \in [2,m], \forall c_t \in r_i$, $Pr_i(c_t) = \frac{\sum_{h=1}^{i} |\mathbb{N}^{h,t}_{i,i}|}{|\mathbb{N}_{i,i}|} = \frac{(|\mathbb{N}_{i,j}| - \sum_{h=1}^{j} |\mathbb{N}^{h,t}_{i,j}|) + d_i \cdot \sum_{h=1}^{j} |\mathbb{N}^{h,t}_{i,j}|}{|\mathbb{N}_{i,j}| \cdot d_i} =$ $\frac{\sum_{h=1}^{j} |\mathbb{N}^{h,t}_{i,j}|}{|\mathbb{N}_{i,j}|} + \frac{1 - \frac{\sum_{h=1}^{j} |\mathbb{N}^{h,t}_{i,j}|}{|\mathbb{N}_{i,j}|}}{d_i} = Pr_j(c_t) + \frac{1-Pr_j(c_t)}{d_i}$. Let $I_f(k) = 1$. By Equation 2, $\forall c \in r_k, Pr_k(c) = \frac{1}{|r_k|} = Pr_g(c) + \frac{1-Pr_g(c)}{d_k}$. Similarly, we can prove $\forall i \in [1,m], \forall c \notin r_i, Pr_i(c) = Pr_j(c)$. □

Thus, given a disjoint-superset related RS set, $\mathbb{R}_f$, we can easily retrieve the corresponding super RS set $\mathbb{SRS}$ and calculate the attributes of each super RS. Besides, for any two coins c, c' in the same RS r_i, we proved that if $Pr_i(c) \leq Pr_i(c')$, $\forall I_f(j) \geq I_f(i), Pr_j(c) \leq Pr_j(c')$.

Theorem 3.4. *Given a disjoint-superset related RS set, $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}$, $\forall i \in [1,m], \forall I_f(j) \leq I_f(i), \forall c, c' \in r_j$, if $Pr_j(c) \leq Pr_j(c')$, it holds that $Pr_i(c) \leq Pr_i(c')$.*

Proof. By Definition 7, $\forall i \in [1,m], \forall I_f(j) \leq I_f(i), \forall c, c' \in r_j$, if $c \in r_i$, $c' \in r_i$. Suppose $I_f(j) = I_f(i) - 1$. Thus, $Pr_i(c) - Pr_i(c') = Pr_j(c) + \frac{1-Pr_j(c)}{d_i} - Pr_j(c') - \frac{1-Pr_j(c')}{d_i} = [Pr_j(c) - Pr_j(c')] - \frac{1}{d_i} \cdot [Pr_j(c) - Pr_j(c')]$. Since $d_i \geq 1$ and $Pr_j(c) \leq Pr_j(c')$, $Pr_i(c) - Pr_i(c') \leq 0$. □

Besides, given a disjoint-superset related RS set $\mathbb{R}_f$, for any two coins c, c' in the same RS r_j, we proved that if $Pr_i(r_j|c) \leq Pr_i(r_j|c'), \forall I_f(k) \in [I_f(j)-1, I_f(i)], Pr_k(c) \geq Pr_k(c')$, where $I_f(j) \leq I_f(i)$.

Theorem 3.5. *Given a disjoint-superset related RS set, $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}$, $\forall i \in [1,m], \forall I_f(j) \leq I_f(i), \forall c, c' \in r_j$, if $Pr_i(r_j|c) \leq Pr_i(r_j|c')$, $\forall I_f(k) \in [I_f(j) - 1, I_f(i)], Pr_k(c) \geq Pr_k(c')$.*

Proof. Denote $F(t, j, i)$ as a function to recursively calculate $Pr_i(c_t)$ from $Pr_j(c_t)$. As proved in Theorem 3.4, $F(t, j, i)$ increases when $Pr_j(c_t)$ increases. By Definition 7, $\forall i \in [1,m], \forall I_f(j) \leq I_f(h), \forall c, c' \in r_j$, if $c \in r_i$, $c' \in r_i$. Denote $I_f(g) = I_f(j) - 1$. By Theorem 3.1 and 3.3, $\forall i \in [1,m], \forall I_f(j) \leq I_f(i), \forall c_t \in r_j$, $Pr_i(r_j|c_t) = \frac{Pr_i(r_j, c_t)}{Pr_i(c_t)} = \frac{Pr_j(r_j, c_t)}{F(t,g,i)} = \frac{\frac{1-Pr_g(c_t)}{d_j}}{F(t,g,i)}$. Thus, $Pr_i(r_j|c_t)$ increases when $Pr_g(c_t)$ decreases. Thus, $\forall i \in [1,m], \forall I_f(j) \leq I_f(i), \forall c, c' \in r_j$, if $Pr_i(r_j|c) \leq Pr_i(r_j|c')$, $Pr_g(c) \geq Pr_g(c')$. Then by Theorem 3.4, we have $\forall i \in [1,m], \forall I_f(j) \leq I_f(i), \forall c, c' \in r_j$, $\forall k \in [I_f(j) - 1, I_f(i)], Pr_k(c) \geq Pr_k(c')$. □

Then we prove that, if the conditional probabilities $Pr_h(r_j|c_t)$ of two coins in a RS satisfy ϵ-CI, the conditional probabilities of these two coins in subsequent RSs also satisfy ϵ-CI.

Theorem 3.6. *Given a disjoint-superset related RS set, $\mathbb{R}_f = \{r_1, r_2, \cdots, r_m\}$, $\forall i \in [1,m], \forall I_f(j) \leq I_f(i), \forall c, c' \in r_j$, if $Pr_i(r_j|c) \geq Pr_i(r_j|c')$ and $\frac{Pr_i(r_j|c)}{Pr_i(r_j|c')} = \beta$, $\forall k \in (I_f(h), m] \wedge r_i \subseteq r_k, Pr_k(r_j|c) \geq Pr_k(r_j|c')$ and $\frac{Pr_k(r_j|c)}{Pr_k(r_j|c')} = \beta' \leq \beta$.*

Proof. By Theorem 3.5, if $Pr_i(r_j|c) \geq Pr_i(r_j|c')$, $Pr_i(c) \leq Pr_i(c')$. Suppose r_k is the RS with the lowest $I_f(k)$ where $I_f(k) \geq I_f(i)$ and $r_i \subseteq r_k$. By Theorem 3.4, $Pr_k(c) \leq Pr_i(c')$ and $Pr_k(r_j|c) \geq Pr_k(r_j|c')$. Then, by Theorem 3.1 and 3.3, $\beta' = \frac{Pr_k(r_j|c)}{Pr_k(r_j|c')} = \frac{Pr_i(r_j,c)}{Pr_i(c) + \frac{1-Pr_i(c)}{d_h}} \cdot \frac{Pr_i(c') + \frac{1-Pr_i(c')}{d_h}}{Pr_i(r_j,c')} = \frac{Pr_i(r_j,c)}{Pr_i(r_j,c')} \cdot \frac{(d_i-1)Pr_i(c')+1}{(d_i-1)Pr_i(c)+1}$. When $d_i - 1 \geq 1$, $\beta' \leq \frac{Pr_i(r_j,c)}{Pr_i(r_j,c')} \cdot \frac{Pr_i(c')}{Pr_i(c)} = \beta$. When $d_i - 1 = 0$, since $\frac{Pr_i(c')}{Pr_i(c)} \geq 1$, $\beta' = \frac{Pr_i(r_j,c)}{Pr_i(r_j,c')} \leq \frac{Pr_i(r_j,c)}{Pr_i(r_j,c')} \cdot \frac{Pr_i(c')}{Pr_i(c)} = \beta$. □

Suppose the corresponding disjoint-superset related RS set of the given super RS set $\mathbb{SRS}$ is $\mathbb{R}_\tau$. Thus, if the new RS r_τ is composed by some super RSs in $\mathbb{SRS}$ and some fresh coins in $\mathbb{F}$, by Definition 8, r_τ is a super RS in $\mathbb{R}_\tau \cup r_\tau$. Then by Theorem 3.6, since RSs in $\mathbb{SRS}$ all satisfy the ϵ-CI, if r_τ satisfies ϵ-CI, it is a ϵ-CIK-RS. In other words, to verify if r_τ is a ϵ-CIK-RS, we just need to verify if it satisfies ϵ-CI, which decreases the time complexity form $\mathcal{O}(\sum_{i=1}^{|\mathbb{R}_\tau|+1} |r_i|^2)$ to $\mathcal{O}(|r_\tau|^2)$. Besides, suppose $pr_{max} = \max\{pr_m(c) | c \in r_\tau\}$ and $pr_{min} = \min\{pr_m(c) | c \in r_\tau\}$. Thus, by Definition 5, Theorem 3.3 and Theorem 3.2, to verify if r_τ satisfies ϵ-CI, we just need to verify if $\frac{(1-pr_{max}) \cdot e^\epsilon}{(d_\tau - 1) \cdot pr_{max} + 1} \geq \frac{(1-pr_{min})}{(d_\tau - 1) \cdot pr_{min} + 1}$, which further decreases the time complexity from $\mathcal{O}(|r_\tau|^2)$ to $\mathcal{O}(|r_\tau|)$, where d_τ is the degree of r_τ and is calculated as the number of coins in r_τ minus the number of its subsets in $\mathbb{R}_\tau$.

3.4 Hardness of the CIA-MS-DS Problem

In this subsection, we prove that the CIA-MS-DS problem is NP-hard by reducing 0-1 knapsack problem [21].

Theorem 3.7. *The CIA-MS-DS problem is NP-hard.*

Proof. We prove the theorem by a reduction from the 0-1 knapsack problem [21], which can be described as follows: Given a set, I, of n items i numbered from 1 up to n, each with a weight w_i and a

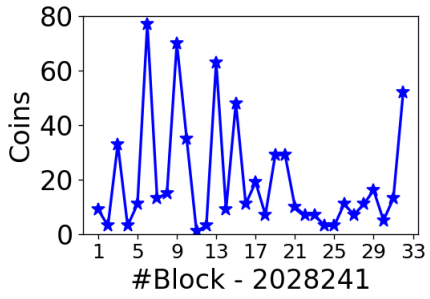


Figure 4: The Number of Coins in blocks

value x_i, along with a maximum weight capacity C, the 0-1 knapsack problem is to find a subset I' of I that maximizes $\sum_{i \in I'} x_i$ subjected to $\sum_{c_j \in r_k} w_i \leq C$.

For a given 0-1 knapsack problem instance, we can transform it into a special CIA-MS-DS problem instance as follows: we generate a super RS set, $\mathbb{SRS}$, where there are n super RS numbered from 1 up to n, each with a size x_i and diversity w_i. The fresh coin set is empty. The ϵ is large enough that any combination of super RSs in $\mathbb{SRS}$ is a ϵ-CIK-RS. Besides, for any two coins, the historical transactions which outputs them are different. Thus, $dive_{r_\tau} = \sum_{srs_i \in r_\tau} dive_i$.

Thus, to find the new RS r_τ whose diversity is maximum is equal to find a maximum assignment of 0-1 knapsack problem. Given this mapping, we can reduce the 0-1 knapsack problem to the CIA-MS-DS problem. Since the 0-1 knapsack problem is known to be NP-hard [21], the CIA-MS-DS problem is also NP-hard. □

4. COINMAGIC FRAMEWORK

Since the new RS's effectiveness on privacy preserving is impacted by the related RSs, before selecting mixins, a user needs to retrieve the mixin universe $\mathbb{C}$ and the related RS set $\mathbb{R}$. The mixin universe can be retrieved according to the user's interest or the blockchain system's requirement. For instance, in Monero [15], the system requires that in each RS, half of the mixins should be the coins which are less than 1.8 days old [23]. However, this method has two defects. Firstly, since the number of transactions in each block is unstable (especially, some miners even generate empty blocks), the cardinality of $\mathbb{C}$ is also not constant. For instance, Figure 4 shows the number of coins in blocks between 2028242 and 2028273 of Monero. Among these blocks, the block 2028252 only contains 1 coin but the block 2028247 contains 77 coins. Secondly, the size of the related RS set can unlimitedly increased over time, which is not efficient for constructing the spent coin permutation tree and generating new RSs.

In this paper, we propose a framework, namely CoinMagic, to retrieve the mixin universe $\mathbb{C}$ and related RS set $\mathbb{R}$, then generate new RSs. Specifically, CoinMagic partitions the blocks in a blockchain into disjoint and sequential batches, then generates new RSs only upon the coins and RSs in the same batch. The number of coins in each batch is bounded in range specified by the system. For each batch, its mixin universe is consisted of coins in the batch, Thus, the mixin universes for different batches are disjoint. In addition, as each RS selects mixins from the mixin universe in its batch, the related RS sets in different batches are also disjoint. The size of each related RS set is bounded by the cardinality of the mixin universe for its batch.

As shown in Algorithm 1, mixin universes are retrieved by batches. We start at the last block of the last batch before c_τ (line 1). Then, we initialize the mixin universe $\mathbb{C}$ and the related RS set $\mathbb{R}$ as empty (line 1). Then we continually read blocks and add the outputs of

Algorithm 1: CoinMagic Framework

Input: the spent coin c_τ, blocks in the blockchain system, $\mathbb{B}$, the threshold, λ, of the number of coins in a batch, budget, and the system's required CI level
Output: the new RS

1 set b_i as the last block of the last batch before c_τ, $\mathbb{C} = \emptyset$, $\mathbb{R} = \emptyset$;
2 **while** b_i *is not the last block in* $\mathbb{B}$ *and* $|\mathbb{C}| < \lambda$ **do**
3 $b_i \leftarrow$ next block;
4 **if** b *is not an empty block* **then**
5 add the output coins of transactions in block b to $\mathbb{C}$;
6 **while** b_i *is not the last block in* $\mathbb{B}$ **do**
7 $b_i \leftarrow$ next block;
8 **foreach** *RS* r *in* b_i **do**
9 **if** r *is a subset of* $\mathbb{C}$ **then**
10 add r to $\mathbb{R}$;
11 run a RS generation algorithm to generate a new RS satisfying the constraints of SRS, budget and ϵ-CIK in Definition 10.

transactions in blocks to the mixin universe $\mathbb{C}$ until we reach the last block in the blockchain or the cardinality of $\mathbb{C}$ is large enough (line 2-5). Then, we continually read blocks and add the RSs in blocks which are the subsets of $\mathbb{C}$ to $\mathbb{R}$ (line 6-10). Since the size of every block is limited, the number of coins in a block has an upper bound λ'. Thus, $|\mathbb{C}| \in [\lambda, \lambda + \lambda' - 1]$ and $|\mathbb{R}| \leq |\mathbb{C}| \leq \lambda + \lambda' - 1$. This procedure (line 1 -10) can be accomplished when the user updates her/his local blockchain status to the global status. In other words, the user does not need to pay much extra cost for implementing this framework. Finally, we run a RS generation algorithm to generate a new RS satisfying the constraints of SRS, budget, and ϵ-CIK.

Since the mixins of a RS can only be selected within the same batch, there could be a situation that a user cannot find a RS with ϵ-CI to spend a coin. For example, in Example 2, if the user uses the RS $\tau_5 = \{c_1, c_2, c_3\}$, then when another user wants to spend c_4, she/he cannot find a RS r with ϵ-CI, since witnesses can easily find that c_1, c_2, and c_3 has been spent in previous three RSs (i.e., r_1, r_2, r_3) and c_4 must be the spent coin in the new RS r. We denote the set of fresh coins in a batch as $\mathbb{F}$. To avoid this problem, we required that at any moment, the $\mathbb{F}$'s cardinality cannot be 1. Next, we formally prove that if $|\mathbb{F}| \neq 1$, for any c_τ and a budget, there always exists at least one RS which satisfies the constraints of SRS, budget and ϵ-CIK.

Theorem 4.1. *If $|\mathbb{F}| \neq 1$, for any c_τ and a budget B, there always exists at least one RS which satisfies the SRS constraint, ϵ-CIK-RS constraint, and the budget constraint, where m_τ is the module which contains c_τ and $B \geq \max\{|m_\tau|, 2\}$.*

Proof. When m_τ is a super RS, according to Theorem 3.6, $r_{m+1} = m_\tau$ is a ϵ-CIK-RS and its size satisfies the budget constraint. When m_τ is a fresh coin, since $|\mathbb{F}| \neq 1$, there is at least one fresh coin, denoted by fc. By Definition 5, $r_{m+1} = m_\tau \cup fc$ is also a CIK-RS and its size satisfies the budget constraint. □

5. PROGRESSIVE APPROACH

In this section, to tackle the CIA-MS-DS problem, we propose an approach, namely the "Progressive" Algorithm. The main idea of the approach is that we first use a 0-1 knapsack algorithm to find a selection which satisfies the ϵ-CIK-RS constraint and the SRS

Algorithm 2: Progressive Algorithm

Input: A spent coin c_τ, a budget B, a super RS set $\mathbb{SRS}$, and a fresh coin set $\mathbb{F}$

Output: An eligible RS r_τ

1 $\mathbb{M} = \mathbb{SRS} \cup \mathbb{F}$, $r_\tau = m_\tau$;
2 **for** $i = 1$ *to* $|M|$ **do**
3 $M_i = M \backslash \bigcup_{m_k \in M \wedge Pr^k_{max} > Pr^i_{max}} m_k$;
4 **for** $j = 1$ *to* $|M|$ **do**
5 **if** $m_j \in M_i$, $Pr^i_{max} \geq Pr^\tau_{max}$ *and* $Pr^j_{min} \leq Pr^\tau_{min}$ **then**
6 $\omega_{i,j} = m_i \cup m_j \cup m_\tau$;
7 $M_{i,j} = M_i \backslash (\omega_{i,j} \cup \bigcup_{m_k \in M_i \wedge Pr^k_{min} < Pr^j_{min}} m_k)$;
8 update the diversity of each m in $M_{i,j}$;
9 calculate $\bar{d}$ and $O^{i,j}_{max}$;
10 $C_{i,j} = \delta$-KP$(M_{i,j}, \bar{d}, O^{i,j}_{max})$;
11 **if** $|C_{i,j}| + |\omega_{i,j}| > B$ **then**
12 $\psi_{i,j} = \omega_{i,j}$;
13 **while** $|\psi_{i,j}| \leq B$ **do**
14 greedily add $\bar{m}$ in $M_{i,j}$ whose ratio of the increase of $\psi_{i,j}$'s diveristy to its size is the largest to $\psi_{i,j}$;
15 $M_{i,j} = M_{i,j} \backslash \bar{m}$;
16 **else**
17 $\psi_{i,j} = \omega_{i,j} \cup C_{i,j}$;
18 **return** *the* $\psi_{i,j}$ *with maximal diversity as* r_τ

constraint. Then we greedily change the selection to let it satisfy the budget constraint.

5.1 Progressive Algorithm

The algorithm is inspired by the following three lemmas.

The first lemma shows that the degree of the new RS r_τ in $\mathbb{R}_f$, denoted by d_τ, is the sum of the degree of each super RS and fresh coin which is contained in r_τ, i.e., $d_\tau = \sum_{srs_i \subset r_\tau} d_i + \sum_{fc_i \subset r_\tau} d_i$. Similar with the definitions of a super RS's attributes in Subsection 3.1, we define the attributes of a fresh coin as following. For each fresh coin, fc_i, its diversity $dive_i = 1$, $ns_i = 0$ and its diversity $d_i = 1$, since there is only one fresh coin in fc_i which is not contained in any RS in $\mathbb{R}_f$. Besides, since fc_i is not contained in any RS in $\mathbb{R}_f$, its $Pr^i_{max} = Pr^i_{min} = 0$.

Lemma 5.1. *For the new RS r_τ, its degree d_τ is the sum of the degree of each super RS and fresh coin which is contained in r_τ, i.e., $d_\tau = \sum_{srs_i \subset r_\tau} d_i + \sum_{fc_i \subset r_\tau} d_i$.*

Proof. Suppose Sub_i is the set of RSs in $\mathbb{R}_f$ which are the subsets of srs_i and Sub_τ is the set of RSs in $\mathbb{R}_f$ which are the subset of r_τ. Thus, $d_\tau = |srs_i| - |Sub_\tau| = (\sum_{srs_i \subset r_\tau} |srs_i| + \sum_{fc_i \subset r_\tau} 1) - |Sub_\tau|$. By Definition 7, $\forall srs_i, srs_j \in \mathbb{SRS}_f$, $Sub_i \cap Sub_j = \emptyset$. Besides, $\forall srs_i \subseteq r_\tau, \forall r \in Sub_i$, $r \in Sub_\tau$, since $r \subseteq srs_i \subseteq r_\tau$. Therefore,$|Sub_\tau| = \sum_{srs_i \subset r_\tau} ns_i$. Therefore, $d_\tau = (\sum_{srs_i \subset r_\tau} |srs_i| + \sum_{fc_i \subset r_\tau} 1) - \sum_{srs_i \subset r_\tau} ns_i = (\sum_{srs_i \subset r_\tau} |srs_i| \sum_{srs_i \subset r_\tau} ns_i) + \sum_{fc_i \subset r_\tau} 1 = \sum_{srs_i \subset r_\tau} d_i + \sum_{fc_i \subset r_\tau} d_i$. □

As defined in Subsection 3.3, for a RS r_τ, $Pr_{max} = \max\{pr_m(c) | c \in r_\tau\}$ and $pr_{min} = \min\{pr_m(c) | c \in r_\tau\}$. We will show that when $e^\epsilon \cdot Pr_{min} \cdot (1 - Pr_{max}) - Pr_{max} \cdot (1 - Pr_{min}) \geq 0$, the new RS must be a ϵ-CIK-RS (which is defined in Definition 6). Besides, when $e^\epsilon \cdot Pr_{min} \cdot (1 - Pr_{max}) - Pr_{max} \cdot (1 - Pr_{min}) \leq 0$, if the degree d_τ of the new RS r_τ is smaller than $\frac{(e^\epsilon - 1) \cdot (1 - Pr_{max}) \cdot (Pr_{min} - 1)}{e^\epsilon \cdot Pr_{min} \cdot (1 - Pr_{max}) - Pr_{max} \cdot (1 - Pr_{min})}$, the new RS r_τ must be a ϵ-CIK-RS.

Lemma 5.2. *For a new RS r_τ, it is a ϵ-CIK-RS, if $e^\epsilon \cdot Pr_{min} \cdot (1 - Pr_{max}) - Pr_{max} \cdot (1 - Pr_{min}) \geq 0$. Besides, it is a ϵ-CIK-RS, if $e^\epsilon \cdot Pr_{min} \cdot (1 - Pr_{max}) - Pr_{max} \cdot (1 - Pr_{min}) < 0$ and $d_\tau \leq \frac{(e^\epsilon - 1) \cdot (1 - Pr_{max}) \cdot (Pr_{min} - 1)}{e^\epsilon \cdot Pr_{min} \cdot (1 - Pr_{max}) - Pr_{max} \cdot (1 - Pr_{min})}$.*

Proof. As proved in Subsection 3.3, to verify if the new RS r_τ is a ϵ-CIK-RS, we just need to verify if $\frac{(1 - pr_{max}) \cdot e^\epsilon}{(d_\tau - 1) \cdot pr_{max} + 1} \geq \frac{(1 - pr_{min})}{(d_\tau - 1) \cdot pr_{min} + 1}$. Since $(d_\tau - 1) \cdot Pr_{max} + 1$ and $(d_\tau - 1) \cdot Pr_{min} + 1$ must be positive, to verify $\frac{(1 - pr_{max}) \cdot e^\epsilon}{(d_\tau - 1) \cdot pr_{max} + 1} \geq \frac{(1 - pr_{min})}{(d_\tau - 1) \cdot pr_{min} + 1}$ is the same as to verify $e^\epsilon \cdot (1 - Pr_{max}) \cdot [(d_\tau - 1) \cdot Pr_{min} + 1] \geq [(d_\tau - 1) \cdot Pr_{max} + 1] \cdot (1 - Pr_{min})$.

Suppose $X = e^\epsilon \cdot Pr_{min} \cdot (1 - Pr_{max}) - Pr_{max} \cdot (1 - Pr_{min})$ and $Y = (e^\epsilon - 1) \cdot (1 - Pr_{max}) \cdot (Pr_{min} - 1)$. Thus, to verify if $\frac{(1 - pr_{max}) \cdot e^\epsilon}{(d_\tau - 1) \cdot pr_{max} + 1} \geq \frac{(1 - pr_{min})}{(d_\tau - 1) \cdot pr_{min} + 1}$, we just need to verify if $X \cdot d_\tau \geq Y$. Since $e^\epsilon \geq 1$, $Pr_{max} \leq 1$ and $Pr_{min} \leq 1$, $Y \leq 0$. Thus, if $X \geq 0$, it always hold that $X \cdot d_\tau \geq Y$, which means, r_τ is a ϵ-CIK-RS. Besides, if $X < 0$ and $d_\tau \leq \frac{Y}{X}$, it holds that $X \cdot d_\tau \geq \frac{Y}{X}$, which means r_τ is a ϵ-CIK-RS. □

Besides, suppose O_{max} is the maximal number of coins which are outputted by the the same historical transaction. We show the relationship between the diversity of the new RS r_τ and the summation of the diversity of each super RS and fresh coin which is contained in r_τ.

Lemma 5.3. *For a new RS r_τ, we have $\sum_{srs_i \subseteq r_\tau} dive_i + \sum_{fc_i \subseteq r_\tau} 1 \geq dive_\tau \geq \sum_{srs_i \subseteq r_\tau} \frac{dive_i}{O_{max}} + \sum_{fc_i \subseteq r_\tau} \frac{1}{O_{max}}$, where d_τ is the diversity of the new RS r_τ.*

Proof. In a set of coins, some coins may be outputted by the same historical transaction. Since O_{max} is the maximal number of coins which are outputted by the same historical transaction, it must hold that $dive_\tau \geq \frac{|r_\tau|}{O_{max}}$. Since for any super RS srs_i its diversity $dive_i \leq |srs_i|$, $\frac{dive_i}{O_{max}} \leq \frac{|srs_i|}{O_{max}}$. Therefore, $dive_\tau \geq \frac{|r_\tau|}{O_{max}} = \frac{\sum_{srs_i \subseteq r_\tau} |srs_i| + \sum_{fc_i \subseteq r_\tau} 1}{O_{max}} \geq \sum_{srs_i \subseteq r_\tau} \frac{dive_i}{O_{max}} + \sum_{fc_i \subseteq r_\tau} \frac{1}{O_{max}}$.

In addition, denote the set of historical transactions which output coins in srs_i as HT_i and the set of historical transactions which output coins in r_τ as HT_τ. Since a historical transaction set HT_i may intersect with another historical transaction set HT_j on some historical transactions, $|HT_i| + |HT_j| \geq |HT_i \cup HT_j|$. Thus, $d_\tau = |HT_\tau| = |\bigcup_{srs_i \subseteq r_\tau} HT_i \cup \bigcup_{fc_i} t_i| \leq \sum_{srs_i \subseteq r_\tau} |HT_i| + \sum_{fc_i \subseteq r_\tau} 1 = \sum_{srs_i \subseteq r_\tau} dive_i + \sum_{fc_i \subseteq r_\tau} 1$. □

Thus, by Lemma 5.1 and Lemma 5.2, we can transform the ϵ-CIK-RS constraint to the constraint of degree. Specifically, we can enumerate all $Pr_{max} - Pr_{min}$ pairs, and for each pair, we calculate the upper bound of the degree of the new RS, denoted as $\bar{d}$. We require the new degree of the new RS cannot exceed $\bar{d}$. In addition, by Lemma 5.3, we can approximate estimate the diversity of the new RS.

Inspired by the aforementioned lemmas, we design the Progressive Algorithm. Algorithm 2 shows the pseudo-code of our Progressive Algorithm. By the SRS constraint, r_τ is composed of some super RS in $\mathbb{SRS}$ or some fresh coins in $\mathbb{F}$. In other words, each super RS and fresh coin is a candidate module of the new RS r_τ. Thus, we get the module set $\mathbb{M}$ by combing $\mathbb{SRS}$ and $\mathbb{F}$ (line 1). Specifically, like the definitions of a super RS in Subsection 3.1, we define the attributes of each fresh coin. Since a fresh coin fc_i is not

been contained in any RS, its $Pr^i_{max} = Pr^i_{min} = 0$. Besides, the diversity of fc_i is $dive_i = 1$ and its degree is $d_i = 1$. By the SRS constraint, r_τ has to contain m_τ, where m_τ is the module in $\mathbb{M}$ that contains c_τ (line 1). Then we enumerate all Pr_{max}-Pr_{min} pairs (line 2 -17), where m_i is the module in the new RS whose Pr^i_{max} is the maximum and m_j is the module in the new RS whose Pr^i_{min} is the minimum. Thus, for the given Pr_{max}-Pr_{min} pair, $\omega_{i,j} = m_i \cup m_j \cup m_\tau$ must be contained in the new RS (line 6). Besides, the set of candidate modules, which can be selected in the new RS, is $M_{i,j} = M \backslash (\bigcup_{Pr^k_{max} > Pr^i_{max}} m_k \cup \omega_{i,j} \cup \bigcup_{Pr^k_{min} < Pr^j_{min}} m_k)$ (line 7). In addition, we update the diversity of each module in $M_{i,j}$ (line 8). Specifically, the diversity of each module is updated by the number of transactions outputting the coins in modules, excluding the transactions outputting the coins in $\omega_{i,j}$. Then, we calculate the upper bound of the degree of the new RS and the $O^{i,j}_{max}$, which is the maximal number of coins in $M_{i,j}$ which are outputted by the same transaction (line 9). Then, we run the δ-KP Algorithm (line 10), where the item set is $M_{i,j}$, the weight of the item m_i is d_i, the value of the item m_i is $\frac{dive_i}{O^{i,j}_{max}}$, and the capacity of the knapsack is $\bar{d}$. The δ-KP Algorithm is the dynamic programming algorithm [21] whose precision parameter is δ. The selection $C_{i,j}$ from the δ-KP Algorithm may violate the budget constraint. Then, we greedily select modules in $C_{i,j}$ to let $\psi_{i,j}$ satisfy the budget constraint (line 11-15). For each module m_t in the selection, we calculate its increase ratio $\rho_t = \frac{dive_{\psi'_{i,j}} - dive_{\psi_{i,j}}}{|m_t|}$, where $\psi'_{i,j} = \psi_{i,j} \cup m_t$. For each iteration, we add the module m_t with the largest ρ_t (14). If the selection $C_{i,j}$ from the δ-KP Algorithm satisfies the budget constraint, we set $\psi_{i,j} = \omega_{i,j} \cup C_{i,j}$ (line 16-17). Finally, we return the $\psi_{i,j}$ with the largest diversity as r_τ (line 18).

5.2 Theoretic Analyses

We first give the proof of the approximate ratio of the δ-KP Algorithm.

Theorem 5.1. *Suppose $C^*_{i,j}$ is the selection of modules in $M_{i,j}$ whose degree is smaller than $\bar{d}$ and the number of historical transactions outputting coins in $C^*_{i,j}$ is the largest. Denote the diversity of $C^*_{i,j}$ as $dive^*$ and the diversity of $C_{i,j}$ as $dive_C$. It holds that $\frac{dive_C}{dive^*} \geq \frac{1-\delta}{O_{max}}$.*

Proof. Denote the value of each module m_t in δ-KP as $dive^\#_t$. Denote the optimal selection of δ-KP when the input item set is $M_{i,j}$ as OPT.

Since O_{max} is the maximal number of coins which are the outs of a same transaction, $dive_C \geq \sum_{m_t \in C} \frac{dive_t}{O_{max}} = \sum_{m_t \in C} dive^\#_t$ and $\sum_{m_t \in OPT} dive^\#_t \geq \frac{dive^*}{O_{max}}$. Therefore, by [21], $dive_C \geq \sum_{m_t \in C_{i,j}} dive^\#_t \geq (1-\delta) \cdot \sum_{m_t \in OPT} dive^\#_t \geq (1-\delta) \cdot \frac{dive^*}{O_{max}}$. Thus, $\frac{dive_C}{dive^*} \geq \frac{1-\delta}{O_{max}}$. □

Then, based on Theorem 5.1, we give the proof of the approximate ratio of the Progressive Algorithm.

Theorem 5.2. *The approximate ratio of the Progressive Algorithm is $\min\{\frac{1-\delta}{O_{max}}, \frac{O_{min}}{O_{max}} \cdot \frac{B-S^+}{B}\}$, where S^+ is the maximal size of a module in $\mathbb{M}$.*

Proof. Suppose the ring signature with maximal diversity is r_{opt} and the ring signature generated by the Progressive Algorithm is r_p. Suppose m_h is the module whose pr^i_{max} is the highest in the r_{opt}, and m_s is the module whose pr^i_{min} is the smallest in the r_{opt}. Denote the diversity of r_{opt} as $dive_{opt}$, the diversity of r_p as $dive_p$, the diversity of $\omega_{h,s}$ as $dive^\omega_{h,s}$, and the diversity of $\psi_{h,s}$ as $dive^\psi_{h,s}$. Besides, denote the number of historical transactions which outputs the coins in $C_{h,s}$ and does not outputs the coins in $\omega_{h,s}$ as $dive_{h,s}$. Thus, $dive_p \geq dive^\psi_{h,s}$.

When $|C_{h,s}| + |\omega_{i,j}| \leq B$, $dive^\psi_{h,s} = dive^\omega_{h,s} + dive^c_{h,s}$. Suppose when $i = h$ and $j = s$, the optimal result of δ-KP is OPT' and its diversity is denoted as $dive'_{opt}$. Thus, $dive_{opt} \leq dive'_{opt} + dive^\omega_{h,s}$. As proved in Theorem 5.1, $dive^c_{h,s} \geq dive'_{opt} \cdot \frac{1-\delta}{O^{i,j}_{max}}$. Thus, $\frac{dive_p}{dive_{opt}} \geq \frac{dive^\omega_{h,s} + dive^c_{h,s}}{dive^\omega_{h,s} + dive'_{opt}} \geq \frac{dive^\omega_{h,s} + dive'_{opt} \cdot \frac{1-\delta}{O^{i,j}_{max}}}{dive^\omega_{h,s} + dive'_{opt}} \geq \frac{1-\delta}{O^{i,j}_{max}} \geq \frac{1-\delta}{O_{max}}$.

When $|C_{h,s}| + |\omega_{i,j}| > B$, $B \geq |\psi_{h,s}| \geq B - S^+$. Since $dive^\psi_{h,a} \geq \frac{|\psi_{h,s}|}{O_{max}}$, $dive_p \geq dive^\psi_{h,s} \geq \frac{B-S^+}{O_{max}}$. Since $dive_{opt} \leq \frac{|r_{opt}|}{O_{min}}$ and $|r_{opt}| \leq B$, $dive_{opt} \leq \frac{B}{O_{min}}$. Thus, $\frac{dive_{ts}}{dive_{opt}} \geq \frac{O_{min}}{O_{max}} \cdot \frac{B-S^+}{B}$. □

Next, we give the proof of the time complexity of the Progressive Algorithm.

Theorem 5.3. *The time complexity of the Progressive Algorithm is $\mathcal{O}(\frac{n^5}{\delta})$, where $n = |\mathbb{M}| = |\mathbb{SRS}| + |\mathbb{F}|$ and δ is the precision parameter of the δ-KP Algorithm.*

Proof. There are $\mathcal{O}(n^2)$ (i,j) pairs. For each (i,j) pair, the δ-KP Algorithm cost $\mathcal{O}(\frac{n^3}{\delta})$. Thus, the total time complexity is $\mathcal{O}(\frac{n^5}{\delta})$. □

6. GAME THEORETIC APPROACH

Although the Progressive Algorithm can generate a new RS with a theoretic guaranteed approximate ratio, its time complexity is high. To solve the CIA-MS-DS problem more efficiently, in this section, we develop an approach based on the game theory. Specifically, we model the CIA-MS-DS problem as a strategic game, where each super RS and each fresh coin corresponds to a player: its goal is to find a strategy that maximizes its own utility. However, to develop such an approach, there are two challenges need to be solved: 1) design utility functions of players to let the sum of their objective is the same as the objective of the CIA-MS-DS problem; and 2) prove the algorithm can achieve a Nash equilibrium with guaranteed quality within polynomial time. We solved these two challenges in the following subsections.

6.1 Game Theoretic Algorithm

In this subsection, we solve the first challenge and design utility functions of players.

In strategic games [25], players compete with each other to optimize their individual objective functions. Under this framework, each player always tries to choose a strategy that minimizes her/his own cost without taking the effect of her/his choice on the objectives of other players into consideration. The input of the framework is a strategic game, which can be formally represented by a tuple $\langle P, \{S_p\}_{p \in P}, \{C_p : \times_{p \in P} S_p\}_{p \in P} \rangle$ where P is a set of players and S_p represents all the possible strategies that a player p can take during the game to optimize her/his function C_p. The optimization of C_p depends on the own strategy of p, as well as the strategies of other players. In [26], Nash points out that a strategic game has a pure Nash equilibrium, if there exists a specific choice of strategies $s_p \in S_p$ such that the following condition is true for all $p_i \in P$: $C_i(s_1, \cdots, s_i, \cdots, s_{|P|}) \leq C_i(s_1, \cdots, s'_i, \cdots, s_P), \forall s'_i \in S_{p_i}$. Thus, no player has the incentive to deviate from her/his current strategy. To express the objective functions of all players, [27] proposes a single function $\Phi : \times_{p \in P} S_p$, called the potential function

Algorithm 3: Game Theoretic Algorithm

Input: A spent coin c_τ, a budget B, a super RS set $\mathbb{SRS}$, and a fresh coin set $\mathbb{F}$

Output: An eligible RS r_τ

1 $\mathbb{M} = \mathbb{SRS} \cup \mathbb{F}$, $r_\tau = m_\tau$;
2 **for** $i = 1$ *to* $|M|$ **do**
3 $P_i = M \backslash \bigcup_{m_k \in M \wedge Pr^k_{max} > Pr^i_{max}} m_k$;
4 **for** $j = 1$ *to* $|M|$ **do**
5 **if** $m_j \in P_i$, $Pr^i_{max} \geq Pr^\tau_{max}$ *and* $Pr^j_{min} \leq Pr^\tau_{min}$ **then**
6 $\omega_{i,j} = m_i \cup m_j \cup m_\tau$;
7 $P_{i,j} = P_i \backslash (\omega_{i,j} \cup \bigcup_{m_k \in P_i \wedge Pr^k_{min} < Pr^j_{min}} m_k)$;
8 initialize the strategy of each player;
9 **repeat**
10 **foreach** *play* $p \in P_{i,j}$ **do**
11 $s_i = s^0$;
12 **if** *the utility of* s^1 *is higher* **then**
13 $s_i = s^1$;
14 **until** *reaching an Nash equilibrium*
15 get $\psi_{i,j}$ by strategies of players, combining $\omega_{i,j}$;
16 **return** *the* $\psi_{i,j}$ *with maximal diversity as* r_τ

in potential games, which constitutes a special class of strategic games. Let $\overline{s_i}$ denote the set of strategies followed by all players except p_i (i.e., $\overline{s_i} = \{s_1, \cdots, s_{i-1}, s_{i+1}, \cdots, s_{|P|}\}$). A potential game is exact if there exists a potential function Φ, such that for all s_i and all possible combinations of $\overline{s_i}$ $C_i(s_i, \overline{s_i}) - C_i(s'_i, \overline{s_i}) = \Phi(s_i, \overline{s_i}) - \Phi(s'_i, \overline{s_i})$. In [27], it is proved that for potential games, the framework always converges to a pure Nash equilibrium.

We model our problem as a game. Each player has two strategies, s^1 and s^0, which are being selected and not being selected in the new RS, respectively. Given s_i and $\overline{s_i}$, if r_τ is eligible, the utility of p_i is $U_i(s_i, \overline{s_i}) = \frac{|\{t_i | c_i \in r_\tau\}|}{|P|}$, otherwise $U_i(s_i, \overline{s_i}) = 0$. Thus, the objective function of the CIA-MS-DS problem is equal to the summation of the utility of each individual player. The goal of each player is to find the strategy that maximizes its own utility. This decomposition of the objective of the CIA-MS-DS problem into the summation of individual utility functions provides a natural motivation for modeling the CIA-MS-DS problem as a game. Further, we define the potential function as

$$\Phi(S) = \begin{cases} \frac{|\{t_i | c_i \in r_\tau\}|}{|P|}, & r_\tau \text{ is eligible} \\ 0 & \text{otherwise} \end{cases}$$

Algorithm 3 shows the pseudo-code of our Game Theoretic Algorithm. Algorithm 2 shows the pseudo-code of our Progressive Algorithm. By the SRS constraint, r_τ is composed of some super RS in $\mathbb{SRS}$ or some fresh coins in $\mathbb{F}$. In other words, each super RS and fresh coin is a candidate module of the new RS r_τ. Thus, we get the module set $\mathbb{M}$ by combing $\mathbb{SRS}$ and $\mathbb{F}$ (line 1). Besides, by the SRS constraint, r_τ has to contain the module m_τ which contains c_τ (line 1). Then we enumerate all Pr_{max}-Pr_{min} pairs (line 2 -15), where m_i is the module in the new RS whose Pr^i_{max} is the maximum and m_j is the module in the new RS whose Pr^i_{min} is the minimum. Thus, for the given Pr_{max}-Pr_{min} pair, $\omega_{i,j} = m_i \cup m_j \cup m_\tau$ must be contained in the new RS (line 6). Besides, for each Pr_{max}-Pr_{min} pair, the player set is $P_{i,j} = M \backslash (\bigcup_{Pr^k_{max} > Pr^i_{max}} m_k \cup \omega_{i,j} \cup \bigcup_{Pr^k_{min} < Pr^j_{min}} m_k)$ (line 7). For each player, we randomly assign a strategy (line 8). Next, the algorithm starts the best-response procedure (line 9-14). In each iteration, for each player p, it selects the strategy with the highest utility, when the two utilities are the same, it selects s^0 (line 10-13). When reaching the Nash equilibrium, according to strategies of players, we get a candidate RS $\psi_{i,j}$ (line 15). Since $r_\tau = m_\tau$ satisfies the required constraints and when r_τ is eligible, the utility of each player is higher, there is at least one eligible RS $\psi_{i,j}$.

6.2 Theoretic Analyses

In this subsection, we solve the second challenge and prove the algorithm can achieve a Nash equilibrium with guaranteed quality within polynomial time.

We first prove that for each Pr_{max}-Pr_{min} pair, our game is an exact potential game.

Theorem 6.1. *For each Pr_{max}-Pr_{min} pair, the game in the best-response procedure is an exact potential game.*

Proof. We denote a Pr_{max}-Pr_{min} pair by (i, j), where $Pr_{max} = Pr^i_{max}$ and $Pr_{min} = Pr^j_{min}$. For any (i, j) pair, we first proving $U_k(s_k, \overline{s_k}) - U_k(s'_k, \overline{s_k}) = \Phi(s_k, \overline{s_k}) - \Phi(s'_k, \overline{s_k})$.

Suppose $r^{i,j}_{\tau,1}$ is the RS which is generated by the strategies $\overline{s_k}$ and s_k, combining $\omega_{i,j}$. Suppose $r^{i,j}_{\tau,2}$ is the RS which is generated by the strategies $\overline{s_k}$ and s'_k, combining $\omega_{i,j}$.

When $r^{i,j}_{\tau,1}$ is eligible and $r^{i,j}_{\tau,2}$ is not eligible, $U_k(s_k, \overline{s_k}) = \Phi(s_k, \overline{s_k}) = \frac{|\{t_h | c_h \in r^{i,j}_{\tau,1}\}|}{|P_{i,j}|}$ and $U_k(s'_k, \overline{s_k}) = \Phi(s'_k, \overline{s_k}) = 0$. Thus, $U_k(s_k, \overline{s_k}) - U_k(s'_k, \overline{s_k}) = \frac{|\{t_h | c_h \in r^{i,j}_{\tau,1}\}|}{|P_{i,j}|} - 0 = \Phi(s_k, \overline{s_k}) - \Phi(s'_k, \overline{s_k})$.

When $r^{i,j}_{\tau,1}$ is eligible and $r^{i,j}_{\tau,2}$ is eligible, $U_k(s_k, \overline{s_k}) = \Phi(s_k, \overline{s_k}) = \frac{|\{t_h | c_h \in r^{i,j}_{\tau,1}\}|}{|P_{i,j}|}$ and $U_k(s'_k, \overline{s_k}) = \Phi(s'_k, \overline{s_k}) = \frac{|\{t_h | c_h \in r^{i,j}_{\tau,2}\}|}{|P_{i,j}|}$. Thus, $U_k(s_k, \overline{s_k}) - U_k(s'_k, \overline{s_k}) = \frac{|\{t_h | c_h \in r^{i,j}_{\tau,1}\}|}{|P_{i,j}|} - \frac{|\{t_h | c_h \in r^{i,j}_{\tau,2}\}|}{|P_{i,j}|} = \Phi(s_k, \overline{s_k}) - \Phi(s'_k, \overline{s_k})$.

When $r^{i,j}_{\tau,1}$ and $r^{i,j}_{\tau,1}$ are not eligible, $U_k(s_k, \overline{s_k}) = \Phi(s_k, \overline{s_k}) = 0$ and $U_k(s'_k, \overline{s_k}) = \Phi(s'_k, \overline{s_k}) = 0$. Thus, $U_k(s_k, \overline{s_k}) - U_k(s'_k, \overline{s_k}) = 0 - 0 = \Phi(s_k, \overline{s_k}) - \Phi(s'_k, \overline{s_k})$.

When $r^{i,j}_{\tau,1}$ is not eligible and $r^{i,j}_{\tau,1}$ is eligible, $U_k(s_k, \overline{s_k}) = \Phi(s_k, \overline{s_k}) = 0$ and $U_k(s'_k, \overline{s_k}) = \Phi(s'_k, \overline{s_k}) = \frac{|\{t_h | c_h \in r^{i,j}_{\tau,1}\}|}{|P_{i,j}|}$. Thus, $U_k(s_k, \overline{s_k}) - U_k(s'_k, \overline{s_k}) = 0 - \frac{|\{t_h | c_h \in r^{i,j}_{\tau,1}\}|}{|P_{i,j}|} = \Phi(s_k, \overline{s_k}) - \Phi(s'_k, \overline{s_k})$.

Then, for any (i, j) pair, by [27], since $U_k(s_k, \overline{s_k}) - U_k(s'_k, \overline{s_k}) = \Phi(s_k, \overline{s_k}) - \Phi(s'_k, \overline{s_k})$, the game in the best-response procedure is an exact potential game. □

Since the game in each best-response procedure is an exact potential game, and the set of strategic configurations S is finite, by [27], a Nash equilibrium can be reached after players changing their strategies a finite number of times. For simplicity, we prove the upper bound for the number of rounds required to reach the convergence of Game Theoretic Algorithm by a scaled version of the problem where the objective function takes integer values. We assume an equivalent game with potential function $\Phi_\mathbb{Z}(S) = d \cdot \Phi(S)$ such that $\Phi_\mathbb{Z}(S) \in \mathbb{Z}, \forall S$, whcih does not scale with the size of the problem. Then, we can prove the following lemma.

Lemma 6.1. *The number of rounds required by each best-response procedure to converge to an equilibrium is $\mathcal{O}(d \cdot n)$, where $n = |\mathbb{M}| = |\mathbb{SRS}| + |\mathbb{F}|$.*

Proof. The scaled version of the Game Algorithm with the potential function $\Phi_\mathbb{Z}(S) = d \cdot \Phi(S)$ will converge to a Nash equilibrium

Table 1: Experimental Settings (Real).

Parameters	Values
the budget B	40, 60, **80**, 100, 120
the CI level ϵ	1.3, 1.4, **1.5**, 1.6, 1.7
the range of the degree of each module $[d^-, d^+]$	[1,9], [1,8], **[1,7]**, [1,6], [1,5]
the range of Pr_{max} of each module, $[PM^-, PM^+]$	[0.1, 0.5], [0.1, 0.55], **[0.1, 0.6]**, [0.1, 0.65], [0.1, 0.7]

in the same number of rounds as the Game algorithm. Since the change of $\Phi_{\mathbb{Z}}(S)$ is at least 1, and $0 \leq \Phi_{\mathbb{Z}}(S) \leq n$, the number of rounds is at most $\frac{d \cdot n - 0}{1} = d \cdot n$. □

Then, we give the proof of the time complexity of the Game Theoretic Algorithm as follows.

Lemma 6.2. *The time complexity of the Game Theoretic Algorithm is $\mathcal{O}(d \cdot n^4)$, where $n = |\mathbb{M}| = |\mathbb{SRS}| + |\mathbb{F}|$.*

Proof. There are $\mathcal{O}(n^2)$ Pr_{max}-Pr_{min} pairs and for each pair, the best-response procedure requires $\mathcal{O}(d \cdot n)$ rounds iteration. Furthermore, in each iteration of the best-response procedure (line 9 - 14), there are $\mathcal{O}(n)$ players. Thus, the time complexity of the Game Theoretic Algorithm is $\mathcal{O}(d \cdot n^4)$. □

After proving Game Theoretic Algorithm can converge within polynomial time, we discuss how good the resulting solution is. Usually, researchers use **social optimum (OPT)**, **price of stability(PoS)**, and **price of anarchy(PoA)** to evaluate the quality of an equilibrium. The OPT is the solution that yields the optimal values to all the objective functions, so that their total utility is maximum. The PoS of a game is the ratio between the best value among its equilibriums and the OPT. The PoA of a game is the ratio between the worst value among its equilibriums and the OPT.

Theorem 6.2. *The PoS is bounded by $\frac{1}{n}$ and the PoA is bounded by $\frac{O_{min} \cdot dive_\tau}{B}$, where $n = |\mathbb{M}| = |\mathbb{SRS}| + |\mathbb{F}|$, O_{min} is the minimal number of coins which are outputted by the same transaction, and $dive_\tau$ is the diversity of the module which contains c_τ.*

Proof. Let $S_{i,j}$ be the set of strategies of players in $P_{i,j}$ and $U(S_{i,j})$ = $\sum_{k=1}^{|P_{i,j}|} U_k(s_k, \overline{s_k})$. Thus, $\frac{U(S_{i,j})}{|P_{i,j}|} \leq \Phi(S_{i,j}) \leq U(S_{i.j})$. Let $S^*_{i,j}$ be the globally optimal set of strategies of players in $P_{i,j}$ that maximizes $U(S_{i,j})$ and let $OPT_{i,j}$ = $U(S^*_{i,j})$. Let $S^{\#}_{i,j}$ be the set of strategies of players in $P_{i,j}$ that yields the maximum of $\Phi(S_{i,j})$, i.e., the best Nash equilibrium of a game. Thus, $U(S^{\#}_{i,j}) \geq \Phi(S^{\#}_{i,j})$ $\geq \Phi(S^*_{i,j}) \geq \frac{U(S^*_{i,j})}{|P_{i,j}|} = \frac{OPT_{i,j}}{|P_{i,j}|}$. Thus, $PoS \geq \frac{U(S^{\#}_{i,j}) + dive^{\omega}_{i,j}}{OPT_{i,j} + dive^{\omega}_{i,j}}$ $\geq \frac{\frac{1}{|P_{i,j}|} \cdot OPT_{i,j} + dive^{\omega}_{i,j}}{OPT_{i,j} + dive^{\omega}_{i,j}} \geq \frac{1}{|P_{i,j}|} \geq \frac{1}{n}$. Since $m_\tau \subseteq r_\tau$ and $OPT \leq \frac{B}{O_{min}}$, $PoA \geq \frac{dive_\tau}{OPT} \geq \frac{O_{min} \cdot dive_\tau}{B}$. □

7. EXPERIMENTAL STUDY

7.1 Experiment Configuration

We use both real and synthetic data sets to test our proposed approaches. Specifically, for real data sets, we retrieve the coins in the blocks between 2028242 and 2028273 from the Monero System. The time gap between the block 2028242 and the block 2028273 is one hour. There are 285 transactions and 633 coins. Figure 5 shows the distribution of the number of coins in a transaction. Most transactions only output two coins and there are 4 transactions which output 16 coins. Since in Monero System, most RS's size is 11, we retrieve 627 coins among them and generate 57 super RSs. For each super RSs, it randomly selects 11 coins as its coin set and its transaction set contains the transactions outputting the selected 11 coins. We uniformly generate the degree of each super SR within the range $[d^-, d^+]$. We generate the Pr_{max} of each super RS within the range $[PM^-, PM^+]$ following the uniform distribution. Since each super RS satisfies the ϵ-CI, we generate the Pr_{min} of each super RS by $\frac{e^\epsilon \cdot (1 - Pr_{max})}{d_i \cdot Pr_{max} + 1} = \frac{1 - Pr_{min}}{d_i \cdot Pr_{min} + 1}$. Besides, we vary the budget from 40 to 120 and the CI level ϵ from 1.3 to 1.7.

To examine the effects of the number of modules, the number of historical transactions, and each module's size, we generate the synthetic dataset and run the experiments on it. For synthetic data sets, we generate n modules. For each module, we randomly generate its degree, size, mixin set, Pr_{max} and Pr_{min}. Specifically, we uniformly generate the degree of each module within the range $[d^-, d^+]$. We uniformly generate the size of each within the range of $[s^-, s^+]$. We generate the Pr_{max} of each each module within the range $[PM^-, PM^+]$ following the uniform distribution. Since each module satisfies the ϵ-CI, we generate the Pr_{min} of each module by $\frac{e^\epsilon \cdot (1 - Pr_{max})}{d_i \cdot Pr_{max} + 1} = \frac{1 - Pr_{min}}{d_i \cdot Pr_{min} + 1}$. Besides, for each coin, among o historical transactions, we randomly select a historical transaction outputting it. We vary the budget from 110 to 190 and the CI level ϵ from 1.6 to 2. And we vary the n from 50 to 90. Since the average size of each module is 16, 50 modules cover more than 800 coins, which is large enough for real-world applications. In Monero [15], as shown in the real data sets, the average number of coins in each hour is less than 800.

We conduct experiments on both the real data sets and the synthetic data sets to evaluate the effectiveness and efficiency of our two approaches, the Game Theoretic Algorithm and the Progressive Algorithm, in terms of the new RS's diversity and the running time. As proved in Theorem 3.7, the CIA-MS-DS problem is NP-hard, thus, it is infeasible to calculate the optimal result as the ground truth in large scale datasets. Alternatively, we compare our approaches with two baseline methods, the Greedy Algorithm and the Random Algorithm. The Greedy Algorithm initializes r_τ as m_τ and then greedily add the candidate module which can bring the largest increase of the diversity of the temporary r_τ among the modules which would not make the temporary r_τ ineligible to the temporary r_τ. The Random Algorithm initializes r_τ as m_τ and then randomly add a candidate module among the modules which would not make the temporary r_τ ineligible to r_τ.

Table 1 and Table 2 introduce our experiment settings on real data sets and synthetic data sets respectively, where the default values of parameters are in bold font. In each set of experiments, we vary one parameter, while setting other parameters to their default values. For each experiment, we sample 50 problem instances and run the algorithms. We report the average value of the running time and the RS's diversity. All our experiments were run on an Intel CPU @2.2 GHz with 16GM RAM in Java.

7.2 Results on Real Data Sets

Effect of the Budget, B. Figure 6 illustrates the experimental result on different budgets, B, from 40 to 120. In Figure 6(a), when the budget gets larger from 40 to 80, the diversities of the new RSs that generated by the four approaches increase; then, they almost keep stable. The reason is that at the beginning, with the increase of B, the new RS can contain more mixins. Nevertheless, the new RS is also constrained by the ϵ-CIK-RS constraint. When the budget

Table 2: Experimental Settings (Synthetic).

Parameters	Values
the number, n, of modules	**50**, 60, 70, 80, 90
the number, o, of transactions	50, 60, **70**, 80, 90
the budget B	110, 130, **150**, 170, 190
the CI level ϵ	1.6, 1.7, **1.8**, 1.9, 2
the range of the degree of each module $[d^-, d^+]$	**[1,9]**, [1,8], [1,7], [1,6], [1,5]
the range of the size of each module $[s^-, s^+]$	[11,15], **[14,18]**, [17,21] [20,24], [23,27]
the range of the Pr_{max} of each module, $[PM^-, PM^+]$	[0.1, 0.2], [0.1, 0.35], **[0.1, 0.5]**, [0.1, 0.65], [0.1, 0.8]

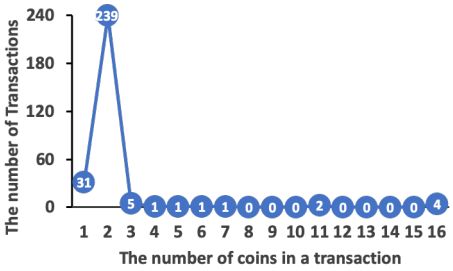
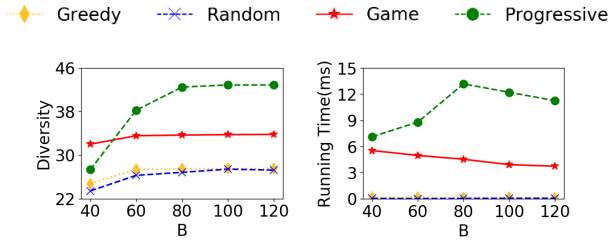


Figure 5: The Distribution of the Number of Coins in a Transaction

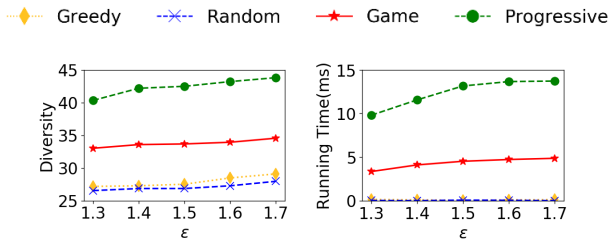


(a) Diversity (b) Running Time

Figure 6: Effect of the Budget (Real)

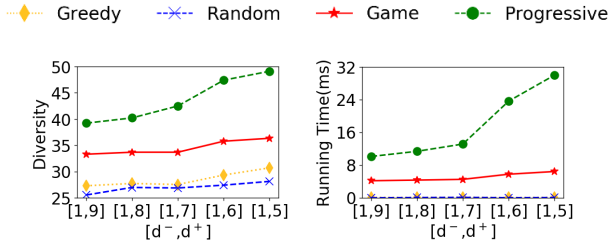
is large enough, the diversity of the new RS is limited by the ϵ-CIK-RS constraint. And particularly, when the budget is very low, the diversity of the new RS which is generated by the Progressive Algorithm is lower than that of the new RS which is generated by the Game Theoretic Algorithm. The reason is when the budget is low, the Progressive Algorithm needs to adjust the selection from the δ-KP algorithm, $C_{i,j}$, for the budget constraint (line 11-15). When the budget is very low, the diversity of the new RS which is adjusted by the greedy procedure in the Progressive Algorithm is not as good as that of the Game Algorithm. The reason is that, the greedy procedure is easier to fall into the local optimal trap while the Game Theoretic Algorithm can release this by the games between players.

As shown in Figure 6(b), when the budget increases, the running time of two baseline approaches also increases. Because, the increase of B allows the new RS to contain more mixins, which thus leads to the higher complexity of the CIA-MS-DS problem and the increase of the running time. However, the running time of the Game Theoretic Algorithm decrease. The reason is that, when the budget constraint is relaxed, the game in the Game Theoretic Algorithm is easier to reach the Nash equilibrium. The running time of the Progressive Algorithm increases at the beginning. The reason is that, when the budget increases but still is a strict constraint, the Progressive Algorithm spends more time on the greedy procedure.



(a) Diversity (b) Running Time

Figure 7: Effect of the CI Level (Real)



(a) Diversity (b) Running Time

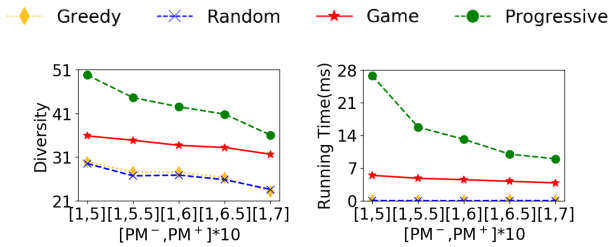
Figure 8: Effect of the Range of the Degree of each Super RS (Real)

But later, when the budget is large enough, the running time of the Progressive Algorithm decreases. The reason is that, when the budget is large enough, the budget constraint is relaxed and there are more $C_{i,j}$ whose size are smaller than the budget and the Progressive Algorithm do not need to run the greedy procedure for these candidate RSs.

Effect of the CI Level, ϵ. Figure 7 illustrates the experimental result on different CI levels, ϵ, from 1.3 to 1.7. In Figure 7(a), when the ϵ increases, the diversities of the new RSs that generated by the four approaches also increase. The reason is that the increase of ϵ makes the ϵ-CIK-RS constraint more relaxed and the new RS has more valid modules. As shown in Figure 7(b), when the ϵ increases, the running time of four approaches also increases. Because, the increase of ϵ let the new RS has more valid modules, which thus leads to the higher complexity of the CIA-MS-DS problem and the increase of the running time.

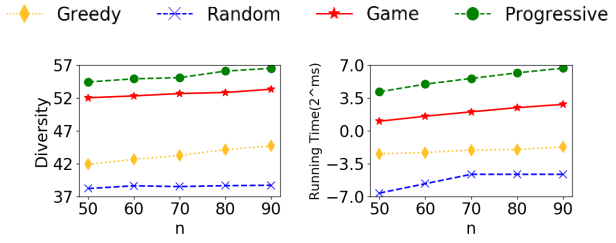
Effect of the Range of the Degree of each Super RS, $[d^-, d^+]$. Figure 8 illustrates the experimental results on different ranges, $[d^-, d^+]$, of the degree of each super RS, from [1,9] to [1,5]. In Figure 8(a), the diversities of the new RSs that generated by our four approaches increase, when the average value of degrees of modules decreases. The reason is that, when the average value of degrees of modules is higher, it is more difficult to satisfy the ϵ-CIK-RS constraint. In other words, the new RS has fewer valid modules. In Figure 8(b), the running time of our four approaches increases when the average value of degrees of modules decreases. The reason is that, when the average value of degrees of modules, the new RS has fewer valid modules, which thus leads to lower complexity of the CIA-MS-DS problem and the decrease of the running time.

Effect of the Range of the Pr_{max} of each Super RS, $[PM^-, PM^+]$. Figure 9 illustrates the experimental results on different ranges, $[PM^-, PM^+]$, of the Pr_{max} of each module, from [0.1,0.5] to [0.1,0.7]. In Figure 9(a), the diversities of the new RSs that generated by our four approaches decrease, when the range is wider. The reason is that, when the range of the Pr_{max} of each module is wider, it is more difficult to satisfy the ϵ-CIK-RS constraint. In other words, the new RS has fewer valid modules. In Figure 9(b), the running time of our four approaches decreases when the difference of the Pr_{max} of modules increases. The reason is that, when



(a) Diversity (b) Running Time

Figure 9: Effect of the Range of the Pr_{max} of each Super RS (Real)



(a) Diversity (b) Running Time

Figure 10: Effect of the Number of Modules (Synthetic)

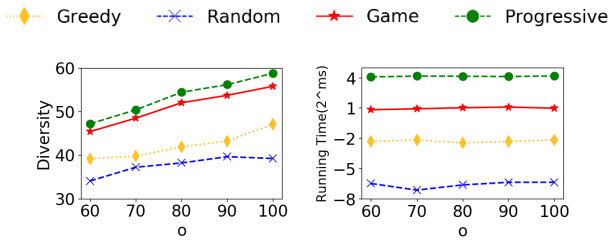
the range of the Pr_{max} of modules is narrower, the new RS has less valid modules, which thus leads to higher complexity of the CIA-MS-DS problem and the increase of the running time. Specifically, compared with the Game Theoretic Algorithm, the Progressive Algorithm is more sensitive about the change of the range. The reason is that, when the range is wider, the cardinality of $M_{i,j}$ is smaller and the running time of the δ-KP Algorithm is smaller.

7.3 Results on Synthetic Data Sets

To examine the effects of the number of modules, the number of historical transactions, and each module's size, we generate the synthetic dataset and run the experiments on it. We also test the effects of the budget, the CI level, the range of the degree of each module, and the range of the Pr_{max} of each module on the synthetic data sets. Due to the space limitation, please refer to Appendix F of our technical report [24] for details.

Effect of the Number, n of Modules. Figure 10 illustrates the experimental result on a different number, n, of modules from 50 to 90. In Figure 10(a), when the number of modules increases, the diversities of the new RSs that generated by the four approaches also increase. The reason is that the increase of n let the new RS has more valid modules. Specifically, our two approximate algorithms achieve much better results than the two baseline algorithms. The RS which is generated by the Progressive Algorithm has the largest diversity. As shown in Figure 10(b), when the number of modules increases, the running time of four approaches increases. Because when the number of modules increases, the new RS has more valid modules, which thus lead to the increase of the running time. Specifically, the Progressive Algorithm's running time is the highest while the running time of the Game Theoretic Algorithm is much lower.

Effect of the Number, o of Historical Transactions. Figure 11 illustrates the experimental result on a different number, o, of historical transactions from 60 to 100. In Figure 11(a), when the number of historical transactions increases, the diversities of new RSs that generated by the four approaches also increase. The reason is that the increase of o decreases the overlap between historical transaction set of modules, where the historical transaction set of a module is the set of historical transaction outputting the coins in the



(a) Diversity (b) Running Time

Figure 11: Effect of the Number of Historical Transactions (Synthetic)

module. As shown in Figure 11(b), when the number of historical transactions increases, the running time of four approaches almost keeps stable. Because the complexity of the CIA-MS-DS problem and the time complexities of four approaches are all not related to the number of historical transactions.

Effect of the Range of the Size of each Module, $[s^-, s^+]$. Figure 12 illustrates the experimental results on different ranges, $[s^-, s^+]$, of the size of each module, from [11,15] to [23,27]. In Figure 12(a), the diversities of the new RSs that generated by our four approaches decrease, when the size of each module increases. The reason is that, when the average value of sizes of modules is higher, it is more difficult to satisfy the budget constraint. In other words, the new RS has fewer valid modules. In Figure 12(b), the running time of our Progressive Algorithm and two baseline approaches decreases when the average value of sizes of modules increases. The reason is that, when the average value of sizes of modules is higher, the new RS has fewer valid modules, which thus leads to lower complexity of the CIA-MS-DS problem and the decrease of the running time. However, the running time of our Game Theoretic Algorithm increases when the average value of sizes of modules increases. The reason is that, when the average value of sizes of modules is higher, the Game Theoretic Algorithm needs to do more games to reach the Nash equilibrium.

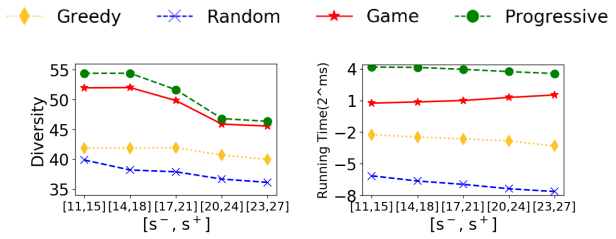
We finally summarize our findings as following:

- Our two approximate algorithms can achieve results with higher diversity compared with that of two baselines.
- The Progressive Algorithm outputs the RS with the highest diversity while it costs much time. But its speed is still acceptable for the current public blockchain systems, like Monero [15].
- The Game Theoretic Algorithm outputs the RS with high diversity quickly. Compared with the Progressive Algorithm, it is more suitable for consortium blockchain systems whose TPSs are high.

8. RELATED WORK

Blockchain technologies are gaining massive momentum in recent years, largely due to its immutability and transparency. Many applications for security trading and settlement [28], asset and finance management [29] [30], banking and insurance [31] are evaluated. However, the transparency character also brings the privacy problem. In many practice applications, users do not want to share all their information with other participants. To solve the privacy problem, some researchers have proposed some privacy-preserved blockchain system.

These works can be classified into two categories. The works in the first category focus on developing mixing protocols [32, 33, 34, 35]. In these protocols, anonymous service providers use mixing protocols to confuse the trails of transactions. The client's funds are divided into smaller parts which are mixed randomly with similar random parts of other clients. This helps to break links between the



(a) Diversity (b) Running Time

Figure 12: Effect of the Range of the Size of each Module (Synthetic)

users and the transactions they purchased while these methods rely on mixers' trustiness and the mixers know the transactions' privacy. These methods weaken the blockchain system's decentration.

The works in the second category focus on developing advanced encryption methods. In the Monero blockchain system [14], the researchers build the privacy-preserved blockchain system based on the Ring Confidential Transaction (RingCT) [12] [13] [14]. In the first version of RingCT [12] (RingCT 1.0), the researchers adapted a RS scheme [38] to protect the transaction's sender's identity. In the second version of RingCT [13], the researchers put forward a new efficient RingCT protocol (RingCT 2.0), which is built upon the well-known Pedersen commitment [40] and saves almost half RS's size compared with former version RingCT. In [14], the researchers put forward the newest RingCT protocol based on Bulletproof [42], which decreases the RS's size from $\mathcal{O}(n)$ to $\mathcal{O}(\log n)$.

While these cryptographic techniques are used to some extent achieve confidentiality, the considerable overhead of such techniques makes them impractical [43]. Besides, these techniques assume the adversary has no extra information. However, since all data in the blockchain system is accessed, adversaries can attack a user's privacy by analyzing the traffic flow on the blockchain system. Our work considers the traffic flow's impact and proposes methods to strengthen RSs' privacy-preserving effect by selecting a set of desirable mixins.

9. CONCLUSION

In this paper, we formulate a differential privacy definition, namely ϵ-coin-indistinguishability, in blockchain scenarios. We show that, if a RS satisfies the ϵ-coin-indistinguishability, it is resistant to the "Chain-Reaction" analysis. And in this paper, we formulate the coin-indistinguishability-aware mixin selection problem with the disjoint-superset constraint (CIA-MS-DS), which aims to find a set of mixins which satisfy the ϵ-coin-indistinguishability constraint, as well as the budget constraint, and has the maximal diversity. We formally prove that the CIA-MS-DS has some significant properties which can help to simplify the problem, while the CIA-MS-DS problem still is an NP-hard problem. To efficiently and effectively solve the CIA-MS-DS problem, we propose a novel framework, CoinMagic, and propose two approximate algorithms, namely the Progressive Algorithm and the Game Algorithm, with theoretical guarantees. When evaluated on the real and synthetic data sets, our approaches achieved clearly better performance than two baseline algorithms.

10. REFERENCES

[1] S. Nakamoto *et al.*, "Bitcoin: A peer-to-peer electronic cash system," 2008.

[2] G. Wood *et al.*, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, no. 2014, pp. 1–32, 2014.

[3] "[online] Thunderchain." `https://www.lianxiangcloud.com`, 2017.

[4] C. Xu, C. Zhang, and J. Xu, "vchain: Enabling verifiable boolean range queries over blockchain databases," in *Proceedings of the 2019 International Conference on Management of Data*, pp. 141–158, ACM, 2019.

[5] C. Zhang, C. Xu, J. Xu, Y. Tang, and B. Choi, "Gemˆ 2-tree: A gas-efficient structure for authenticated range queries in blockchain," in *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 842–853, IEEE, 2019.

[6] K. Korpela, J. Hallikas, and T. Dahlberg, "Digital supply chain transformation toward blockchain integration," in *proceedings of the 50th Hawaii international conference on system sciences*, 2017.

[7] A. Azaria, A. Ekblaw, T. Vieira, and A. Lippman, "Medrec: Using blockchain for medical data access and permission management," in *2016 2nd International Conference on Open and Big Data (OBD)*, pp. 25–30, IEEE, 2016.

[8] Y. Guo and C. Liang, "Blockchain application and outlook in the banking industry," *Financial Innovation*, vol. 2, no. 1, p. 24, 2016.

[9] "[online] Mass vehicle ledger." `https://mvlchain.io/`, 2017.

[10] M. E. Andrés, N. E. Bordenabe, K. Chatzikokolakis, and C. Palamidessi, "Geo-indistinguishability: Differential privacy for location-based systems," *arXiv preprint arXiv:1212.1984*, 2012.

[11] T. Okamoto and K. Ohta, "Universal electronic cash," in *Annual international cryptology conference*, pp. 324–337, Springer, 1991.

[12] N. Van Saberhagen, "Cryptonote v 2.0," 2013.

[13] S.-F. Sun, M. H. Au, J. K. Liu, and T. H. Yuen, "Ringct 2.0: A compact accumulator-based (linkable ring signature) protocol for blockchain cryptocurrency monero," in *European Symposium on Research in Computer Security*, pp. 456–474, Springer, 2017.

[14] T. H. Yuen, S.-f. Sun, J. K. Liu, M. H. Au, M. F. Esgin, Q. Zhang, and D. Gu, "Ringct 3.0 for blockchain confidential transaction: Shorter size and stronger security," 2019.

[15] "[online] Monero." `https://www.getmonero.org/`, 2017.

[16] "[online] Bytecoin." `https://bytecoin.org/`, 2017.

[17] L. Sweeney, "k-anonymity: A model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

[18] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*, pp. 106–115, IEEE, 2007.

[19] M. Möser, K. Soska, E. Heilman, K. Lee, H. Heffan, S. Srivastava, K. Hogan, J. Hennessey, A. Miller, A. Narayanan, *et al.*, "An empirical analysis of traceability in the monero blockchain," *Proceedings on Privacy Enhancing Technologies*, vol. 2018, no. 3, pp. 143–163, 2018.

[20] J. O. M. Chervinski, D. Kreutz, and J. Yu, "Floodxmr: Low-cost transaction flooding attack with monero's bulletproof protocol.," *IACR Cryptology ePrint Archive*, vol. 2019, p. 455, 2019.

[21] V. V. Vazirani, *Approximation algorithms*. Springer Science & Business Media, 2013.

[22] C. Dwork, "Differential privacy," *Encyclopedia of Cryptography and Security*, pp. 338–340, 2011.

[23] A. Hinteregger and B. Haslhofer, "Short paper: An empirical analysis of monero cross-chain traceability," in *International Conference on Financial Cryptography and Data Security*, pp. 150–157, Springer, 2019.

[24] "[online] Technical Report." `https://cspcheng.github.io/pdf/DASC.pdf`.

[25] N. Armenatzoglou, H. Pham, V. Ntranos, D. Papadias, and C. Shahabi, "Real-time multi-criteria social graph partitioning: A game theoretic approach," in *ACM SIGMOD*, pp. 1617–1628, 2015.

[26] J. F. Nash *et al.*, "Equilibrium points in n-person games," *PNAS*, vol. 36, no. 1, pp. 48–49, 1950.

[27] D. Monderer and L. S. Shapley, "Potential games," *Games and economic behavior*, vol. 14, no. 1, pp. 124–143, 1996.

[28] "[online] Ripple." `https://ripple.com`, 2017.

[29] "[online] Blockchain software for asset management." `http://melonport.com`, 2017.

[30] J. Morgan and O. Wyman, "Unlocking economic advantage with blockchain. a guide for asset managers.," *New York: JP Morgan Reports*, 2016.

[31] G. Group *et al.*, "Blockchain: Putting theory into practice," 2016.

[32] G. Maxwell, "Coinjoin: Bitcoin privacy for the real world," in *Post on Bitcoin forum*, 2013.

[33] T. Ruffing, P. Moreno-Sanchez, and A. Kate, "Coinshuffle: Practical decentralized coin mixing for bitcoin," in *European Symposium on Research in Computer Security*, pp. 345–364, Springer, 2014.

[34] T. Ruffing, P. Moreno-Sanchez, and A. Kate, "P2p mixing and unlinkable bitcoin transactions.," in *NDSS*, 2017.

[35] P. Moreno-Sanchez, T. Ruffing, and A. Kate, "Pathshuffle: Credit mixing and anonymous payments for ripple," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 3, pp. 110–129, 2017.

[36] E. B. Sasson, A. Chiesa, C. Garman, M. Green, I. Miers, E. Tromer, and M. Virza, "Zerocash: Decentralized anonymous payments from bitcoin," in *2014 IEEE Symposium on Security and Privacy*, pp. 459–474, IEEE, 2014.

[37] E. Ben-Sasson, A. Chiesa, E. Tromer, and M. Virza, "Succinct non-interactive zero knowledge for a von neumann architecture," in *23rd {USENIX} Security Symposium ({USENIX} Security 14)*, pp. 781–796, 2014.

[38] R. Cramer, I. Damgård, and B. Schoenmakers, "Proofs of partial knowledge and simplified design of witness hiding protocols," in *Annual International Cryptology Conference*, pp. 174–187, Springer, 1994.

[39] W. Diffie and M. Hellman, "New directions in cryptography," *IEEE transactions on Information Theory*, vol. 22, no. 6, pp. 644–654, 1976.

[40] T. P. Pedersen, "Non-interactive and information-theoretic secure verifiable secret sharing," in *Annual international cryptology conference*, pp. 129–140, Springer, 1991.

[41] G. Maxwell and A. Poelstra, "Borromean ring signatures," *Accessed: Jun*, vol. 8, p. 2019, 2015.

[42] B. Bünz, J. Bootle, D. Boneh, A. Poelstra, P. Wuille, and G. Maxwell, "Bulletproofs: Short proofs for confidential transactions and more," in *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 315–334, IEEE, 2018.

[43] E. Androulaki, A. Barger, V. Bortnikov, C. Cachin, K. Christidis, A. De Caro, D. Enyeart, C. Ferris, G. Laventman, Y. Manevich, *et al.*, "Hyperledger fabric: a distributed operating system for permissioned blockchains," in *Proceedings of the Thirteenth EuroSys Conference*, p. 30, ACM, 2018.