As a new privacy protection technology, local differential privacy is the current research hotspot, which is mainly used in the field of statistical database. At present, the research direction of local differential privacy mainly involves disturbance mechanism, single value frequency distribution, multi value frequency distribution and mean value distribution. Random response technology, as one of the disturbance mechanisms, has a simple and intuitive framework and the degree of disturbance can be directly quantified. Therefore, most of the research work of local differential privacy is based on random response technology, including the frequency distribution of discrete data.

In our project, we learn from local differential privacy algorithm PrivKVM and PrivKVM', whose disturbance mechanism is random response mechanism, but the protected value is continuous. Therefore, we set a certain range of discrete values according to the selection in specific social situations or various circumstances when events happen and apply the improved local differential privacy algorithm PrivKVD to realize the privacy protection of discrete data.

The specific situation is as follows: in the internal system of a hospital, there are documents recording disease situations of all kinds of patients, which can be classified as malignant tumor, benign tumor and non-tumor. Based on the humanitarian protection and privacy protection of patients, when processing their data, we first obtain [[0,0], [1,1], [0,0]] through the mapping function mapping() to indicate that the current patient has a benign tumor; then use the algorithm PrivKVD to disturb the dataset, whose result is that external personnel can not know the specific patient's disease situation, but can analyze the proportion of a certain disease situation.

Table II Notations

|  |  |
| --- | --- |
| **Symbol** | **Description** |
| U | the set of users |
| n | the number of users, n=|U| |
| k | the number of discrete values for choosing |
| ui | the i-th user in U |
| Si | the set of KV pairs possessed by ui |
| ⟨kj,vj⟩ | the j-th KV pair in Si |
| fk | the frequency of users' choice |

This project studies the problem of distributed discrete data aggregation in the context of LDP. Without loss of generality, let the universe consist of a set of users U = {u1,u2,...,un} whose value domain is the discrete domain containing a few numbers the situation specifies. The i-th user ui possesses [[0,0],...,[0,0],[1,1],[0,0],...,[0,0]] as key-value (KV) pair Si = ⟨kj,v\*⟩ where 1 ≤ j ≤ k, kj ∈{0,1}, v\* ∈ {0,1}. The main notations are listed in Table II.

An untrusted data collector needs to estimate some statistics of these discrete data from all users. In this project, we focus on one fundamental estimation: frequency estimation.

* + Frequency estimation. The frequency of users' choice, fk, is defined as the portion of users who possess a KV pair whose choice in the list has the index j. Formally for any choice j,

fk =| { ui | ∃⟨kj,v\*⟩ } | / n