LELEC2770 – Practical Sessions

Practical Session 6: Privacy

1. Download the datasets *diabetes1.csv* and *diabetes2.csv*. It corresponds to medical data for 384 persons. The datasets are decomposed as follow:

diabetes1.csv	*diabetes2.csv*
Cholesterol	Height [inches]
Location	Frame
Age [years]	Waist [inches]
Gender	

   (a) Compute the K-anonymity metric for both datasets.

   (b) Apply a protection mechanism on some variables in the datasets in order to get a K higher than 5.

   (c) Evaluate your transformed data with an utility function based on the distortion. Which means the sum of the differences between initial and transformed values. Look again about your protection mechanism such that you maximize your utility.

   (d) Merge the two datasets and compute again the K-anonymity metric.

   (e) What's the advantage of differential privacy and how could it help here?

2. Download the dataset *basket.csv*. It corresponds to purchases in a supermarket for 10 users. Each observation is a set of products resumed as a basket.

   Observations are in rows while variables are in columns. The columns are:

   Milk / Meat / Apple / Bread / Pizza / Beer / Banana / Fish / Sugar / Corn Flakes / ID

   Notice that the last column is for the users IDentification.

(a) From the dataset, compute the Hypothetical Information (HI) between users and observations given by :

$$HI[U;O] = H[U] - \sum_{u_i \in \mathcal{U}} Pr[u_i] \cdot \sum_{o \in \mathcal{O}_i} Pr[o \mid u_i] \cdot \log_2 Pr[u_i \mid o].$$

With $H[\cdot]$ the Shannon Entropy, $\mathcal{U}$ the set of users and $\mathcal{O}_i$ the set of observations from user u_i.

(b) Download the dataset *basket_test.csv* and re-identify the observations with regards to the previous dataset and users.

(c) Compute the success rate of your re-identification with the help of the file *user_test.csv* which corresponds to the true *basket_test* users.

(d) Compute the Perceived Information with *basket.csv* the training set to build the model and *basket_test.csv* the test set.