**This is just a rough brainstorm**

1. Use principle components analysis to create PCs of the primary keys. We would somehow need to make these values all numeric or ordinal in order to take this approach, However, this would essentially create a series of loading where the higher loading represent the variables that had the highest degree of variability and therefore have potentially sparse identifiers which would better allow for identification of the tuple. We could then take these attributes and use them to create clusters
2. Cluster the attributes and give the user 2 tables in each data release: 1 with the cluster information, 2: table containing the remaining features of the dataset with an associated cluster and no aggregation. The user could then create a Cartesian join to the properties if more information was needed:
   1. The key to this approach is that the user would have to have a good amount of domain expertise

Very generalized example:

Clusters

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Age range | zipcode region | m/f | frequent flier status |
| 1 | 0-18 | Midwest | 1 | 0 |
| 2 | 19-25 | Northeast | 0 | 0 |
| 3 | 26-38 | South Cental | 0 | A |
| 4 | 39-55 | South east | 1 | A-pref |
| 5 | 55+ | Desert mtn | 1 | A |

Data provided to users

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Cluster | Orig | dest | dep\_date | Dep\_time | Business mkt? | # competitiors | Ontime? | NPS score | Itinerary type | total sked travel time | total actual travel time |
| 2 | DAL | MSY | 2/1/2016 | 6:00 | 0 | 2 | Y | 7 | Nonstop | 1:20 | 1:15 |
| 3 | OKC | HOU | 2/2/2016 | 7:00 | 1 | 2 | N | 2 | Nonstop | 1:45 | 2:15 |

These 2 tables could be joined to create a Cartesian product if the user wanted the anonymized attribute groupings to show up in the recordspace

1. Decision tree grouping. Not sure how this would work, but could use the principles of k-anonymity, l-diversity and t-closeness to determine the size of each node. Could be a really simple solution. Use clustering to group, then use decision trees/random forests to classify unknown observations?

Attacks:

Once we have the model in place, we can try some attacks.

Mixture of background knowledge and public data attack

1. I could pull Julie’s flight record in our anonymized dataset and we could calculate the probability that she was on one of 7 DAL-MSY flights on 2/23/2017 and link that to her facebook status and location changes? I’m sure she won’t mind ;)