Users Columns

Identifiers

1. <ignored trivial>

Quasi-identifiers

1. Gender
2. DOB
3. Zipcode

Sensitive data (to be studied)

1. Infection history (By date)
   1. May not be suitable due to background knowledge attack
2. List of vaccination types taken
   1. Could see relations as to which vaccination is more effective
3. Last close contact date with infected user
   1. Could identify the incubation period of the virus
4. Last infected date
   1. May not be suitable due to background knowledge attack
   2. None / date
5. Number of close contacts since infected
   1. May be interesting to identify safety measures effectiveness
6. Number of infections recorded
7. Number of close contacts history

Attacks on K-Anonymity

1. Homogeneity Attack
   1. This occurs when a group of quasi-identifiers has the same sensitive data. For example, for all people living in Jurong region, age within 90+ has cancer, etc.
   2. Also known as lack of diversity in sensitive data
2. Background knowledge attack
   1. <Very hard to counter, to be ignored for now>

Summary of solutions

1. K-anonymity for quasi-identifiers
   1. This will ensure there are always k records for each distinct quasi-identification groups.
2. L-diversity for sensitive data
   1. This works by checking for diversity of sensitive data given a distinct quasi-identification group.
   2. If there is a case of no diversity found in any given distinct group of quasi-identification, generalization of certain quasi-identification field(s) will be performed.
      1. In summary, the k-anonymized quasi-identifiers will be anonymized even further.

References

1. <https://personal.utdallas.edu/~mxk055100/courses/privacy08f_files/ldiversity.pdf>