*Christian Jarani*

*CS 1699: Privacy*

*Project 3: Anonymity*

**W0:** For my dataset, I chose to use the dog license register from Allegheny County that Prof. Garrison provided for us via email. The dataset contains fields for the type of dog license registered, the breed/color/name of the dog, the owner’s zip code, the expiration year of the license, and the time at which the license was validated.

* *Quasi-Identifiers*: Owner Zipcode, Age Range (Individual vs. Senior or Disability)
* *Sensitive Info*: Breed
  + Dog name and spayed/neutered status of dog are also sensitive information, but will be anonymizing data based on breed in particular for the reasons described in **W1**.
* *Source*: <https://catalog.data.gov/dataset/allegheny-county-dog-licenses/resource/c189c9ad-7f20-4686-ae22-b534af7ad092>

**W1:** For the quasi-identifiers listed above, this information exists in a plethora of sources including census/demographic information collected by the state/federal government, health records, education facilities, social groups, etc. Cross-referencing the chosen data set against any of these aggregates could provide insights into the specific person referred to. The sensitive information I will be focusing on is the breed of dog since it is the most likely identifier to be unique on a person-to-person basis that is general enough such that clusters will not leak information about their contents i.e. larger clusters that are similar in size.

Breed listing could revel connections with dog breed communities i.e. kennel clubs and further identification within based off dog name. This also depends on the specificity of the information collected by said communities for each of their members.

**W3**: Compared to the original output, entries in the new table are much more general than in the original table. Firstly, entries are sorted primarily by breed and further by zip code range, and do not include information such as name of dog, age of owner, specific zip code, or license validity date. This not only provides the information necessary for drawing insights from census contexts, but also protects the anonymity of each person included in the original table.

As far as information leakage, it would be possible to find the number of people who own a certain breed of dog within a range of zip codes, but that would require the number of users in each bin being listed along with the other information in the new table. If the number of users is not included in the data set, that would effectively render this information useless, but a good middle ground would be to only include rows with a number of corresponding entries (from the previous table) greater than 0.

**The comments between lines 73 and 93 display this table format.**

**W5**: Using the insight “The number of users who aren’t disabled, live in a zip code between [15000,15100), and owns a Mini Schnauzer”, if one of the users included in that dataset was later said to own a different breed of dog/disability status was changed/etc. and the table was updated accordingly (or if another user was added), this would show a difference in terms of the number of users i.e. between dataset D with the normal number and D’ with this new addition or removal. If we do not report the number of users in the data set and instead use a query that asks for the breeds that exist in a certain zip code range, the same weakness is true.

**W7**: To satisfy differential privacy, the problem in **W5** could be solved by grouping entries further to group entries by ranges of breeds instead of individual breeds. That way, if one user/breed is removed from the data set, it wouldn’t present any useful information if a user/breed was removed from the table if the number of users per entry was not reported.

**The comments between lines 123 and 219 display this table format.**