Data Anonymization and Aggregation

How can we make data sets public and maintain privacy?

Big data problems

- Data can be used for tons of predictive and descriptive tasks
- Currently 2,500,000,000 Gigabytes of data are created each day.
- 90% of all data in the world were created in the last 2 years.
- Data is really useful because
  - High volume
  - Great variety of data types and topics
  - High Velocity of data

Personally Identifiable Data

- Some data isn’t about individuals
  - E.g. LHC data
  - Nothing really to worry about
- Other data is overtly personal
  - E.g. Medical records
  - Laws like HIPAA and FERPA
- Dangerous land where data is not obviously personal
  - Credit card transactions
  - Taxi cab rides
  - Targeted advertising

Math: how easy is identification

- Information theory: there are about 7.5 billion people in the world.
  - \( \log(\text{base } 2) \) of that is 32.8
  - So we can uniquely identify everyone with only 33 bits
- Each fact you know about someone reduces entropy by \( \log\) (base 2) of the improbability.
  - Gender is \( \log(2) \approx 1 \) bit
  - Birthday (w/out year) is \( \log(365) \approx 8.5 \) bits
  - KU \( \log(4352) \approx 12.08 \) bits (there are about 4352 universities in the world)
  - So birthday + KU reduces entropy from 33 bits to 12.5...

Dangers of Personally Identifiable data

- Even basic information can be personally identifiable when tied to other information
  - For example if I know that a particular data point belongs to a female in WestPoint, I can narrow down who it might be
- People can target individuals in a massive dataset and work to identify them
  - For example, Twitter tried to track down the Charlottesville protesters

Dangers of finding small communities

- One predictive method is to find out features of a particular community and then identify someone as a member of that community
  - You now know things about the new member even if you haven’t observed much about them
- If communities are too small they are too descriptive
  - Example: Facebook advertising groups
Possible solutions

- De-identification
- Data Anonymization
- Data Aggregation

De-identification

- Completely remove or mask the identifying information before publication.
- Instead of things like names, individuals are numbers
- Must remove quasi-identifiers as well
  - Things like zip-code, date of birth etc
- Issue: much of the data removed may be useful for predictions
  - For example, month of birth is correlated to success in sports and year with population differences
- Issue: some data can never be de-identified.
  - Genetic information

Data anonymization

- Encrypting identifiable information
- Difference between De-identification
  - Information structure kept
  - But in a way that is not human interpretable
- Issue: Because the information structure is kept, people may be able to reverse engineer the information.

NYC Taxi rides

- With a freedom of information request NYC taxi trips
  - 20 GB of data
  - 173 million trips
  - Anonymized license number and medallion number (taxi id)
  - Someone noticed that a particular driver had tons more drives than others
  - Someone else realized that that was actually the encryption for 0 or missing information
  - NYC taxi licenses are 6 digits
  - Medallion numbers are also structured
  - Now a cryptographically secure hashing function (which was used for encryption) isn’t secure
  - Guarantees require nothing to be known about the structure or content of the information
  - In the data world, we almost always have an idea about the structure.

NYC Taxi rides

- Not only could you identify the taxi used for specific routes
  - You could use additional information to find out who drove the taxi
  - And someone even used celebrity blogs, correlating times of photos of stars entering or exiting taxis, to expose the amount of tips particular celebrities paid.
Data Aggregation

- Aggregate a handful of data points so individuals are no longer identified.
- With a statistical approach, you shouldn’t be able to undo this to get information about an individual.
  - For example, aggregate between 3 and 6 people, add a bit of noise $N(0, .05)$, round to 2-3 significant digits.
- Data aggregation means that no individual is identifiable.
- Issues: how should we aggregate?
  - “Data can be either useful or perfectly anonymous but never both.” —Ohm

Data Aggregation: ways to aggregate

- One possibility, aggregate people who are the most similar:
  - Keep the highest correlations between input features and output predictions.
  - Mask individual differences.
  - What happens if you identify a (new) person as being similar to a particular group?
- Another possibility, aggregate randomly:
  - Randomly select a few people, aggregate.
  - Completely impossible to tell who is in the group.
  - But can we extend to new individuals?

Target’s target ads

- Target wanted to encourage pregnant women to shop at Target.
- But after you have a child, the birth is a public record and all sorts of companies can advertise to you.
- Target set out to find if they could tell someone was starting their second trimester.
- Mined their customer database:
  - all purchases women made weeks back.
  - from the birth announcement.

Target’s target ads

- As the story goes, a man came into the store with an advertisement addressed to his high school daughter.
- The manager apologized and called a few weeks later to again apologize.
- Instead the man apologized to the manager:
  - saying that his household was hiding the pregnancy from him.

What does this mean for Data Science

- If the community we are looking for (or use for aggregating) it may be damaging just to identify that someone is a member of a community.
- And misidentifications can be costly:
  - @yesYourARacist incorrectly identified someone.
  - He received threats and was unable to return to his university.

- Should we make data public? Should data be sold?
- If the service is free, you are the product. Should we have more pay for service options?
  - I am not on Facebook, if I could pay to have them not collect my data, would I be on Facebook?
- Should companies be more transparent on what types of information they have on you?
- Should companies be more transparent about the ‘bins’ or communities they are using for prediction?
- Should we have the choice to opt out of individualized results?