CYBER 759 – Privacy Attacks
Gunnar Yonker

1. Membership Inference Attack

A membership inference attack seeks to determine whether a specific data point was used in the training of a machine learning model. This type of attack can be particularly concerning when the model is trained on sensitive data, such as medical or financial records.

i. Gather Information: The attack collects information about the target machine learning model. This includes understanding its architecture, parameters, and any publicly available details about the training dataset.
ii. Shadow Model Creation: The attack constructs a shadow model, which is a replica of the target model. This shadow model is trained on data that mimics the distribution of the target model's training data but doesn't include the exact same data points.
iii. Attack Dataset Formation: The attacker creates an "attack dataset". This dataset consists of data points they're interested in probing for membership information.
iv. Train the Attack Model: Using features extracted from the target model's predictions and other data-related metadata, the attacker trains an attack model. This model's goal is to predict whether a given data point was part of the target model's training set.
v. Evaluate and Predict: The trained attack model is used to predict membership for the data points of interest. If the model predicts with high confidence that a specific data point was used during training, the attacker can assume its presence in the target dataset.

2. Singling Out Attack

A singling out attack focuses on isolating or identifying a specific individual within a dataset. The goal is to gather more details about that individual using either the same dataset or combining it with other external data sources.

For example, imagine a medical dataset containing anonymized patient records, including age, gender, diagnosis, and treatment. If an attack knows that John Doe, a 45-year old male, visited a specific hospital around a certain date, they might filter the dataset to identify records of 45-year old males who visited during that timeframe. If only one or very few records match, the attack can deduce other medical details about John, such as his diagnosis and treatment, thereby singling him out.

3. Strava Heat Map Attack

The Strava heat map attack occurred when Strava, an exercise tracking app, released an anonymized global activity map displaying users' jogging, biking, and hiking paths. Some sharp observers noticed that in regions with fewer Strava users, particularly in Africa and the Middle East, the exercise trails inadvertently revealed the locations of undisclosed US and NATO military bases.

Major Factors contributing to the Strava Heat Map Attack:

i. Insufficient Anonymization: Even though the data was anonymized, the specific routes taken by users in less populated regions made it easy to deduce the locations of military bases.
ii. Underestimation of Data Granularity: The high resolution and granularity of the data shared allowed for easy identification of patterns.

iii. Lack of User Awareness: Many users weren't aware their data could be used in such an aggregated form and presented publicly.

Potential Mitigation Strategies:

i. Improved Data Aggregation: Ensure that data is aggregated to a level where individual patterns cannot be easily discerned.
ii. User Awareness and Consent: Educate users about how their data is used and ensure they provide informed consent.
iii. Geo-Fencing Sensitive Areas: Implement algorithms to detect and exclude potentially sensitive areas from public maps.

4. Data Protection Measures

Given the company's desire to maintain user trust and adhere to privacy regulations, the following data protection measures can be considered:

i. Data Minimization: Only collect and store data that's absolutely necessary for the intended purpose. This reduces the risk associated with potential data breaches.
ii. Implement Differential Privacy: By adding noise to the data, differentia privacy ensures that the privacy of individual data points is maintained even when aggregated.
iii. Regular Privacy Audits: Conduct regular privacy impact assessments and audits to identify potential vulnerabilities. This proactive approach helps in anticipating and mitigating potential threats.

A reasonable data management policy should also involve continuously staying updated on emerging privacy threats and techniques, adapting defenses accordingly, and ensuring that users are educated about best privacy practices.