Research Title:

Temporal Analysis of Connected-Vehicle Location Data and its Implications for Reidentification Risk

Primary Investigator:

Name: **Dr. Ayelet Gal-Tzur**

Faculty: **Department of Industrial and Management Engineering, Faculty of Engineering**

Academic Institute: **Ruppin Academic Center**

As the capabilities of modern automatic data collection and storage systems for large-scale mobility data continue to advance, there is a growing awareness that this data can be leveraged to deduce potentially sensitive personal information. The recent technology of connected vehicles presents a new opportunity as a data source, and this source is anticipated to steadily expand as the proportion of new vehicles grows.

The goal of the current research is to evaluate the probability of reidentifying users based on floating car data (FCD). Within the scope of this investigation, the term “reidentification” is defined as the capability to deduce an entire set of a user's daily locations from a limited set of identified locations, which may be exploited by a potential attacker. This study aims to explore the impact of the number of locations that an attacker successfully identifies and the time-of-day at which this information is uncovered on user's reidentification probability.

The FCD data of four consecutive weeks were partitioned into two periods - first three weeks and fourth week. 24-hours diaries were constructed for each user and each day type for the first three weeks and for the fourth week. The diaries of users during the fourth week were sampled at a finite number of locations. These locations were then used to estimate the probability of each user to be reidentified in the fourth week based on the diaries of the first three weeks, taking into account the number of sampled locations and the time of day at which these locations were sampled.

As expected, this study's results show that when potential attackers identify more locations, the number of matches per user decreases, and simultaneously, the average Reidentification Score (RS) increases.

Thie study reveals that the trends of various phenomena become substantial with three detected locations and exhibit an asymptotic trend as this number increases to four and five detected locations, a finding that is aligned with the work of De Montjoye et al. (2013)[1].

Furthermore, the study indicates that the time of day when a user's location is detected has an impact on the reidentification probability. Generally, detecting a user's location during peak hours has a lesser effect on their reidentification risk compared to detecting it during off-peak hours, especially at night.

These findings emphasize the significance of taking into account the temporal aspect and potential risks associated with varying times of the day when ensuring the privacy of users.