MaskIt: Privately Releasing User Context Streams for Personalized Mobile Applications

Reviewed by Liyue Fan

Summary:
This paper studies a problem of sharing user context streams with a privacy constraint. The authors propose that the user contexts can be modeled as states and the transition between contexts can be captured by a first-order Markov model. For each incoming raw context, the proposed system will release or suppress the user state, depending on whether the \( \delta \)-privacy constraint is satisfied. The privacy constraint guarantees that the adversary’s posterior belief of user context with the released states (original or suppressed) should not exceed the prior belief by more than \( \delta \). This is because releasing a suppressed state will disclose user information too. Two algorithms are proposed under the same system framework: one is a search-based algorithm to find the optimal set of probabilities to release each state that maximizes utility under privacy constraint; the other is to simulate the adversary behavior and use the Markov model to predict user state based on previous observations. Results show simulatable (the second) algorithm and a hybrid algorithm outperforms the probabilistic (the first) algorithm.

Discussions: (related to my own research)
1. The authors stated that many human activities can be captured by Markov model, such as data collected by smartphone sensors, web search, etc. This may support us to model the web browsing trajectories with a Markov chain.
2. Their privacy constraint is an additive constraint that should be evaluated at every time point, which is the reason why to search for feasible solutions in the probabilistic algorithm. However, differential privacy provides automatic guarantee for aggregates. Thus we are spared from searching/simulating.
3. In the problem they define, each user may have a different state transition matrix. The utility metric is the ratio of unsuppressed states against the length of state sequence, which is also specific to each user. In the monitoring application, we will use the Markov chain to capture mass behavior/transition probabilities, which allows us to describe precisely the crowd movement/trend when the population is large.
**Challenges:**
Some challenges still exist:

1. We are to study the mass behavior of web browsing/or other similar activities and to predict the number of users on each page/state at every time stamp. It can be difficult due to the arbitrary length of each session. Most sessions are short and some, but rarely, sessions can be extremely long. This imposes a problem of data sparsity, after a few steps, very few users/sessions remain active and the majority will be “OFF” or “END”. My question is, does it make sense to consider “END” as a valid state as well? If it is, the total number of sessions will not decrease in this case.

2. Another relevant challenge is to model the number new sessions that may start (in proportion to the existing sessions, or fixed amount) at each time point. This may also help with the data sparsity issue with long sequence, but may not be needed if “END” is considered as a state.

3. The process noise for each state may be different, e.g. the pages visited more frequently show more variance in count and less visited pages show little variance. How to find out the process noise variance for each state?

4. The measurement noise captures the perturbation added to each count. Which perturbation mechanism should be used, Laplace or Gaussian?

5. Since recursive estimation now requires matrix-matrix multiplication, the complexity for every time stamp is roughly $O(m^3)$, with $m$ being the number of states/pages.

6. The training of Markov chain may arguably violate differential privacy.