NIST Differential Privacy Synthetic Data Challenge

Q & A Session for Challenge Webinar

January 15, 2019

Q: I am participating in this challenge as part of a team. What does our Team Captain need to do and what do our team members need to do in order to participate?
A: The team captain fills out the team form and lists the Topcoder handle (and optionally, the name) of each member. The team captain should be the one to make submissions to Topcoder platform and the Differential Privacy Algorithm write ups for review. All members of the team should sign up with Topcoder and register their handles. The team captain will be receiving the awards from NIST. The team captain must be a US citizen.

Q: What information about the dataset is considered public and does not need to be protected by DP? In addition, can we utilize additional public datasets to develop our approach and how do we ensure that doing so doesn't violate DP?
A:

- The table schema (the names of the columns, and the set of *possible* values in each column) is public and constant--it does not change between provisional and sequestered phases. The table schema is available in the documentation that accompanies the data. You should absolutely not hard-code any part of your solution to the set of values that actually appear in the training data-set, rather than the set of possible values.
- Public information relevant to the San Francisco data, such as precinct maps, etc, can be used to improve accuracy in your algorithm. In your write up, include which information you're using, links to sources online, and a clear explanation of how/where in the algorithm you're applying this information.
- Our target application is a data owner who's released anonymized data annually for years and is now switching to differentially private synthetic data (specifically, we're looking at cases where the data-sets cover a very small sub-sample of the population and reflect largely non-overlapping sets of individuals from year to year). In this case, in the provisional phase you are developing and refining your algorithms using the already released data from previous years. When you're invited to submit your code to the

sequestered phase, you're essentially transitioning your prototype system to production, and at this point you need to be strictly preserving the differential privacy guarantee over the sequestered data-set. Your performance on the sequestered data is what determines your final ranking and prize eligibility.

- With the statement that the provisional phase data is considered publicly available, it's important to note that… you still do have to build a differentially private ***synthetic data generator***, i.e. a tool that takes as input a sensitive/private data-set and builds a new synthetic data-set based on this input, while satisfying differential privacy. In the sequestered phase your solution must use the input sequestered dataset: it's not sufficient to simply return the provisional data in place of generating a synthetic data-set, or build a non-DP synthesizer trained only on the provisional data. In general, it's a good idea to assume that the provisional dataset may have a significantly different distribution than the sequestered dataset.

Q: How will the contest communicate to us about important updates, questions with our entry, final scoring, etc.?
A: The forum! Especially watch the forum at the end of the match, and during the sequestered phase.

Q: Any advice for submitting docker executables for the sequestered testing?
A: Make sure they run reliably, and within a reasonable number of hours (< 24), even on arbitrary epsilon.

Q: Will the final evaluation use eps > 0.01?
A: The final/sequestered evaluation exists to confirm the relative ranking and prize eligibility of the entrants. It will initially use values of epsilon between 1 and 0.01, looking at discrete AUC to determine the final score, however a wider range of epsilon will be used if two teams have such similar performance that it's not possible to clearly determine their ordering.

Q: Is this match using event data, or microdata tables such as those released by statistical agencies?
A: We're using event data again in this match. In Match #3 we plan to use statistical data.

Q: Do those who participate in first round still need to submit the prescreen even if the algorithm does not change? Also, are we allowed to modify our algorithm after submission prescreen?
A: If the algorithm really hasn't changed from Match #1, that's fine, you don't need to submit to prescreen and your score boost will carry over. However, if you do make any significant changes (ex: use of new public data, or changes that affect function sensitivity or noise addition processes), then we recommend submitting an updated write-up highlighting the changes for pre-screening. If you're invited to the final evaluation and there's a violation of differential privacy in your solution at that point, you will be ineligible for a prize and you'll retroactively lose your prescreen bonus (affecting your final ranking for the match). Although prescreens are quick spot-checks and aren't guaranteed to catch all mistakes, keeping your write-up and prescreen up to date makes it more likely any issues will be caught early.

Q: Can non-US citizens participate?
A: Yes, non-US citizens can participate on a team as long as the team captain is a U.S. citizen. If you do not have a U.S. citizen as your team captain, then you may participate but you will not be eligible for prize money, as stated in Match Contest Rule #8. Per the NIST Official rule #3B on challenge.gov, the official representative (team captain) must be age 18 or older and a U.S. Citizen or permanent resident of the United States or its territories to receive a prize payment. Prior to payment, the winners will be required to verify eligibility.
IMPORTANT NOTE: All contest submissions must be submitted from the team captain's account; this includes algorithm write-ups submitted to the DP Pre-screen process.