NIST Differential Privacy Synthetic Data Challenge

Q & A Session for Challenge Webinar

November 13, 2018

Q: Does the algorithm have to be original? Can the implementation be based on existing algorithms?
A: Your implementation can be based on an existing algorithm or algorithms. If that's what you choose to do, you should be sure to cite your references, and you should still create your own algorithm write-up and privacy proof to submit for DP screening.

Q: Can R software be used for this competition?
A: Yes, there are no restrictions on the software that you use, as long as TopCoder is able to run your code on their machines if you're selected to compete in the final/sequestered round.

Q:There are well-known papers that solved this problem. I wrote papers to solve this problem when that data is partitioned between two entities. Can I participate?
A: Yes, you may participate, but note that we're using the central model for differential privacy here: We assume that the raw data is owned by a single trusted data owner, and that owner is creating a differentially private synthetic data set for public release.

Q: How much can I refer to a previous paper's proof? Or must I write up my own proof in my own words? How detailed are you expecting (say, level of peer-reviewed conference academic paper)?
A: Please write your proof in your own words *unless* you are the author of the proof from the published paper. The proof should be clear, complete and correct. Rather than thinking about peer-reviewed conference papers (which can have strict page-limits that result in foreshortened proofs with omissions), consider how you would present the proof during a classroom lecture: Focus on being explanatory and complete.

Q: Scoring - when partitioning continuous values into 100 buckets, the range is from my data, correct?
A: The bucket range is from the original data, not the synthetic data. In general, scoring looks at a set of analyses (in this Match, bucketed 3-marginals) on the original data, then runs those *identical* analyses on your synthetic data (i.e., the same marginals, with the same buckets), and compares the results between the two data-sets.

Q: Scoring: Given the scoring scheme is public in a .zip file, what is the advantage to submitting online? I can check my score using the algorithm in the .zip file
A: Submitting your code online will allow you to be benchmarked on the leaderboard. Only the top 10 on the leaderboard at the end of the match will be invited to compete in the final sequestered phase, in which the prizes will be awarded.

Q: How will you check privacy? My algorithm is randomized so I might, by luck, sample very small noise values and generate something very close to the real data.
A: We check privacy by evaluating your algorithm rather than its output, through the DP Prescreening and DP Final Validation processes (which you can read about on the problem statement or in the rules). As far as luck: at the end of the marathon match we ask the top 10 teams to submit their code for a thorough evaluation (the Sequestered Phase). During this evaluation we'll run everyone's algorithms many times, over a range of epsilon values, and identify each algorithm's average privacy/utility trade-off curve. The top 5 teams after this evaluation will be eligible for prizes. Luck might get someone a good spot on the leaderboard during the Testing Phase, but it won't easily help during the Sequestered Phase.

Q: How is the differential privacy bound (epsilon) checked on submitted files?
A: During the Testing Phase, we're going by the honor system for epsilon values (although if we see a team significantly outperforming what we expect to be possible we may ask them to submit again to DP prescreening). When we invite the top 10 teams to the Sequestered Phase at the end of the match, they will be submitting their code and final write-ups, and we'll be doing a source code review and validation, in addition to a much more thorough performance evaluation.

Q: Since there is some randomness in result from randomized algorithms, in the small chance that the first output when you test my algorithm behaves differently, would you rerun my algorithm multiple times to check?
A: Yes, we will run algorithms multiple times during the final scoring in the Sequestered phase.

Q: Is the execution time of the algorithm being scored?
A: Nope! Not at all; that's out of scope for this competition. However, if you're invited to the Sequestered Phase, we will need to be able to run your code without inducing any memory errors.

Rules & Regulatory Questions

Q: My team captain won't play any role in the design of the algorithm, but will only handle administrative matters i.e. payment. Is that OK?
A: Yes, that is fine. Your team captain must make all submissions to the Topcoder platform and will be responsible for all official communications and administrative matters.

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.

Q: Regarding tax: will the entire prize amount forwarded to other teammates necessarily be considered by IRS as income for the team captain himself? If not, what is the legal status of this payment, and how could this be explained to the IRS?
A: For the prize winners, NIST will make one single payment per cash prize. We recommend that the team captain (Official Representative) and members meet with their tax advisor in order to properly receive the prize funds. There may be tax implications for the Official Representative or organization's tax id number receiving the funds.

Q: Will my submission be public after the deadline or will it be kept private?
A: Your submission will remain private and will only be accessible to the Challenge Sponsor as stated in NIST Official Rule #4 on challenge.gov.

Q: Are algorithms submitted still owned (copyrighted) by the submitted? Or will you own (in whole or in part) it?
A: All intellectual property created in the submission will remain with the competitor. NIST will not own any IP submitted to this challenge as stated in NIST Official Rule #6 on challenge.gov.