The NIST Collaborative Research Cycle Explanatory Workshop 2023 Call for Papers

17 Aug 2023 - v1.1
NIST Contact: Gary Howarth

Event	Date
Submission Deadline	7 NOV 2023
Review Deadline	28 NOV 2023
Notification	1 DEC 2023
Workshop	18 DEC 2023
Proceedings camera ready	5 JAN 2023
Proceedings published	24 JAN 2023

Key Goals and Aims

The NIST Collaborative Research Cycle (CRC) Explanatory Workshop is a venue to discuss and present topics related to the NIST Diverse Community Excerpts data and the CRC Data and Metrics Archive.

The Explanatory Workshop invites the research community to submit 'tiny' papers (≤ 4 pages + appendices). Our workshop's tiny papers are inspired by the ICLR tiny paper initiative and TPDP. Submissions will undergo a light, single-blind (anonymous reviewers) peer review process.

Featured submissions will present their papers on the workshop date. Accepted submissions will be prepared into a *non-archival set of proceedings* made available on the CRC website. Participation in the CRC Workshop is not intended to preclude authors from publishing their research elsewhere.

Details

Why should we care?

The NIST CRC seeks to strengthen our understanding of tabular data deidentification technologies. Deidentified data is an ambitious attempt to democratize the benefits of big data; it uses generative models, noise infusion strategies, or anonymization algorithms to recreate sensitive personal data with sanitized or synthetic records suitable for public release. However, it is vulnerable to the same bias and privacy issues that impact other data applications, and

can even amplify those issues. When deidentified data distributions introduce bias or artifacts, or leak sensitive information, they propagate these problems to downstream applications. Furthermore, real-world survey conditions such as diverse subpopulations, heterogeneous non-ordinal data spaces, and complex dependencies between features pose specific challenges for data deidentification algorithms.

What are we trying to achieve?

Shared benchmarks promote understanding and exploration of a problem by providing common resources, vocabulary, and analytic framework. These features allow a community to coordinate efforts on a particular problem.

The CRC seeks to advance understanding of deidentification technologies by analyzing deidentified instances of the NIST Diverse Communities Data Excerpts (Excerpts). The Excerpts consist of a small curated geography and feature set derived from the significantly larger 2019 American Community Survey (ACS) Public Use Microdata Sample (PUMS), a publicly available product of the U.S. Census Bureau.

From February 2023 to the present, NIST has collected over 350 deidentified instances of the Excerpts from the research community. We have performed a robust evaluation of the deidentified data using the SDNist Report Tool, which provides a host of visualizations, statistical evaluations, and explorations of the deidentified data.

We have released the CRC Data and Metrics Archive containing the contributed data alongside detailed evaluation metrics for each set in human- and machine-readable format. We have also developed a variety of tools and notebooks to assist exploration.

The NIST Collaborative Research Cycle Explanatory Workshop seeks original research (in the tiny paper format) using these resources.

Scope and Possible Paper Topics

We warmly welcome all perspectives. Potential paper topics include but aren't limited to:

- Definition of a somewhat novel yet simple idea (ex: metric, statistic, privacy enhancement), and application of this idea to the CRC archive data, including implementation and exploration/evaluation.
- Comparison of an existing or new deidentification strategy to others in the archive (both similarities and differences), or an improved parameter configuration for a deidentification strategy that's already in the archive.
- A modest and self-contained theoretical result illustrated by evaluations on the Diverse Community Excerpts data or CRC archive data.
- Evaluation of a concept in a *previously published paper*, as applied to the Diverse Community Excerpts data, SDNist metrics, and/or CRC archive. We welcome new evaluations of existing ideas in light of the archive—these are especially valuable as examining research against the common benchmark framework can help us combine,

compare and find broader implications of otherwise disparate ideas. The content of the original paper can be limited to a short summary and citation/link, with the focus of the CRC paper being on a new evaluation and discussion section grounding the previous work on our benchmark data.

It is recommended to use multiple submissions to cover multiple ideas—we expect these tiny four page papers to be relatively quick to write. You're on a scavenger hunt in our Research Acceleration Bundle and we're interested in anything you encounter that you would like to share.

We do encourage analytical thinking rather than just high score checking (see scavenger hunt advice below). Try to show how ideas fit together, talk about their implications, point out anomalies and areas of uncertainty.

Submission Preparation Guidelines

- Papers must relate to the Diverse Communities Excerpts benchmark data (Use of the SDNist evaluation report generator, PCA Metric Explorer, or the CRC Data and Metrics archives are optional and encouraged)
- Tiny papers: 4 pages + appendix. (see ICLR DEI track, TPDP).
- Recommend picking one focused topic per paper, multiple submissions for multiple ideas are encouraged.
- There is no prescribed format, though our Latex template is recommended (source files; Overleaf link). Text should be ≥ 10 pt
- The essential claims, arguments, and conclusions should be in the four pages of the tiny paper. Please do not treat the tiny paper as a four-page abstract with a copious body in the appendix.
- Submissions are welcome to build on previous publications with new analysis on Excerpt data. Please link to a publicly accessible copy/preprint of the original paper, but please don't include the text of previously published (and potentially copyrighted) papers in your submission.
- You are welcome to include links to Github (or similar) repositories if you would like to provide code supplementing your submission.
- Submit at: https://cmt3.research.microsoft.com/NISTCRC2023/

Submission Checklist

- ☐ Read and accepted the submission terms and conditions (in CMT)
- ☐ Appropriately cited the works of others
- ☐ Did not include full text of any previously published papers (used link instead)
- ☐ Verified that all text is proofread and legible
- ☐ Labeled figures and axes, used sensible axis and heatmap ranges

Analytical meta-analysis—Scavenger Hunt Advice:

The CRC Archive is challenging and exciting for the same reason: It has a lot of things in it—many parameters, many approaches, many types and concepts of privacy. The only common element is the Excerpts benchmark data. This means there's a wide array of interesting observations to be made, but it can also make exploration more difficult. So here's a few pointers to keep in mind that may help:

- First try to suspend your judgment– begin exploring with a broad curiosity rather than a specific goal.
- Pick a couple significant factors (metric results, data properties, etc) and think about how they may be related.
- Focus not just on 'good' or 'bad' algorithm performance, but also look for details and observations that are interesting, strange or revealing.
- Make the implicit explicit– state your ideas clearly, and then work to form chains of questions and implications. Is your idea really accurate? How do you find out if it is really accurate? What do you do about what you find out– what else does your idea imply? If it is quite accurate, how do you prove its accuracy?
- Look for patterns of repetition, contrast and anomalies– how can you generalize your idea? Where else would you expect it to appear (and can you check to see if it actually does)? Are there any exceptions or areas where it doesn't apply?

Review Guidelines

All submissions will go through a friendly, lightweight review process. The reviews are intended to select papers that are appropriate for this workshop and to improve submissions. We have no specific limitations on the number of papers that will be accepted.

- Appropriateness:
    - Does the submission use the Diverse Communities Excerpt Data?
    - Is the submission germane, coherent, and free of offensive material?
    - Does the submission situation itself within the context of other research through citations?
    - Does the submission follow the format guidelines?
        - Main arguments, claims, and conclusions are contained within the four-page tiny paper format.
        - Formatting is clear, legible, and well labeled.
- Clarity:
    - Does the submission make novel and interesting observations and/or arguments?
    - Does the submission demonstrate clear, straightforward reasoning with easy-to-follow connections?
- Correctness:
    - Are there sufficient evaluation and/or theoretical work to support the claims and conclusions in the submission?
    - Is the submission free of major errors or omissions?

Call For Papers Release Notes

v1

- Initial publication

v1.1

- Addition of submission link
- Remove invitation to submit Python notebooks
- Formatting adjustments