An Exploratory Meta-Analysis to Identify Outlying Behavior in the NIST Collaborative Research Cycle Archive

Dhruv Kapur[1], Jeremy Seeman*[2]

Abstract
The NIST Collaborative Research Cycle (CRC) archive provides a collection of deidentification techniques applied to three Diverse Community Excerpts (DCE) datasets. In this exploratory meta-analysis, we propose a metric for evaluating errors in pairwise features from the DCE datasets to assess the quality of categorical pairwise associations relative to their prevalence in the dataset. Using this metric, we identify outlying pairs in where deidentification algorithms overperform or underperform relative to the pairwise association's prevalence. We conclude by proposing follow-up work to leverage these metrics as more generalizable evaluation tools.

Keywords
Meta-Analysis,Evaluation

[1] *College of Engineering, University of Michigan; Intern, Knexus Research Coperation*
[2] *Michigan Institute for Data Science, University of Michigan*
***Corresponding author**: jhseeman@umich.edu

Contents

1. Introduction

Deidentification techniques for producing synthetic data have been well-studied for their empirical [1] and formal [2] confidentiality protections. In this paper, we focus on empirical measures of synthetic data utility, specifically the ways that deidentification techniques tend to produce synthetic datasets that are more homogeneous than their target datasets. While making synthetic data appear more homogenous is a necessary consequence of applying any privacy-enhancing technology with confers disclosure avoidance protections [3–6], associations between smaller subsets of the data are harder to preserve while protecting data subject confidentiality, a phenomenon yielding inequitable privacy-utility trade-offs for different demographic subgroups [7, 8]. Different algorithms for deidentifying data may differ in how effectively they do or do not preserve these relationships. Simultaneously, certain relationships within the data may be a priori more difficult to preserve, regardless of which algorithm gets applied. It therefore behooves us to develop metrics to disentangle relationships in the data that are easily preserved by many mechanisms versus those where we can differentiate algorithm performance.

As an example testbed to develop such metrics, the NIST Collaborative Research Cycle (CRC) archive [9] contains examples of deidentification algorithms applied to the Diverse Communities Excerpts (DCE) datasets drawn from public-use microdata areas (PUMAs) in the American Community Survey (ACS) [10]. Each implementation in the archive contains privacy and utility metrics comparing the target data to the deidentified data. In this paper, we perform an exploratory analysis to empirically investigate the deidentification techniques in the archive. We propose metrics for identifying outlying relationships across algorithmic techniques, allowing us to isolate anomalous structures in the DCE datasets that may be of interest for evaluating synthetic data and potential equity issues in subpopulation data quality.

Note that this empirical meta-analysis of the NIST CRC archive is meant to serve as a description of submissions to the NIST CRC, not a description of the algorithms themselves. Because the set of possible implementation parameters was not exhaustively explored in ways that compare utility guarantees at the same privacy level, the analysis does not immediately yield conclusions about algorithmic performance more generally. Future work to make the preliminary analysis more comprehensive is discussed in Section 3.

2. Analysis Description

2.1 Setup and Metrics

We examine CRC archive entries which synthesize all features in the DCE datasets. Table 1 describes the number of submissions by target dataset, algoritm type, and privacy category, and Table 4 describes the privacy loss budget configurations for differentially private (DP) algorithms. In particular, we focus on the subset of features described in Table 3.

Table 1. Description of CRC algorithm implementations by target dataset, algorithm type, and privacy category.

target dataset	algorithm type	privacy category	count
ma2019	neural net	dp	3
		non-DP	7
	query matching	dp	2
	sdc	sdc	1
	stat model	dp	1
		non-DP	1
national2019	neural net	dp	5
		non-DP	15
	query matching	dp	4
	sdc	sdc	3
	stat model	dp	1
		non-DP	1
tx2019	neural net	dp	3
		non-DP	7
	query matching	dp	2
	sdc	sdc	1
	stat model	dp	1
		non-DP	1

For each dataset $s \in \mathscr{S} \triangleq \{\mathrm{NA}, \mathrm{MA}, \mathrm{TX}\}$ in the CRC, each of size N_s containing features $d \in \mathscr{D}$, we consider a subset of categorical feature pairs $(d_1, d_2) \in \mathscr{D}_{\text{pairs}}$ defined in Table 3. For each feature pair, we refer to the levels on both features as "nodes", forming node pairs $(n_1, n_2) \in \mathscr{N}_{d_1,d_2}$. For example, when considering the feature pair $(d_1, d_2) = (\texttt{EDU}, \texttt{OWN\_RENT})$, one example node pair sets $n_1 = \texttt{GED}$ and $n_2 = \texttt{'rent'}$ so $(n_1, n_2) \in \mathscr{N}_{\texttt{EDU},\texttt{OWN\_RENT}}$. For brevity, we refer to the collection of (n_1, n_2, d_1, d_2) as $\vec{n} \in \vec{\mathscr{N}}$ where contextually appropriate.

We are interested in pairwise counts, or the co-occurrence of node pairs, in the target data (for example, the number of college-educated renters in the national PUMAs). These serve as sufficient statistics for models that investigate relationships between these discrete variables [11]. We represent these counts as $x_{(\vec{n},s)}$. For a given dataset $s \in \mathscr{S}$ and deidentification algorithms applied to that dataset $a \in \mathscr{A}_s$, let $\tilde{x}_{(\vec{n},s,a)}$ be the count constructed from the deidentified data. We then first focus on absolute error, i.e.,

$$\mathrm{AbsErr}(\vec{n}, s, a) \triangleq |x_{(\vec{n},s)} - \tilde{x}_{(\vec{n},s,a)}| \tag{1}$$

To establish the relative difficultly of generating counts for certain node pairs over others, we consider the following metric, $\mathrm{AbsErrScore}(\vec{n}, s, a)$, defined below as

$$\begin{aligned}&\mathrm{AbsErrScore}(\vec{n}, s, a) \triangleq \\ &= \frac{1}{|\mathscr{N}_{d_1,d_2}||\mathscr{A}_s|} \sum_{(n_1^*, n_2^*) \in \mathscr{N}_{d_1,d_2}, a \in \mathscr{A}_s} \mathbb{1}\Big\{ \\ &= \mathrm{AbsErr}((n_1^*, n_2^*, d_1, d_2), s, a) \leq \mathrm{AbsErr}(\vec{n}, s, a)\Big\}\end{aligned} \tag{2}$$

This metrics computes the percentile of score for a particular node pair relative to all other node pairs for that feature $\mathscr{N}_{d_1,d_2}$ and algorithms applied to the same dataset $a \in \mathscr{A}_s$. We then define $\mathrm{AbsErrScoreMed}(\vec{n}, s)$ as the per-node-pair median score taken over all algorithm implementations for that particular node pair and dataset, i.e.,

$$\mathrm{AbsErrScoreMed}(\vec{n}, s) \triangleq \underset{a \in \mathscr{A}_s}{\mathrm{Median}}(\mathrm{AbsErrScore}(\vec{n}, s, a)) \tag{3}$$

This metric captures the relative difficulty to preserve the relationship between a particular feature and node-pair across all algorithms and node-pairs, for a given fixed dataset.

As is true with many deidentification algorithms that aim to minimize uniform error measures, absolute errors associated with larger target counts tend to be larger. For the NIST CRC archive, we empirically observe that, for non-zero target counts, our proposed metric and $\log_{10}(\vec{x}_{\vec{n},s} + 1)$ are approximately linearly correlated for each node pair, i.e., for a fixed feature pair d_1, d_2 and dataset s, we can approximate

$$\begin{aligned}&\mathrm{AbsErrScoreMed}(\vec{n}, s) \approx \\ &\beta_{s,(d_1,d_2),0} + \beta_{s,(d_1,d_2),1} \log_{10}(x_{(\vec{n},s)} + 1)\end{aligned} \tag{4}$$

Figure 1 shows one such regression model for the feature pair `[RAC1P, HISP]`, which shows the approximately linear relationships between target counts and $\mathrm{AbsErrScoreMed}(\cdot)$ across all three target datasets. Similarly, in Figure 2, we see that these large Pearson correlations persist across feature pairs in our empirical investigations and across datasets, yielding correlations greater than .9 for the majority of dataset and feature pair combinations. Note that because we are primarily interested in identifying outlying behavior, additional nonlinear complexity in the model form has a negligible effect on outlier identification, but would be relevant for future work.

We use this relationship to interpret model residuals. Large positive residuals refer to node pairs where the CRC submitted algorithms overall perform worse than we would expect for node pairs of similar size in the DCE datasets. Conversely, large negative residuals refer to node pairs where the deidentification algorithms overall perform better than we would expect for node pairs of similar size in the DCE datasets. This helps us determine pairwise structures in the DCE datasets where the CRC submitted algorithms overperformed (large negative residuals) or underperformed (large positive residuals) on the particular tasks.

We select the top 1% of model residuals by absolute value across all feature pairs and datasets and partition them into overperformers and underperfomers based on their sign, negative or positive, respectively. We then investigate how different classes of methods perform on these outlying node pairs.

2.2 Example Outliers by Algorithm Type

First, in Table 2, we select and interpret the top 3 outliers by absolute residual value over different subsets of the feature pairs. The node pairs selected by our data-driven method reveal issues with deidentification techniques that generally agree with long-standing difficulties in demographic data collection and modeling [12]. First, looking at Table 2 a), we see that algorithms in the NIST CRC archive struggled to capture pairs involving multiracial or multiethnic respondents, mirroring challenges in designing items to capture complex racial and ethnic identities and their relationships to other covariates [13, 14]. Second, we see that the algorithms in the CRC archive struggled to capture behaviors for young adults, both in Table 2 b) and Figure 3, showing large numbers of outlying residuals for young adults between 18-30; again, young adults pose distinct demographic challenges due to their (generally) higher variability in geography, income, and housing types [15, 16]. Finally, we see that node pairs that conform to broad population trends tend to be easier to estimate across algorithm implementations. For example, in Table 2 c), we see that implementations in the CRC archive overperformed on nodes where education and income were highly correlated but underperformed where education and income were less correlated. Overall, our metric isolates node pairs whose overperformance or under-performance cannot be explained by size alone, revealing the empirical consequences of synthetic data's normalizing effect on hard-to-study subpopulations.

Turning now to algorithm types, Figure 4 shows how the methods perform relative to one another within the NIST CRC archive. In the top subfigure, we evaluate overperforming outlying node pairs and make a few observations. First, we notice that for the implementations in the archive on overperforming algorithms, non-DP methods tend to outperform DP methods within the same algorithm type (`neural nets` and `stat model` algorithms have examples of DP and non-DP algorithms). This sanity check agrees with DP theory, where the additional noise injected into model training produces lower-accuracy results than their non-DP counterparts. For underperforming nodes, though, we do not observe a major difference between DP and non-DP methods. Without more comprehensive coverage in the NIST CRC archive across algorithm implementation differences, we cannot attribute these effects to the algorithms themselves.

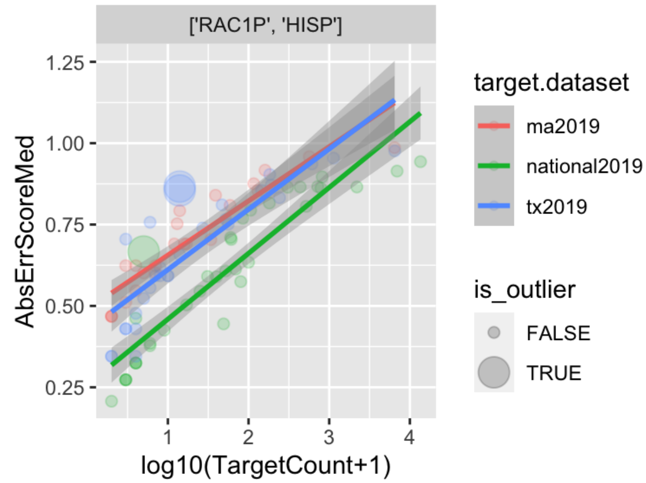

Figure 1. Example scatterplot of AbsErrScoreMed$(\cdot)$ versus $\log_{10}(x_{(\cdot)}+1)$ for `RAC1P x HISP` node pairs by dataset, with outliers flagged. Shaded error bars correspond to a 95% prediction interval from the model in Equation 4.

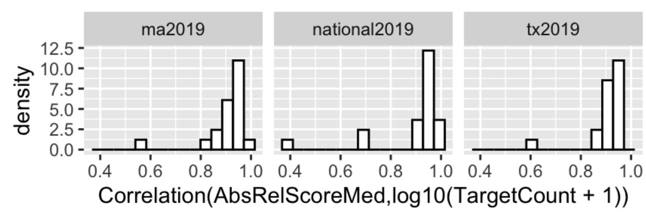

Figure 2. Histograms of Pearson correlations between AbsErrScoreMed$(\cdot)$ versus $\log_{10}(x_{(\cdot)}+1)$

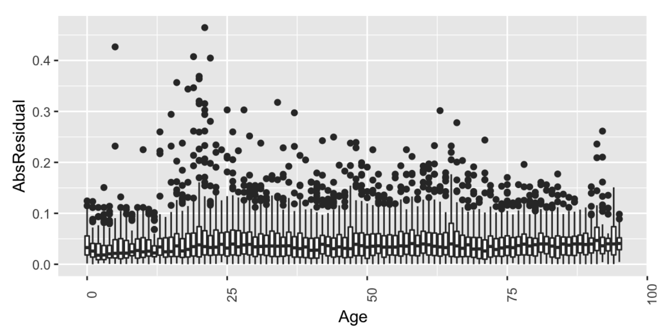

Figure 3. Residual absolute values for `Age` node pairs.

3. Discussion and Future Work

This empirical, exploratory meta-analysis of the DCE archive allows us to examine new metrics for evaluating how well deidentification techniques perform relative to one another *and* relative to the difficulty of preserving particular pairwise feature relationships in the data. Our goal with this preliminary work was to identify overall patterns in the CRC archive, but future work is needed to make more generalizable claims that aid in the evaluation and policy assessment of deidentification techniques. To this end, we propose a few follow-up directions for how such analyses could be extended as general

Table 2. Top 3 residual outliers by feature pair groups.

a) Top 3 residual outliers for all categorical feature pairs

left_feat	right_feat	left_label	right_label	abs_residual	overperf
RAC1P	HISP	Asian alone	All other Spanish/Hispanic/Latino	0.27	FALSE
RAC1P	HISP	Some Other Race alone	Not Spanish/Hispanic/Latino	0.23	FALSE
MSP	OWN_RENT	Widowed	Own housing unit	0.22	TRUE

b) Top 3 residual outliers for all categorical x AGEP pairs

left_feat	right_feat	left_label	right_label	abs_residual	overperf
AGEP	MSP	21	Now married, spouse present	0.46	FALSE
AGEP	EDU	5	Grade 1 to grade 8	0.43	FALSE
AGEP	MSP	19	Now married, spouse present	0.41	FALSE

c) Top 3 residual outliers for all categorical x PINCP_DECILE pairs

left_feat	right_feat	left_label	right_label	abs_residual	overperf
EDU	PINCP_DECILE	No schooling completed	80th percentile	0.35	FALSE
PINCP_DECILE	SEX	20th percentile	Female	0.31	TRUE
EDU	PINCP_DECILE	Grade 9 to grade 12, no diploma	20th percentile	0.31	TRUE

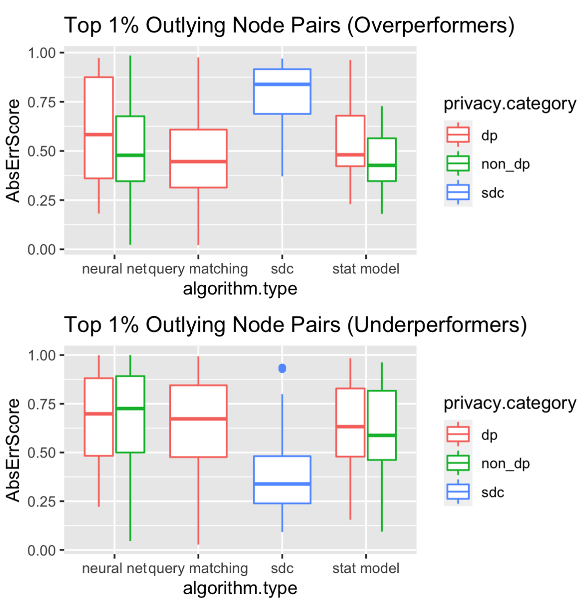

Figure 4. AbsErrScore$(\cdot)$ distributions by algorithm type for all entries in the CRC archive for overperforming node pairs (top) and underperforming node pairs (bottom).

evaluation tools.

First, because the CRC archive only contains one application of each algorithm to the DCE datasets, we have no way of empirically accounting for randomness in the applied algorithms, particularly for methods which require additional noise for privacy preservation. For example, it could be the case that some node pairs identified as outliers are artifacts of the particular pseudo-random number generation yielding the synthetic data in the CRC archive. Future would will need to consider multiple replicates of each algorithm implementation to ensure evaluation interpretations are robust to randomness in the synthetic data generation process.

Next, our meta-analysis did not consider how different algorithmic techniques provide different privacy guarantees. For example, we do not consider how differences in disclosure risk vary across formal and empirical methods, opting to instead analyze the NIST CRC archive as a whole. Similarly, some DP algorithms are implemented at different privacy loss budgets, injecting different levels of noise into the output statistics. As a result, differences in utility have been decontextualized from differences in privacy risks, requiring apples-to-apples comparisons at the same level of disclosure risk for these tools to yield effective evaluations.

By enabling more complete coverage and additional replications of each algorithm, we could develop more sophisticated models for relative node pair performance and outlier detection metrics. Such tools would help practitioners identify dataset structures and associated algorithm classes that are likely or unlikely to be sufficiently captured by synthetic data.

References

[1] Leon Willenborg and Ton De Waal. *Elements of statistical disclosure control*, volume 155. Springer Science & Business Media, 2012.

[2] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[3] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[4] Cynthia Dwork and Moni Naor. On the difficulties of disclosure prevention in statistical databases or the case for differential privacy. *Journal of Privacy and Confidentiality*, 2(1), 2010.

[5] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 193–204, 2011.

[6] John Abowd, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Daniel Kifer, Philip Leclerc, William Sexton, Ashley Simpson, Christine Task, and Pavel Zhuravlev. An uncertainty principle is a price of privacy-preserving microdata. *Advances in neural information processing systems*, 34:11883–11895, 2021.

[7] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct publication of the 27th conference on user modeling, adaptation and personalization*, pages 309–315, 2019.

[8] David Pujol and Ashwin Machanavajjhala. Equity and privacy: More than just a tradeoff. *IEEE Security & Privacy*, 19 (6):93–97, 2021.

[9] Christine Task, Karan Bhagat, Streat Damon, and Gary Howarth. NIST Diverse Community Excerpts Data, December 2022. URL `https://data.nist.gov/od/id/mds2-2895`.

[10] U.S. Census Bureau. American community survey, 2019. URL `https://www.census.gov/programs-surveys/acs`.

[11] Alan Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.

[12] Seymour Sudman and Graham Kalton. New developments in the sampling of special populations. *Annual review of Sociology*, 12(1):401–429, 1986.

[13] Eileen Patten. Who is multiracial? depends on how you ask. *Pew Research Center [Internet]. Available: https://www. pewresearch. org/social-trends/2015/11/06/who-is-multiracial-depends-on-how-you-ask*, 2015.

[14] Miri Song. Who counts as multiracial? *Ethnic and Racial Studies*, 44(8):1296–1323, 2021.

[15] Richard A Settersten Jr. The contemporary context of young adulthood in the usa: From demography to development, from private troubles to public issues. In *Early adulthood in a family context*, pages 3–26. Springer, 2011.

[16] Frances Goldscheider. Why study young adult living arrangements? a view of the second demographic transition. In *Workshop-Leaving home: A European focus*, 2000.

A. Additional Tables and Figures

Table 3. Feature names, types, and domains for the DCE datasets.

feature name	type	domains
AGE	integer ordinal	$\{0-99\}$
EDU	categorical	$\{N,1,\ldots,12\}$
HISP	categorical	$\{0,\ldots,4\}$
HOUSING_TYPE	categorical	$\{1,2,3\}$
MSP	categorical	$\{N,1,\ldots,6\}$
NOC	integer ordinal	$\{N,0-19\}$
OWN_RENT	categorical	$\{1,2,3\}$
PINC_DECILE	integer ordinal	$\{N,0-9\}$
RAC1P	categorical	$\{1,\ldots,9\}$
SEX	categorical	$\{1,2\}$

Table 4. Description of CRC DP algorithms by privacy loss budget.

epsilon	delta	count
1	10^{-5}	5
	$3.6*10^{-6}$	1
	NaN	3
5	NaN	6
10	$1/n^2$, where n is the data size	1
	10^{-5}	3
	NaN	3

Table 5. Description of all-features deidentification algorithm implementations in the NIST CRC archive by target dataset, algorithm type, privacy category, and privacy loss parameters for DP algorithms.

target dataset	algorithm type	privacy category	epsilon	delta	count
ma2019	neural net	dp	1	NaN	1
			5	NaN	1
			10	NaN	1
		non-DP	NaN	NaN	7
	query matching	dp	1	10^{-5}	1
			10	10^{-5}	1
	sdc	sdc	NaN	NaN	1
	stat model	dp	5	NaN	1
		non-DP	NaN	NaN	1
national2019	neural net	dp	1	10^{-5}	2
				NaN	1
			5	NaN	1
			10	NaN	1
		non-DP	NaN	NaN	15
	query matching	dp	1	10^{-5}	1
				$3.6*10^{-6}$	1
			10	$1/n^2$, where n is the data size	1
				10^{-5}	1
	sdc	sdc	NaN	NaN	3
	stat model	dp	5	NaN	1
		non-DP	NaN	NaN	1
tx2019	neural net	dp	1	NaN	1
			5	NaN	1
			10	NaN	1
		non-DP	NaN	NaN	7
	query matching	dp	1	10^{-5}	1
			10	10^{-5}	1
	sdc	sdc	NaN	NaN	1
	stat model	dp	5	NaN	1
		non-DP	NaN	NaN	1

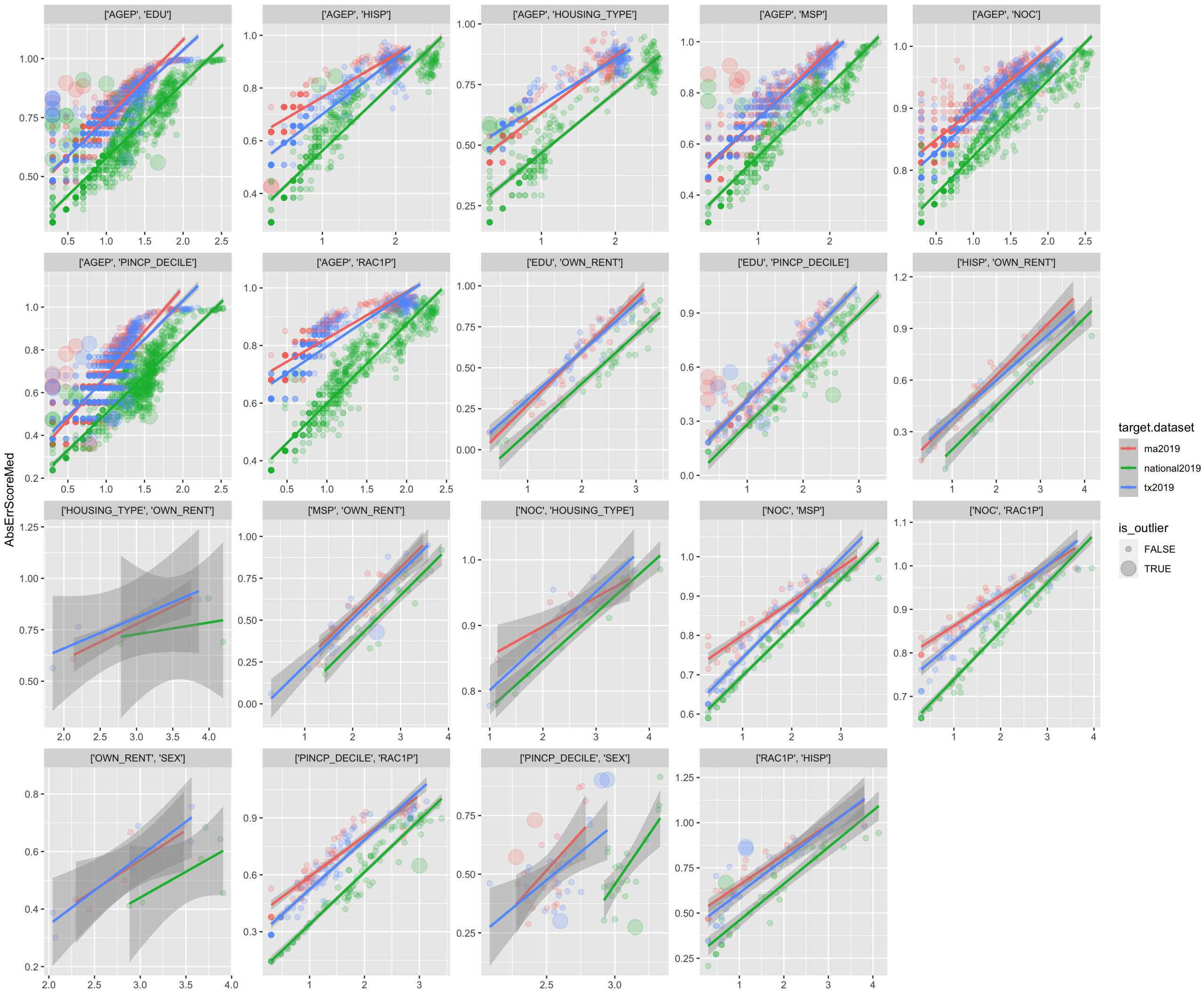

Figure 5. Scatterplots and linear model fits of $\mathrm{AbsErrScoreMed}(\cdot)$ versus $\log_{10}(x_{(\cdot)}+1)$ for all feature node pairs by dataset and feature pair, with outliers flagged. Shaded error bars correspond to a 95% prediction interval from the model in Equation 4.