Statistical Evaluation of Privacy-preserving Publication and Sharing of Three Types of COVID-19 Pandemic Data: Methods and Case Studies

Fang Liu¹, Dong Wang², Tian Yan¹

¹Department of Applied and Computational Mathematics and Statistics
University of Notre Dame, IN 46556, USA
²College of Cyberspace Security
Hangzhou Dianzi University, Wuhan, 430079, China

Abstract

A huge amount of data of various types are collected during the COVID-19 pandemic, the analysis and interpretation of which have been indispensable for curbing the spread of the coronavirus. As the pandemic moves to an endemic state, the data collected during the pandemic will continue to be rich sources for further studying and understanding the impacts of the pandemic on various aspects of our society. On the other hand, naïve release and sharing of the information can be associated with serious privacy concerns. In this study, we use three common but distinct data types collected during the pandemic (case surveillance tabular data, case location data, and contact tracing networks) to demonstrate publication and sharing of granular information or individual-level pandemic data in a privacy-preserving manner. We leverage and build upon the concept of differential privacy to generate and release privacy-preserving data for each data type. All the approaches employed in the study are straightforward to apply. We investigate the inferential utility of the privacy-preserving information through simulation studies at different levels of privacy guarantees and demonstrate the applications of approaches with examples and real-life pandemic data. Our study generates statistical evidence on the practical feasibility of sharing pandemic data with privacy guarantees and on how to balance the statistical utility of released information during this process.

keywords: COVID-19 pandemic, differential privacy, geo-indistinguishability, hot spot heat maps, contact tracing network, randomized response, synthetic data

1 Introduction

1.1 Background

As the world continues to cope with the COVID-19 pandemic and some regions are on track of slowly and cautiously going back to the pre-pandemic life, the impact brought by the pandemic

*Corresponding author: Fang Liu (fang.Liu.131@nd.edu). Fang Liu was supported by NSF Grant #1717417 and the University of Notre Dame Asia Research Collaboration Grant. Dong Wang was supported by the China Scholarships Council program (NO. 201906270230) and NSFC Grant #41971407; Tian Yang was supported by the University of Notre Dame Asia Research Collaboration Grant and the Chinese Scholarship Council,
is here to stay for a very long time. A huge amount of data of various types have been collected
during the pandemic, the analysis and interpretation of which has been indispensable to health
authorities and experts in various fields to gain understanding of the disease, to identify risk
factors, to monitor and forecast the spread of COVID, to evaluate the impacts of the pandemic
on the capacity of healthcare systems, public health, economics and employment, education,
and to plan and implement strategies to mitigate negative impacts. As the COVID-19 pandemic
shifts to an endemic state, the collected data will continue to serve as rich sources for further
studying of the disease and understanding of its impacts on societies.

On the other hand, naïve release and sharing of the information can be associated with serious
privacy concerns, especially considering the fact a huge amount and a great variety of data were
collected very quickly in a short period of time to deal with the global health crisis and the data
privacy and ethics regulations were lagging behind at least in the initial stage of the pandemic.
Many types of data collected during the pandemic are known to be associated with high privacy
risk, such as disease status, medical history, insurance status, location history, close contacts,
employment/income status, etc. Balance must be found between individual privacy protection
and sharing the data for research use.

Fortunately, this is not an unsolvable problem, due to the fact that many questions of research
interest revolve around extracting population-level and aggregate information and understanding
patterns rather than focusing on individual-level information which is often the goal of privacy attacks. If a privacy-preserving data releasing procedure can maintain accurate and useful aggregate information while guaranteeing individual-level privacy, it would make an effective approach for data sharing though further research is still needed to ensure the solutions are practically feasible and software and tools are available to facilitate practical applications.

In this study, we examine three specific types of data collected during the pandemic, develop
and apply privacy-preserving approaches for publishing or sharing each data type, and evaluate
the statistical utility of released privacy-preserving data. We hope our work sheds light on
sharing COVID-19 data and helps promote and encourage more data sharing to assist us to
better understand the disease and its current and future impact on our society.

1.2 Related Work

Privacy-preserving collection and analysis of COVID-19 data have been developed and ap-
pplied during the pandemic. Google research teams apply differential privacy (DP) to generate
anonymized metrics from the data of Google users who opted in for the Location History setting
in their Google accounts and produce the COVID-19 community mobility reports (Aktay et al.,
2020), to understand the impacts of social distancing policies on mobility and COVID-19 case
growth in the US (Wellenius et al., 2021), to generate anonymized trends in Google searches
for COVID-19 symptoms and related topics (Fabrikant et al., 2020), and to forecast COVID-19
trends using spatio-temporal graph neural networks (Kapoor et al., 2020). DP is also inte-
grated in deep learning to predict COVID-19 infections from imaging data (Müftüoğlu et al.,
2020; Ulhaq and Burmeister, 2020). Butler et al. (2020) apply DP to generate individual-level
health tokens/randomized health certificates while allowing useful aggregate risk estimates to
be calculated. The methods in all the work above are designed to provide the usefulness of
collective information in the released data without disclosing individual-level information – such
as locations, behaviors, contacts, medical status, etc upon release.
Privacy-preserving technologies and tools are also adopted in contact tracing (CT) apps and software around the world during the pandemic to track and curb the spread of coronavirus. The apps collect users’ location data (e.g., GPS) or proximity data (e.g., Bluetooth) to identify and notify those who might have been near a COVID-19 patient and at high risk of contracting the disease (Wang and Liu, 2020). On the other hand, location data and relational information can be highly revealing of personal information in general. (Coppens et al., 2014; De Montjoye et al., 2013; Dredze et al., 2016; Felmarie and Kreager, 2017; Kesler and McKenzie, 2018; Kim and Hastak, 2018; Shokri et al., 2011). To protect the sensitive information collected for CT purposes, either a centralized or decentralized data model can be used. In a centralized data model, data are collected and stored on servers owned and maintained by government or health authorities, who process the data, identify close contacts, and notify users of potential exposure. This model is adopted by Alipay Health Code (China), WeChat (China) (Ghaffary, 2020), Corona100m (South Korea) (Wray, 2020), COVIDTracker (Thailand) (5Lab, 2020), ProteGo (Poland) (Gad-Nowak and Grzelak, 2020), and the Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) (EU) (Parliament, 2020). By contrast, in a decentralized data model, only users know the potential exposure and the contact logs are never transmitted to or stored in a central server, such as in Safe paths (US) (Raskar et al., 2020) and the proximity-based Google/Apple Exposure Notification (GAEN) system (Apple and Google, 2020). The decentralized model arguably provides better privacy protection than the centralized model given data is collected, stored, processed, and shared in the former; plus, its privacy protection measures are also often encryption-based.

1.3 Our Contributions

Many privacy-preserving methods and applications adopted during the pandemic, including the work reviewed above, focus on data and information to be shared with health officials, clients or the public to facilitate either quick decision making and action taking or to notify the public in a timely manner. Privacy-preserving release of individual-level COVID-19 data has received less attention, which is the major focus of our work. Individual-level data are valuable for secondary analysis and more in-depth research researchers can perform data analysis on their own to answer specific research questions, the outcomes from which may not only have scientific merits but also produce real-world evidence of broader impacts, including but limited to better understanding of the pandemic, solving problems that were not addressed during the pandemic due to time or resource limitations, and generating new insights into how we can better handle similar crisis in the future. On the other hand, sharing data at a granular level or at individual-level level also entails privacy risks that must be mitigated in data release.

We leverage and build upon existing formal privacy concepts and techniques and apply them to several common but distinct data types collected pandemic data for publication and sharing with privacy guarantees. Specifically, we examine case surveillance data, case location data, and contact tracing networks (CTN). It is not our goal to cover all types of data collected during the pandemic, but to use these three types, which are common pandemic data types and but also different enough in terms of data structure and statistical analysis and models, to demonstrate how to apply formal privacy concepts to release privacy-preserving information. In all three data types, granular or individual-level information may not be shared due to privacy concerns. On the other hand, making the information available to the research community can help us to better understand the risk factors of COVID-19 and identify sub-populations that are vulner-
able to the disease, among others, to explore relationships between hot spots and residential characteristics to study issues such as residential racism and structural segregation during the pandemic and, to study how clustering of COVID-19 cases and how physical proximity may affect the spread of the disease, among others.

Case surveillance data are often shared as tabular data or contingency tables. The data may contain demographic, socioeconomic, or other attributes that can be used as pseudo-identifiers for re-identification or disclosure of sensitive information. We apply the flat Laplace sanitizer with DP guarantees to generate privacy-preserving case surveillance data and examine the statistical utility of log-linear models based on sanitized data, as a function of sample sizes and pre-specified privacy budgets in simulation studies and real-life data published by at the CDC.

Location data can be collected as GPS coordinates (Wang and Liu 2020) or nominal data (city, places, etc). Sharing Location data may lead to disclosure of behavioral and social activities, and other private information about an individual. We apply the planar Laplace mechanism with geo-indistinguishability guarantees to release case location data. We conduct simulation studies and apply the method to a real-life South Korean case location dataset to examine the inference from cluster point process models and the accuracy of hot spot heat maps based on sanitized location data.

CTNs are relational data and represented as graphs. Relational information is viewed by many as private information. We apply two methods to sanitize CTNs constructed from location data and from proximity data, respectively. For the former, we employ the planar Laplace mechanism to sanitize location data first and then form CTNs per the public definition of close contacts. In the latter, we examine the feasibility of a previously studied graph synthesis method – differentially private exponential random graph models – to generate privacy-preserving networks by directly perturbing the relational data in original networks. We conduct simulation studies to examine the utility of sanitized networks in inference from exponential random graph models and the preservation of descriptive structural information in networks after sanitization.

In all the data types and examples examined in this paper, released data are synthetic data generated at a pre-specified privacy budget rather than a set of statistics. With synthetic data, data users may perform analysis on their own. Readers who are interested in differentially private data synthesis may refer to Bowen and Liu (2020) for an overview on this topic.

The rest of the paper is organized as follows. Section 2 provides an overview of the basic concepts in DP, some commonly used randomized mechanisms for achieving DP, and an approach for obtaining valid inferences from sanitized data. Section 3, 4 and 5 examine case surveillance data, case location data, and CTNs, respectively and apply DP procedures to release privacy-preserving data, conduct simulation studies to examine the statistical utility of the privacy-preserving data and apply the DP procedures to real-life pandemic data. Section 6 summarize the study and provides some final remarks on the implementations of DP methods in sanitizing and releasing COVID-19 data in general.
2 Preliminaries

We provide a brief overview of the concepts, mechanism, and approaches related to DP that are used or referred to in the later sections of the paper. Specifically, we introduce the original concept of DP, a common randomized mechanism for achieving DP, extension of the DP notion to location data, the connection of randomized response with DP, as well as an approach for obtaining valid privacy-preserving inference. In each case, we present the formal definition, and provide the motivation behind the concept, and interpret what it means in layman’s terms.

2.1 Differential privacy

DP is a mathematically formulated privacy notion to provide robust privacy guarantees. It aims at protecting every individual in a dataset in the worst-case scenario. In layman’s terms, DP implies the chance that any individual in a dataset can be identified based on released sanitized results is low since the released results are about the same with or without that individual in the data.

Definition 1 ($\epsilon, \delta$)-DP. (Dwork et al., 2006a,b) A randomized algorithm $\mathcal{M}$ is of $(\epsilon, \delta)$-DP if for all dataset pairs of neighboring data sets $(D, D')$ differing by one record and for all subsets $S \subseteq \text{image}(\mathcal{M})$,

$$\Pr(\mathcal{M}(D) \in S) \leq e^\epsilon \Pr(\mathcal{M}(D') \in S) + \delta. \quad (1)$$

$D$ and $D'$ differing by one record (denoted by $d(D, D') = 1$) may refer to the case that they are of the same size but differ in at least one the attribute value in exactly one record, or the case that $D'$ has one record less than $D$ or vice versa.

$\epsilon > 0$ and $\delta \geq 0$ are privacy budget or privacy loss parameters. When $\delta = 0$, $(\epsilon, \delta)$-DP reduces to $\epsilon$ pure DP; the smaller $\epsilon$ is, the more privacy protection there is on any individual in the data, as the released results $\mathcal{M}(D)$ and $\mathcal{M}(D')$ are similar in the sense that their probability density/mass function ratio is bounded with $(e^{-\epsilon}, e^\epsilon)$. When $\epsilon \rightarrow \infty$, there is no privacy protection and all original information is preserved; on the other hand, the released outcome would become very noisy as $\epsilon \rightarrow 0$. In other words, there is always a trade-off between utility and privacy (Liu, 2020). There is no consensus and lacks a universal guideline on the choice of $\epsilon$ (Dwork et al., 2019). $\epsilon$ typically ranges from $10^{-3}$ to 10 in empirical studies in the DP literature on, depending on the type of information released, social perception of privacy, expected accuracy of released data, among others. Real-life of applications of DP often employs larger for better utility (e.g., US Census uses $\epsilon$ of 19.61 and Apple Inc. sets $\epsilon$ at 2, 4, or 8 for different Apps). When $\delta \neq 0$, it is often a very small value (inverse proportional to poly($n$)) and can be interpreted as the probability that the pure $\epsilon$-DP is violated.

Definition 1 is the basic notion of DP; relaxed versions and extensions exist, such as $(\epsilon, \delta)$-probabilistic DP (Machanavajjhala et al., 2008), $(\epsilon, \tau)$-concentrated DP (CDP) (Dwork and Rothblum, 2016), zero-concentrated DP (Bun and Steinke, 2016) (zCDP), Rényi DP (RDP) (Mironov, 2017), and Gaussian DP (GDP) (Dong et al., 2021).

DP provides a mathematically rigorous framework for protecting individual privacy when releasing and sharing information. Many mechanisms and procedures exist that achieve DP, both

1 The brief overview is not comprehensive and nowhere near covering every major concept in DP. We focus on only the concepts used or mentioned in this paper.
the basic notion as well as its various extensions. In this paper, we focus on pure \( \epsilon \)-DP and the Laplace mechanism (Section 3) to illustrate how to apply DP concepts and procedures to protect individual privacy when releasing COVID-19 data. When other types of DP guarantees are desired, corresponding mechanisms for achieving guarantees; for example, the Gaussian mechanisms \cite{Dwork2014, Liu2018} can be used to achieve \((\epsilon, \delta)\)-DP guarantees when releasing information.

**Definition 2 (Laplace mechanism).** \cite{Dwork2006} Let \( s = (s_1, \ldots, s_r) \) be a statistic calculated from a dataset. The Laplace mechanism of \( \epsilon \)-DP releases \( s^* = s + e \), where \( e \) contains \( r \) independent samples from \( \text{Laplace}(0, \Delta \epsilon^{-1}) \), where \( \Delta_1 = \max_{x, x', d(x, x') = 1} \|s(x) - s(x')\|_1 \) is the \( \ell_1 \) global sensitivity of \( s \).

The \( \ell_1 \) global sensitivity represents the maximum \( \ell_1 \) change in \( s \) between two neighboring data sets.\(^2\) The larger the sensitivity, the more impact a single individual will have on the value of \( s \) and more noise would be needed to make \( s(x) \) and \( s(x') \) similar to lower privacy risk.

Every time a dataset is queried, there is a privacy cost (loss) on the individuals in the dataset. Data curators need to track the privacy cost for all the queries and analysis on the data to ensure the overall privacy spending does not exceed the pre-specified level. Two basic composition principles in DP, **parallel composition** and **sequential composition** \cite{McSherry2007}, can be used in privacy loss accounting, which are also used in later sections of the paper.

**Definition 3 (Basic privacy loss composition).** \cite{McSherry2007} If mechanism \( M_j \) is of \((\epsilon_j, \delta_j)\)-DP and applied to a disjoint subset \( D_j \) for \( j = 1, \ldots, P \) of a dataset \( D \), the parallel composition states that applying \( M_j \) to \( D_j \) simultaneously \( \prod_j M_j(D_j) \) is \( \max(\epsilon_j, \delta_j)\)-DP; if \( M_j \) is applied to the same dataset \( D \), the sequential composition states that \( \prod_j M_j(D_j) \) is \((\sum_j \epsilon_j, \sum_j \delta_j)\)-DP.

In layman’s terms, the two privacy loss composition principle basically states as long as there is no overlapping individuals between two datasets to which two DP mechanisms is applied to, the overall loss for releasing the queries results is the maximum privacy spending between the two; otherwise, the overall loss adds up. The sequential composition can be conservative as a lower bound on the overall privacy loss. To improve the bound, advanced composition \cite{Dwork2010} for \((\epsilon, \delta)\)-DP is available. This motivates the development of various DP extensions, including those mentioned above (e.g., CDP, zCDP, RDP, GDP), which allow for tighter privacy loss composition on the privacy loss over multiple releases.

DP is a main-stream concept in privacy research and applications nowadays. Besides its mathematical rigor and robustness to various privacy attacks, it also possesses other critical properties that permits design of sophisticated DP procedures and algorithms for complicated problems (e.g., differentially Private gradient-based optimization). Among those, privacy loss composition and immunity to post-processing and future-proofing are two well-known. The former is presented above; the latter means that information released from a DP mechanism won’t leak additional information about the individuals in the dataset on which the information is based when it is further processed after the release or when there is additional information on these individuals in the future from other sources, as long as there is no access to the original dataset.

\(^2\)In general, one can define \( \ell_p \) \((p \geq 0)\) global sensitivity; see \cite{Liu2018}.
2.2 Location privacy

\cite{Andres2013} extend the pure \( \epsilon \)-DP concept to releasing privacy-preserving location data that represented as pairs of 2-dimensional GPS coordinates, along with the planar Laplace mechanism to achieve such privacy guarantees.

**Definition 4 (Geo-indistinguishability (GI)).** \cite{Andres2013} Let \( d(P, P') \) denote the Euclidean distance between any two distinct locations \( P \) and \( P' \), and \( \epsilon \) be the unit-distance privacy loss. A randomized mechanism \( \mathcal{M} \) satisfies \( \epsilon \)-GI if and only, for any \( \gamma > 0 \), any possible released location \( P^* \), and all possible pairs of \( P \) and \( P' \) that \( d(P, P') \leq \gamma \),

\[
\Pr(\mathcal{M}(P) = P^*|P) \leq e^{\epsilon \gamma} \cdot \Pr(\mathcal{M}(P') = P^*|P').
\] (2)

\( \mathcal{M} \) in Eq (2) enjoys \((\epsilon \gamma)\)-GI for any specified \( \gamma > 0 \) in the sense that the probability of distinguishing any two locations with a radius of \( \gamma \), given the released location \( P^* \), is \( e^{\epsilon \gamma} \)-fold the probability when not having \( P^* \). \( \epsilon \) is regarded as the per-unit-distance loss, and \( \gamma \) denotes the number of unit. The larger \( \epsilon \gamma \) is, the larger the privacy loss \((\epsilon \gamma)\) will be. Suppose the unit is mile. Using \( \epsilon = 0.5 \) per mile would lead to less privacy guarantees compared to using \( \epsilon = 0.1 \) as the former would lead to a higher probability of telling the true location given the perturbed released location. For a fixed \( \epsilon \), using a larger \( \gamma \) also leads to a higher probability of identifying some true location information within a radius of \( \gamma \) mile given the perturbed location information. On the other hand, the identified true location information given the larger \( \gamma \) might not be practically alarming from a privacy perspective. For example, at \( \epsilon = 0.1 \) per mile, the probability of telling the true location apart from any location within a 1-mile radius of the former increases by \( \sim 10\% \) when having \( P^* \) vs not for \( \gamma = 1 \) mile, and the probability of telling the true location apart from any location within a 10-mile radius increases by \( \sim 172\% \) when having \( P^* \) vs not for \( \gamma = 1 \) mile. Though the probability is much larger in the latter, the learned knowledge of location is not specific nor precise (e.g. knowing it is in New York City but not exactly where in the city).

**Definition 5 (planar Laplace mechanism).** \cite{Andres2013} Denote the coordinates of the observed location \( P \) in the Euclidean space by \((x, y)\). The planar Laplace mechanism generates sanitized location \( P^* \) with coordinates with \( \epsilon \)-GI,

\[
(x^*, y^*) = (x + r \cos(\theta), y + r \sin(\theta)), \quad \text{where}
\]

\[
\begin{align*}
& r \sim \text{gamma} (2, \epsilon) = r \epsilon^2 e^{-\epsilon r} \quad \text{and} \quad \theta \sim \text{uniform} \ (0, 2\pi) = 1/(2\pi).
\end{align*}
\] (4)

\( r \) in Eqs (4) is a randomly sampled distance between \( P^* \) and \( P \) and \( \theta \) is the angle of \( P \rightarrow P^* \) in the Euclidean space, and \( r \) and \( \theta \) are independent. The concepts of GI and planar Laplace mechanism are employed in Section 4.

2.3 Re-interpretation of randomized response in the context of DP

The GI is more related to local DP \cite{Duchi2013}, an extension of the original pure \( \epsilon \)-DP, than the latter per se, which is often used for releasing aggregate information rather than individual response.

**Definition 6 (\( \epsilon \)-local DP).** \cite{Duchi2013} A randomization mechanism \( \mathcal{M} \) provides \( \epsilon \) local DP if

\[
\Pr(\mathcal{M}(x) \in \Omega) \leq e^\epsilon \cdot \Pr(\mathcal{M}(x') \in \Omega).
\] (5)
for all pairs of possible data points $x$ and $x'$ from an individual and all possible output subset $\Omega$ from $\mathcal{M}$.

Local DP deals with the case where respondents/data contributors do not trust data collectors. In that sense, the notation of local DP is motivated in a similar manner as the randomized response (RR) procedure (Warner, 1965), a classical research method often used in surveys to allow respondents to answer sensitive questions while maintaining confidentiality. RR satisfies $\epsilon$-DP (Dwork and Roth, 2014; Dwork and Smith, 2010) or $\epsilon$-local DP (Duchi et al., 2013).

We use a basic RR mechanism to illustrate the connection. Before answering a sensitive question (e.g. did you ever steal? what is your sexual orientation?), the respondent flips a fair coin. If it lands on tails, then the respondent answers the question truthfully; if it lands on heads, then the respondent flips a second coin and answers “Yes” for heads and “No” for tails. This particular RR procedure corresponds to $\epsilon = \ln(3)$ (local) DP as

$$
\frac{\Pr(X' = \text{Yes} | X = \text{Yes})}{\Pr(X' = \text{Yes} | X = \text{No})} = \frac{3}{4} = e^{\ln(3)}
\quad \text{and} \quad
\frac{\Pr(X' = \text{No} | X = \text{Yes})}{\Pr(X' = \text{No} | X = \text{No})} = \frac{1}{3} = e^{-\ln(3)}
$$

We will also explore the RR mechanism in Section 5 when releasing CTN based on the work by (Karwa et al., 2017) that perturbs the relational information between two nodes with DP guarantees.

### 2.4 Privacy-preserving statistical inference

Sanitized outputs, compared to the original outputs, are subject to an extra source of variability due to the noise introduced through the randomized algorithm $R$ for achieving DP. To account for the extra source of variability for valid statistical inference, one may directly model the sanitization mechanism, which would complicate the regular inferential procedures either analytically or computationally. An alternative to incorporate the randomness of $R$ is to release multiple sanitized datasets or statistics and employ an inferential combination rule to obtain valid inference (Liu, 2022). The former approach is problem-specific and the second approach is more general and also straightforward to apply. We adopt the later approach in this paper and refer to it as multiple sanitization (MS).

Denote the number of releases by $m$. Per sequential composition, the total privacy budget would be split into $m$ portions, one for each release. $m \in [3, 5]$ is recommended (Liu, 2022). WLOS, suppose the parameter of interest is $\theta$ and its $j$-th sanitized estimate is $\hat{\beta}^{(j)}$ with estimated variance $v^{(j)}$ for $j = 1, \ldots, m$. The final inference of $\beta$, including hypothesis testing and confidence interval (CI) construction, is based on the following inferential rule.

$$
\bar{\beta} = m^{-1} \sum_{l=1}^{m} \hat{\beta}^{(l)}, \quad T = m^{-1}B + W, \quad \text{and} \quad (\theta - \bar{\beta})T^{-1/2} \sim t_{\nu=(m-1)(1+mW/B)^2}, \quad \text{where} \quad \beta = m^{-1} \sum_{l=1}^{m} (\hat{\beta}^{(l)} - \bar{\beta})^2 / (m - 1) \quad \text{(between-set variability)}
$$

$$
B = m^{-1} \sum_{l=1}^{m} v^{(l)} \quad \text{(within-set variability)}
$$

Based on the DP concepts and approaches introduced above, we discuss the privacy-preserving releasing of three pandemic data types: subgroup case surveillance data (Section 3), case location data (Section 4), and CTN data (Section 5). In each case, we describe the data type, introduce methods for sanitizing the data, conduct a simulation study to examine the impact of sanitization on statistical inferences based on the data, and apply a sanitization approach to
a real data set when one is available. All the presented sanitization approaches are either from
the literature or developed/improved by us for the corresponding data type.

3 Privacy-preserving Release of Case Surveillance Data

Case surveillance data are listing of cases, together with various attributes associated with
the cases, such as demographics, exposure histories, disease severity indicators and outcomes,
presence of any underlying medical conditions and risk behaviors, etc. Collection and publication
of COVID surveillance data are necessary during the pandemic to monitor the spread of
the disease and provide necessary information to health authorities for quick decision making,
among others. Case numbers reported at different scales (organization, city, county, state, and
national levels) by demographic groups such as age, gender, race and ethnicity provide valuable
information for identifying risk factors and groups vulnerable to the disease and understanding
the heterogeneity of the susceptibility to the disease. On the other hand, publishing such gran-
ular information may increase re-identification risk and disclosure risk, especially when data are
sparse. In this section, we focus on on publishing subgroup case number with formal privacy
guarantees.

3.1 Methods

Publishing a privacy-preserving subgroup case number dataset can be formulated as releasing
a multi-dimensional histogram or contingency table. The most straightforward approach is the
flat Laplace sanitizer, which injects noise from the Laplace mechanism directly onto each cell
count of a histogram or contingency table. In addition, there also exist many methods for
sanitizing count data that may achieve better utility for certain analyses at the cost of more
complicated implementation (Bowen et al., 2021; Eugenio and Liu, 2021; Geng and Viswanath,
2015; Hay et al., 2010; Li et al., 2018; Xiao et al., 2011, 2012; Xu et al., 2013; Yan et al.,
2018; Zhang et al., 2014).

Given there are so many methods, many of which aim at improving the utility of a certain type
of analysis and are not straightforward to implement, and our main goal is to demonstrate the
application of DP in releasing count data in general without a specific downstream analysis in
mind, we focus on the flat Laplace mechanism. That said, we also examine an existing approach
– the universal histogram (UH) approach (Hay et al., 2010) – and extend the UH approach to
the case where the total sample size of the released data is fixed and public information and
does not require sanitization (the UH-proportion or simply the UHp approach) to compare
with the flat Laplace sanitizer in simulation studies and case studies. However, our simulation
results do not show a superior performance of these two approaches over the simple Therefore,
the descriptions on the UH and UHp approaches are presented in the Supplementary Materials,
along with the simulation results and the results from the application to the real-life CDC data.

The flat sanitizer is a basic sanitization method for releasing numerical statistics in general
including count data. In our problem setting, it would inject noise to the count in each bin/cell
of the multidimensional histogram/contingency table to be released. The sanitizer is straight-
forward to apply for count data as the $l_1$ global sensitivity is 1 if the unbounded DP is used or 2
if the bounded DP is used (WLOS, we use the unbounded DP in the all the cases). Specifically,
a sanitized count in cell $i$ is $\tilde{x}_i \sim \text{Laplace}(x_i, \epsilon^{-1})$ for $i = 1, \ldots, k$ if the Laplace mechanism of
The sanitized counts or proportions from the flat sanitizer can be negative as the support of the Laplace distribution is \( \mathbb{R} \). There are two ways to deal with negative sanitized values if the data curator would prefer to release the negative values as is – to replace the negative values by 0, and to re-draw until the sanitized value is non-negative (Liu, 2019) for more discussion. For the former, additional normalization might be needed in the case of a fixed upper bound, such as when the total sample size \( n \) is fixed. Real non-negative sanitized counts can be rounded to obtain integer count data without compromising privacy due to the immunity to post-processing property.

To obtain sanitized counts for a lower-dimensional contingency table from the sanitized counts at the most granular level in the data set, one may sum over the sanitized counts from all the cells with the attributes in the lower-dimensional contingency table. Per the immunity to post-processing property, the summed counts are also privacy-preserving.

### 3.2 Simulation Study

DP sanitization perturbs the original data and would affect the statistical inference of an analysis procedure, depending on the size of the original data, the privacy loss parameters, the sensitivity of the output, and the analysis procedure itself. We use a simulation study to study on how the three approaches in the previous section for sanitizing count data affect the inferences from a log-linear model.

We simulated 1,000 data sets of counts from \( y_k \sim \text{multinomial}(n, p_i) \), where \( p_k = \lambda_k/(1 + \lambda_k) \), where \( \log(\lambda_k) = \beta_0 + \beta_1 x_{k1} + \beta_2 x_{k2} + \beta_3 x_{k3} + \beta_4 x_{k1} x_{k2} + \beta_5 x_{i1} x_{k3} + \beta_6 x_{i2} x_{k3} \) for \( k = 1, \ldots, 8 \), where \( X_1 = \{0, 1\}, X_2 = \{0, 1\}, X_3 = \{0, 1\} \) are binary attributes and \( \beta = (\beta_1, \ldots, \beta_6) \). In each repeat, we sanitize \( y = \{y\}_{k=1}^{8} \) via the flat Laplace sanitizer independently for \( m = 3 \) times to obtain differentially private \( \tilde{y}^{(j)} \), each with a privacy budget of \( \epsilon/m \), where \( \epsilon \) is the total privacy budget. We assume the total sample size \( n \) is fixed and normalize the raw sanitized counts from the flat sanitizer via \( n \tilde{y}_k^{(j)}/\sum_i \tilde{y}_i^{(j)} \), where \( \tilde{y}_k^{(j)} \) for \( k = 1, \ldots, 8 \) is the sanitized count in cell \( k \). We then run the loglinear model on \( \log(\lambda_k) = x_i^T \beta \) with \( y_k \sim \text{Poisson}(\lambda_k) \) on each set of sanitized \( \tilde{y}^{(j)} \) to obtain estimates on \( \beta \) and their estimated variance and use Eq (6) to generate the final privacy-preserving inference on \( \beta \). We examine two different sample sizes at \( n = 200 \) and \( n = 1,000 \) and four levels of privacy guarantees at \( \epsilon = 0.5, 1, 2 \) and \( 5 \). For comparison, we also run the same loglinear model on the original \( y \) and obtain the inferences from the original data.

The results are presented in Figure 1 and the main observations are summarized as follows. The smaller \( \epsilon \) is and the smaller \( n \) is, the more impact the DP procedure has on the inference; i.e., larger bias and larger root MSE (RMSE). Regardless of \( n \) or \( \epsilon \), the coverage probability (CP) of the 95% CIs is always at nominal coverage. At \( n = 1,000 \), the inference is barely affected by the DP sanitization even for \( \epsilon = 0.5 \). At \( n = 200 \), the bias and the RMSE values are noticeable for \( \epsilon = 0.5 \), acceptable at \( \epsilon = 1 \), and almost ignoble for \( \epsilon > 1 \), compared to the original inference.

We also run the UH and UHp approaches in the simulation study. The results are presented in the Supplementary Materials. In summary, the UHp approach is comparable to the flat
sanitizer in bias and RMSE for most of the parameters, but it has slight under-coverage for the latter at $\epsilon = 1$ and 0.5. The UH approach performs the worst (largest bias, RMSE, and some notable under-coverage) in this simulation study.

3.3 Application to CDC case surveillance data

We downloaded a death count dataset (dated May 24, 2022) for COVID-19 from the CDC website\(^4\). We removed the race group ‘unknown’ and collapsed age groups (0, 4] and [5, 17] to a single < 18 group, and age groups [75, 84] and $\geq 85$ to a single > 74 group. The data contains two attributes – age group and race/ethnicity; each has 7 levels, leading to a $7 \times 7$ contingency table. The data are provided in Table 1. The total sample size $n = 998,262$, which is assumed to be fixed and public.

If the data are not expected to be used for statistical inference with uncertainty quantification, the data curator may release a single sanitized data, that is, all privacy budget $\epsilon$ assigned to this dataset can be used to generate a single privacy-preserving tabular dataset to release; that is, $n_k^* = n_k + e_k$, where $e_k \sim \text{Laplace}(0, \epsilon^{-1})$, for $k = 1, \ldots, 49$ independently. The total $n = 998,262$ is assumed to be public knowledge and is fixed during sanitization, meaning that sanitized $n_k^*$ is normalized as $n_k^* \leftarrow n_k^*[n^{-1}\sum_k n_k^*]$ before release. An example release via the Laplace sanitizer at $\epsilon = 0.5$ is given in Table 2. There is some fluctuation in each cell count due to the sanitization, as expected. The column and row marginals are calculated by summing over the corresponding cell counts after sanitization and are also privacy-preserving per the immunity to post processing properties. Also noted that the sanitization error term on the marginals is

\(^4\)Table 2 at [https://www.cdc.gov/nchs/nvss/vsrr/covid19/health_disparities.htm](https://www.cdc.gov/nchs/nvss/vsrr/covid19/health_disparities.htm)
Table 1: US COVID-19 death counts by age group and race/ethnicity on May 24, 2022

| Age (ys) | NH White | NH Black | NH AIAN | NH Asian | NH NHPI | NH Mix | Hispanic | Total |
|----------|----------|----------|---------|----------|---------|--------|----------|-------|
| <17      | 387      | 274      | 15      | 36       | 11      | 30     | 303      | 1056  |
| 18-29    | 2263     | 1492     | 187     | 190      | 49      | 73     | 2015     | 6269  |
| 30-39    | 6661     | 4144     | 560     | 558      | 151     | 157    | 5919     | 18150 |
| 40-49    | 17269    | 8937     | 1021    | 1206     | 265     | 309    | 13981    | 42988 |
| 50-64    | 97418    | 35753    | 3198    | 5312     | 715     | 952    | 43657    | 187005|
| 65-74    | 141409   | 37765    | 2901    | 7423     | 501     | 913    | 38422    | 229334|
| >75      | 380630   | 54576    | 3210    | 16504    | 449     | 1380   | 56711    | 513460|
| Total    | 646037   | 142941   | 11092   | 31229    | 2141    | 3814   | 161008   | 998262|

Race/ethnicity = ‘unknown’ is not included in the table.
NH = Non-Hispanic; AIAN = American Indian or Alaska Native; NHPI = Native Hawaiian or Other Pacific Islander;
“Mix” means “more than one race”

If the data are expected for inferential purposes, the information as released in Table 2 can still be used, but data users need to model the sanitization mechanism explicitly to ensure the variability due to sanitization is taken into account. This can significantly complicate an analysis procedure, either analytically or computationally, and may require users to write code (optimization, variance estimation, etc) from scratch; in addition, this needs to be done for each analysis type.

A much more user-friendly approach is the multiple synthesis (sanitization) idea given in Sec 2.4 which is the approach we use here. We set $m = 3$ and used the flat Laplace sanitizer to sanitize $n_k$ with $\text{Laplace}(0, \epsilon/m)$ independently to obtain differentially private $n_{(j)}^*$ for $j = 1, 2, m$. For the analysis of the sanitized data, we fitted a 2-way loglinear model with covariates age group and race/ethnicity to the original and sanitized data.

As a side, UH is designed to improve the MSE of marginal counts as compared to the individual cells if the former statistics are of main interest.

**Table 2:** Flat Laplace sanitized ($\epsilon = 0.5, m = 1$) US COVID-19 death counts by age group and race/ethnicity on May 24, 2022

| Age (ys) | NH White | NH Black | NH AIAN | NH Asian | NH NHPI | NH Mix | Hispanic | Total |
|----------|----------|----------|---------|----------|---------|--------|----------|-------|
| <17      | 385      | 271      | 14      | 37       | 8       | 29     | 308      | 1052  |
| 18-29    | 2258     | 1491     | 186     | 198      | 49      | 72     | 2009     | 6263  |
| 30-39    | 6664     | 4140     | 562     | 558      | 145     | 156    | 5928     | 18153 |
| 40-49    | 17269    | 8937     | 1021    | 1202     | 266     | 299    | 13982    | 42976 |
| 50-64    | 97418    | 35753    | 3195    | 5311     | 713     | 952    | 43658    | 187003|
| 65-74    | 141409   | 37765    | 2897    | 7427     | 501     | 914    | 38425    | 229343|
| >75      | 380642   | 54577    | 3209    | 16505    | 449     | 1379   | 56712    | 513472|
| Total    | 646053   | 142941   | 11084   | 31238    | 2130    | 3801   | 161021   | 998262|

Race/ethnicity = ‘unknown’ is not included in the table.
NH = Non-Hispanic; AIAN = American Indian or Alaska Native; NHPI = Native Hawaiian or Other Pacific Islander;
“Mix” means "more than one race"
The parameter estimates from the log-linear model based on the original and sanitized counts are presented in Figure 2. There are 48 regression coefficient associated with the risk factors – 6 associated with age (<18 years is the reference group), 6 associated with race (non-Hispanic white is the reference group), and 36 parameters representing the interaction between the two. In summary, the privacy-preserving inferences based on the sanitized counts from the flat Laplace sanitizer are similar to the original inference at all $\epsilon$ values, due to the large sample size of the dataset.

![Figure 2: Privacy-preserving results from the Log-linear model on sanitized CDC COVID-19 death data via the flat Laplace sanitizer](image)

We also run the UH and UHp approaches on the data set and the results are presented in the Supplementary Materials. The details of the implementations and the results are presented in the Supplementary Materials. In summary, for both approaches, there is some discrepancy between the privacy-preserving point estimates vs the original, especially for UH, where some CIs are noticeably wider than the original mostly in the race/ethnicity groups that are relatively small in size (Non-Hispanic Asian, Non-Hispanic American Indian or Alaska Native, Non-Hispanic Native Hawaiian or Other Pacific Islander, Non-Hispanic Mix Race).

3.4 Summary

Subgroup case numbers contain granular information that permits more complicated analysis and helps us understand better the pandemic, such as quantifying the risk factors associated with COVID-19 as demonstrated in Figure 2. We demonstrate via an simulation study and an real-life application that useful privacy-preserving can be achieved, especially when $n$ is large or people are willing to sacrifice some level privacy ($\epsilon$ is not too small). The results also suggest the flat Laplace sanitizer can be an effective approach for that purposes, despite its simplicity.

4 Privacy-preserving Release of Case Location Data

When a person is diagnosed with COVID-19, health authorities may interview the person for his or her whereabouts and location history in the past few weeks (CDC, 2021; National Health Commission of China, 2021b). Collecting a patient’s location history is critical for health authorities to track and take measures to limit the spread of the disease. However, the exact location information, if shared with the public, may cause serious privacy risk for the patients and can even lead to cyber-bullying (National Health Commission of China, 2021a).

We examine a privacy-preserving approach to releasing location data based on the concept of GI. We focus on static location data at a given time point rather than dynamic location infor-
mation or travel trajectories. Upon the release of individual-level location data, researchers may conduct advanced spatial analysis such as using point process models to understand the spatial trend of the cases or conduct descriptive analysis such as COVID-19 hot spot heat map generation. This type of data can be released on a regular time basis, such as every one or several days, allowing examination of temporal trends of certain estimates or signals.

An example of the data type we work with in this example is given in Figure 3, which shows the locations of 121 COVID-19 patients on Feb 20, 2020. The number of locations per subject ranges from 1 to 11, with about 50% (62 out of 121) has one location, another 34.7% has 2 or 3 locations, and the rest 14% have \( \geq 4 \) locations with one person has 11 locations (all within the city of Gwangju). The timestamp information in hours and minutes of a day is not available in the dataset. We aim to release the geographic locations of the 121 COVID-19 cases.

Figure 3: Location information of 121 COVID-19 patients on Feb 20, 2020 in South Korea

4.1 Method

The approach we propose for releasing privacy-preserving location information is the doppelganger, based on the GI concept in Definition 4. The main idea behind doppelganger, as suggested by the name, is to release \( m \geq 1 \) sanitized versions of the true location \( P \) via the planar Laplace mechanism so to satisfy GI guarantees. The privacy budget per location \( \epsilon \) is split into \( m \) portion for \( m \geq 2 \), \( \epsilon/m \) per release, due to the parallel composition theorem in privacy loss composition. This also guarantees that the privacy loss per location is always kept at \( \epsilon \) regardless of the value of \( m \), one for each sanitization. Similar to the case surveillance data, the main reason for releasing multiple perturbed locations \( (m \geq 2) \) for a given original location is to provide the data for quantifying sanitization uncertainty or drawing statistical inferences using the MS approach presented in Section 2.4.

---

6 We will explore releasing this type of data with privacy guarantees, such as private travel trajectories (Liu et al., 2021), in the future.

7 The data can be found in file “patientroute.csv” at https://www.heywhale.com/mw/dataset/5e797e9e98d4a8002d2c92d3/file

8 Unequal allocation schemes can also be used, as long as the portions add up to \( \epsilon \), though it does not seem necessary in this case.
Specifically, to generate a sanitized location \((x^*, y^*)\) given the original location coordinates \((x, y)\), we apply the planar Laplace mechanism in Eq (3), where \(\epsilon\) is replaced by \(\epsilon/m\). Also noted is that \(\epsilon\) is the per-unit-distance privacy loss, where the unit distance is supplied by the data curator and can be a value deemed appropriate for the task at hand such as 1 meter, 10 meters, 0.5 miles, the choice of which depends on the location types, the area, among other considerations. Figure 4 shows a hypothetical example on how doppelganger works in the area of a local plaza in South Bend, Indiana. The case’s true location (red star) in the left plot is the Martin’s Supermarket, which is not released, with \(m = 5\) released sanitized locations generated with GI privacy guarantees of \(\epsilon = 0.1\) per 10 meters.

**4.2 Simulation Study**

To evaluate statistical utility of sanitized locations via doppelganger, we conducted a simulation study. We simulated 1,000 sets of location data in a square area of \([0, 1] \times [0, 1]\) from an inhomogeneous Matérn cluster point process with the radius the clusters \(= 0.03\) and the non-stationary log-density
\[
\log(\lambda(x, y; \boldsymbol{\beta})) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 y^2 + \beta_5 x y,
\]
where \(\boldsymbol{\beta} = (\beta_0, \ldots, \beta_5) = (4.53, 3.30, 3.43, -0.27, 1.58, 2.24)\) and \(x\) and \(y\) are coordinates. The number of locations ranges from 769 to 1217 across the 1,000 repeats with an average of 970. In each of the 1,000 simulated location sets, we sanitized each location with the planar Laplace mechanism in Eq (3) at \(\epsilon = 5, 2, 1, 0.5\) per 0.01 unit, respectively, and \(m = 3\). We then fitted the inhomogeneous Matérn cluster point process model above to estimate \(\boldsymbol{\beta}\) and applied the inferential rule in Eq 6 to calculate the estimation bias, RMSE, and CP of 95% CIs. The data simulation and analysis were conducted using R package spatstat.core (Baddeley and Turner 2005).

The results are presented in Table 3. In summary, the inferences at \(\epsilon = 5\) and 2 are largely comparable with the original inference – close-to-0 bias, similar RMSE as the original, nominal converge at \(\epsilon = 5\) and slight under-coverage at \(\epsilon = 2\). At \(\epsilon = 1\) and 0.5, the estimation bias is notably much larger; the RMSE values are similar to the original at \(\epsilon = 1\), but much larger at \(\epsilon = 0.5\); the CP is around 83% to 85% at \(\epsilon = 1\) and ranges from 60% to 88% at \(\epsilon = 0.5\). The moderate to severe under-coverage is largely due to the large bias in the \(\boldsymbol{\beta}\) estimates, which in turn may be attributed to the bounding applied to the sanitized locations. Specifically, we
know all locations are bounded with the square of $[0, 1] \times [0, 1]$ (public information) whereas the planar Laplace mechanism, in theory, can generate an infinite $r$ and any angle $\in [0, 2\pi]$. To honor the location boundaries, we a sanitized $x^* < 0$ at 0 and at 1 if it is > 1; similarly in the case of sanitized $y^*$. Bounding sanitized values can lead to biased inference (Liu, 2019).

Table 3: Privacy-preserving inferences of the Matérn cluster point process model on simulated location data (1,000 repeats, $m = 3$)

| metric | parameter | original | $\epsilon = 5$ | $\epsilon = 2$ | $\epsilon = 1$ | $\epsilon = 0.5$ |
|--------|-----------|----------|---------------|---------------|---------------|---------------|
|        | $\beta_0$ | -0.029   | -0.022        | 0.016         | 0.142         | 0.571         |
|        | $\beta_1$ | 0.065    | 0.052         | -0.022        | -0.279        | -1.180        |
| bias   | $\beta_2$ | 0.031    | 0.014         | -0.074        | -0.374        | -1.389        |
|        | $\beta_3$ | -0.085   | -0.077        | -0.028        | 0.154         | 0.801         |
|        | $\beta_4$ | 0.034    | 0.038         | 0.060         | 0.124         | 0.337         |
|        | $\beta_5$ | -0.037   | -0.024        | 0.048         | 0.303         | 1.160         |
|        | $\beta_0$ | 0.466    | 0.465         | 0.459         | 0.457         | 0.680         |
|        | $\beta_1$ | 1.234    | 1.232         | 1.211         | 1.189         | 1.549         |
| RMSE   | $\beta_2$ | 1.164    | 1.162         | 1.152         | 1.166         | 1.693         |
|        | $\beta_3$ | 1.006    | 1.003         | 0.986         | 0.958         | 1.159         |
|        | $\beta_4$ | 0.944    | 0.943         | 0.934         | 0.898         | 0.838         |
|        | $\beta_5$ | 0.985    | 0.982         | 0.972         | 0.989         | 1.431         |
|        | $\beta_0$ | 0.948    | 0.940         | 0.925         | 0.841         | 0.599         |
|        | $\beta_1$ | 0.938    | 0.932         | 0.914         | 0.845         | 0.719         |
| CP     | $\beta_2$ | 0.957    | 0.952         | 0.935         | 0.851         | 0.640         |
|        | $\beta_3$ | 0.938    | 0.929         | 0.909         | 0.842         | 0.769         |
|        | $\beta_4$ | 0.941    | 0.934         | 0.908         | 0.840         | 0.878         |
|        | $\beta_5$ | 0.947    | 0.939         | 0.916         | 0.827         | 0.638         |

4.3 Application to South Korea case location data

We apply the doppelganger to the real-life South Korean case location dataset (Figure 3(a)) to release privacy-preserving sanitized locations at $\epsilon = 5, 2, 1, 0.1$ per 2 miles per individual. Since some individuals have more than one location record, we further divided $\epsilon$ by the number of locations for each individual. That is, if an individual has $h$ original location data points and we release $m$ sanitized locations per original location, then each original location is sanitized with a privacy budget of $\epsilon/(mh)$. Similar to the simulation studies, we honored the fact that all cases are in South Korea and bounded sanitized location with a rectangular approximation of South Korea.

We used the sanitized locations for two analyses: to generate hot spot heat maps and to fit a point process model. The former does not involve statistical inference and we examined two settings of $m$: $m = 1$ and $m = 3$. We used $m = 3$ in the latter case. The privacy-preserving heat maps are displayed in Figure 5 with the same smoothing bandwidth as in Figure 3(b).

The hot spot heat maps based on the sanitized locations are very similar to the original heat map in Figure 3(b) for $\epsilon \geq 1$ for both $m = 1$ and $m = 3$ and are a bit noisy at $\epsilon = 0.5$ especially when $m = 3$; but the major hot spots (the cities of Busan, Seoul, and Daegu) are preserved at $\epsilon = 0.5$ for $m = 1$. In summary, for the purposes to generate heat maps, $m = 1$ is sufficient and each sanitized location is less noisy compared to using $m = 3$ especially for small $\epsilon$.

We fitted an inhomogeneous Matérn cluster point process model with log-density $\log(\lambda(x, y; \beta)) =$
\[ \beta_0 + \beta_1 x + \beta_2 y \]

on the original data and MS data with \( m = 3 \) (for this analysis, we randomly selected one location if an individual has multiple location records, leading to 121 location data points out of 121 individuals). We applied the inferential rule in Eq (6) to obtain the point estimates and 95% CIs for \( \beta \). The results are presented in Table 4. In general, the privacy-preserving inferences are similar to the original, especially for \( \beta_1 \) and \( \beta_2 \) that quantify the linear trends of COVID case intensity along the \( x \) and \( y \) coordinates, respectively. In addition, the privacy-preserving point estimates of the model parameters are robust to \( \epsilon \geq 1 \) and some notable deviation from the original only appears at \( \epsilon = 0.5 \). A surprising observation is the shrinkage in the CIs as \( \epsilon \) decreases for \( \epsilon < 5 \), implying the inferences become more precise, at least for the range of the examined \( \epsilon \) values, though the statistical insignificance remains unchanged across \( \epsilon \). The shrinkage in CIs is counter-intuitive as one would expect the inferences get less precise as the locations are perturbed more at smaller \( \epsilon \). We are not certain about the reasons for this, partly due to the difficulty in finding the cause analytically. We used R package \texttt{spatstat.core} to fit the model, which employs the minimum contrast estimation approach (Waagepetersen, 2007). The approach obtains approximate inferences based on an asymptotic Gaussian distribution assumption with a complicated expression on the variance-covariance of parameter estimates, making it difficult to assess how the sanitization process affects the variance analytically. We plan to design and run more empirical studies to understand better this phenomenon in the future.

Table 4: Privacy-Preserving Matérn cluster point process parameter estimates based multiple sets of sanitized locations in the South Korea case location data (\( m = 3 \), \( \epsilon \) per 2 miles).

| parameter | original estimate (95% CI) | \( \epsilon = 5 \) | \( \epsilon = 2 \) | \( \epsilon = 1 \) | \( \epsilon = 0.5 \) |
|-----------|-----------------------------|-----------------|-----------------|-----------------|-----------------|
| \( \beta_0 \) | -64.2 (-153.5, 25.1) | -65.1 (-157.0, 26.9) | -63.0 (-147.0, 21.0) | -63.8 (-140.6, 13.0) | -57.5 (-129.8, 14.7) |
| \( \beta_1 \) | 0.51 (-0.17, 1.19) | 0.52 (-0.18, 1.21) | 0.50 (-0.14, 1.14) | 0.50 (-0.08, 1.08) | 0.44 (-0.10, 0.99) |
| \( \beta_2 \) | 0.03 (-0.50, 0.56) | 0.03 (-0.52, 0.59) | 0.03 (-0.48, 0.54) | 0.05 (-0.42, 0.51) | 0.07 (-0.39, 0.53) |
4.4 Summary

The doppelganger method aims at privacy-preserving case location data sharing. The simulation study and the real-life application suggest the method can preserve in the released data important statistical signals in the original data at a relatively low-level cost of privacy. The method would be particularly useful for protecting location privacy when sharing information at the local level and releasing hot spot maps on a relatively fine scale. The finer the scale is, the more sparse the data become, the higher the privacy risk for re-identification from releasing location data, and thus the greater the need for effective privacy protection approaches, but also the noisier the released sanitized locations. As the scale gets coarser, say at the city, regional, state, or national levels, the information released by the doppelganger would deviate little from the actual location information.

5 Privacy-preserving Sharing of Contact Tracing Networks

Contact tracing is considered an effective approach for curbing the spread of COVID-19 during the pandemic. Contact tracing can be carried out manually by human contact tracers or digitally via GPS or Bluetooth devices. CTNs can be regarded as a type of social networks with individuals being the nodes and an edge between two people representing they are close contacts (e.g., within 6 feet of each other for a cumulative total of 15 minutes or more over a 24-hour period). CTNs are of research interest since they provide valuable information for us to better understand scientifically how physical proximity affects the spread of the disease and human contact behaviors during the pandemic, among others. However, directly sharing CTNs has privacy concerns. Adversaries may link a CTN with other databases or use background knowledge to infer who were infected with COVID-19 and tell who were close physically (appearing in the same place at the same time) based on the edge information in a CTN.

We examine a couple of approaches for sharing CTNs in a privacy-preserving manner. CTNs can be constructed from various data sources and in various ways and we focus on CTNs constructed for a pre-defined study population (e.g., COVID-19 positive employees in an organization, COVID-19 positive students in a school). And the CTNs include all close contacts that occurred during a pre-specified period time (e.g., one day, 2 weeks, 1 month). People are mobile and may show up in different places at different times. Suppose the time period is one day, starting at noon on June 1, 2020 ending at noon next day and our study population is all students at a college. If a COVID-positive student was in a dining hall from noon to 1pm on June 1, 2020 and had 2 close contacts and was at library from 1:30pm to 5pm and had 1 close contacts, and was in his/her dorm from 5pm to noon next day and had 5 close contacts, then all his/her 8 close contacts would be included in the CTN of the study population. Finally, we consider CTNs with only relational information. Releasing CTNs with nodal attributes (such as demographic information or location information) is a topic for future research.9

5.1 Methods

We examine a few approaches for releasing privacy-preserving CTNs, depending on whether the raw collected information is location data or proximity data. We present two approaches in the main text and two more in the supplementary materials.

---

9If the nodal information to be released is location information, the GI concept may be leveraged for privacy preserving releasing of both location and close contact information.
The first is applicable when collected information is \((x, y, t, \Delta_t)\), where \((x, y)\) is the 2D location information and \((t, \Delta_t)\) records what time and how many the individual is at location \((x, y)\). To decide whether there is close contact between two individuals \((i, j)\), we calculate the Euclidean distance \(d_{ij}\) between the two given the location information at each matched time point the apply the close contact definition (e.g., ≤ 6 feet for a cumulative 15 minutes over a 24-hour period) to construct CTNs graphs. Specifically,

\[
    d_{ij,t} = d_{ji,t} = \sqrt{(x_{i,t} - x_{j,t})^2 + (y_{i,t} - y_{j,t})^2},
\]

\[
    e_{ij} = e_{ji} = 1 \text{ if } \sum_t \left( \min\{\Delta_{i,t}, \Delta_{j,t}\} \cdot \mathbb{1}(d_{ij,t} \leq \tau_d) \right) \geq \tau_t; \quad e_{ij} = e_{ji} = 0 \text{ o.w.},
\]

where \(\tau_d\) and \(\tau_t\) are the cutoff on the distance and cumulative contact time and are public information, \(e_{ij} = e_{ji}\) are entries \([i, j]\) and \([j, i]\) in the CTN from location-time data, and \(e_{ij} = e_{ji} = 1\) means there is close contact between \(i\) and \(j\). Figure 6 displays an example of a location-based CTN with 100 individuals.

![Figure 6: Example of location-based CNT construction](image)

For privacy-preserving CTN construction from location, we first sanitize the location information \((x_i, y_i)\) and \((x_j, y_j)\) and the time duration information \(\Delta_{i,t}\) and \(\Delta_{j,t}\); that is,

\[
    (x^*_i, y^*_i) = (x_i + r_i \cos(\theta_i), y_i + r_i \sin(\theta_i)), \quad (x^*_j, y^*_j) = (x_j + r_j \cos(\theta_j), y_j + r_j \sin(\theta_j))
\]

\[
    \Delta^*_{i,t} = \Delta_{t,i} + \text{Lap}(0, \tau_t/\epsilon_i), \quad \Delta^*_{j,t} = \Delta_{t,j} + \text{Lap}(0, \tau_t/\epsilon_i)
\]

where \(r \sim \text{gamma}(2, \epsilon_d), \theta \sim \text{uniform}(0, 2\pi)\) and \(\epsilon_i + \epsilon_d = \epsilon\) (the total privacy budget), and then substitute \((x^*_i, y^*_i, \Delta^*_{i,t})\) and \((x^*_j, y^*_j, \Delta^*_{j,t})\) for \((x_i, y_i, \Delta_{i,t})\) and \((x_j, y_j, \Delta_{j,t})\), respectively in Eqs (7) and (8) to generate privacy-preserving CTNs. Note that the Laplace mechanism for sanitizing time duration \(\Delta_{t,i}\) and \(\Delta_{t,j}\) only applies in the setting of bounded DP. In addition, we set the \(l_1\) sensitivity of \(\Delta_t\) at \(\tau_t\) as \(\Delta_t\) values that are > \(\tau_t\) are treated the same in CTN constructions. When the nodes are independent of each other, the overall privacy cost for releasing a sanitized CNS is the same as the per-node privacy cost per the parallel compositability.

The second approach is differentially private network synthesis via Bayesian exponential random graph models (DP-ERGM) [Liu et al. 2022]. DP-ERGM perturb network data directly and can be applied when the collected raw information is either location data or binary relational data (i.e. proximity data – within 6 feet or not). In the former, Eqns (7) and (8) are applied first to construct a CTN prior to the implementation of the procedures. The DP-ERGM procedure can be regarded as an application of the model-based differentially private synthesis (MODIPS) approach [Liu 2022] to graph data with ERGMs as the synthesis model. ERGMs are a family
of popular statistical models for analyzing network data (Robins et al., 2007; Snijders et al., 2006). Denote by $e$ the adjacency matrix in a network ($e_{ij} = 1$ if an edge exists between node $i$ and node $j$, $e_{ij} = 0$ otherwise). ERGMs model the conditional distribution of $e$ as

$$p(e|\theta) = \frac{\exp\{\theta^T S(e)\}}{K(\theta)} \text{ with } K(\theta) = \sum_{e'} \exp\{\theta^T S(e')\},$$  \hspace{1cm} (11)

where $S(e)$ is the summary statistics that characterize the network structure such as number of edges, degree distribution, edge-wise shared partnership, etc. $^{10}$ $K(\theta)$ in Eqn (11) is the normalizing constant summed over all possible adjacency matrix $e'$ and is often analytically intractable unless in small networks. Inference of $\theta$ is often based on approaches with approximate $K(\theta)$, such as the Monte Carlo maximum likelihood estimation (Geyer and Thompson, 1992; Hunter and Handcock, 2006).

The steps of a general DP-ERGM procedure are as follows. Given a ERGM (either specified prior to the access to the original data or chosen using a privacy-preserving procedure by costing a portion of the total privacy budget using the original data information), 1) derive the posterior distribution $\theta$ given the likelihood function in Eq (11) and a prior for $\theta$; 2) obtain a sanitized sample $\theta^*$ from the posterior distribution with a pre-specified privacy budget; 3) simulate a network $e^*$ via the ERGM parameterized by $\theta^*$. If multiple sanitized networks are to be released, the above steps are repeated for $m > 1$ times.

In addition to the GI-based method and DP-ERGM, we also examined the RR mechanism for perturbing edge information in graphs with DP guarantees (Karwa et al., 2017) and a debiased version of the RR mechanism (Liu et al., 2022). Both procedures perform significantly worse than the GI-based and DP-ERGM procedures in the statistical analysis we examine in this study, unless the privacy loss is high $\epsilon > 5$. We briefly introduce the RR mechanism and the RR-debias method below and some of the cons; the details on these techniques can be found in the Supplementary Materials.

The RR mechanism we used flips a relation (edge to no-edge or vice versa) in the network with probability $1/(1 + e^\epsilon)$ to satisfy $\epsilon$ edge DP per edge. Given $\epsilon > 0$, the probability of flipping an edge is always $< 0.5$. This seems a very nice property; but if there is an imbalance between the number of edges vs. null edges, the sanitized relational information can still deviate from the original at small $\epsilon$. For example, there are 105 relations for a network with $n = 15$. Suppose $\epsilon = 0.5$ and the observed network has 5 edges. The expected number of edges in a RR-perturbed network is $5 \cdot e^{0.5}/(1 + e^{0.5}) + (105 - 5)/(1 + e^{0.5}) = 41$; in other words, the density of the network jumps from 5.5% to 39% after sanitization. RR in general is known for the bias generated in its implementation and needs correction when drawing inference. In the context of network analysis, the RR mechanism is modelled explicitly and can be quite complicated both analytically and computationally (Karwa et al., 2017). However, for visual display of a RR-sanitized network or just summary statistics, there is impossible to “explicitly modelling” the RR mechanism. In other words, RR-sanitized networks are not suitable for descriptive analysis directly.  

Liu et al. (2022) employs a debiasing approach as an attempt to remove bias in sanitized  

---

$^{10}$Eq (11) is a simplified ERGM as we deal with CTN without nodal attributes in this study. In general, $S$ may contain statistics not only constructed from $e$) but also nodal statistics and metrics for networks with nodal attributes.
networks via the RR mechanism (with edges $e_{ij}^*$) by synthesizing new networks with edges $\tilde{e}_{ij}^*$ given a RR-sanitized network. The debiasing procedure does not use the information from the original network and thus maintains the privacy guarantees, but at the cost of introducing another layer of variability. A “debiased” sanitized network is made of edges drawn from two Bernoulli distributions, depending on whether the synthetic edge $Y'$ from the DWRR is 1 or 0 and the flipping probability used in the RR mechanism. Our study results suggest can mitigate the bias to some degree but cannot completely remove the bias from the RR mechanism.

Note that for the GI-based approach, RR mechanism, and RR-debiased, we can match the nodes in a sanitized CTN with those in the original CTN one by one, but not for DP-ERGM, which samples a whole network from a differentially private ERGM model for release.

### 5.2 Simulation Studies

To evaluate statistical utility of sanitized CTNs, we conducted simulation studies. The data were simulated based on a real-life CTN and are similar to the latter in the distribution of close contacts per individual. The real data are not shareable due to privacy and IRB reasons. For the same reason, there is no real-life application for this data type. We examined all four DP procedures – GI-based approach, DP-ERGM, the RR mechanism, and RR-debiased – in the simulation studies. We summarize the results from the latter two at the end of this section and present their implementation details and detailed results in the Supplementary Materials.

We focus on preservation of qualitative information and descriptive statistics in sanitized CTNs relative to original CTNs first. To be able to run both the location-based and proximity-based privacy-preserving CTN construction, we simulated the coordinates of 100 locations and then applied Eqs (7) and (8) to construct a CTN. For simplicity, we examine the setting where each individual has a single location record and assume the time duration at each location is $> \tau_t$ and is public knowledge. Therefore, we only need to sanitize the location information via Eq (9) and then apply Eqs (7) and (8) on sanitized $(x^*, y^*)$ to generate a privacy-preserving CTN in the GI-based approach with $\epsilon$-GI guarantees. RR, RR-debias, DP-ERGM operate on CTNs directly. For DP-ERGM, we fitted a simple ERGM in Eq (11) with only one covariate – the number of edges $11$ using R package `statnet` (Handcock et al., 2008). The next step is to draw a privacy-preserving posterior sample from $p(\beta | e$ and we used an approach in Liu (2022). We also sanitized the statistic in the ERGM – edge count – via the Laplace mechanism with a sensitivity of 1 under edge DP (removing one edge change the edge count by at most 1). We split the total $\epsilon$ in a 1:1 ratio between the two sanitizations. Given the privacy-preserving sample of $\beta$, we generated a privacy-preserving CTN under the constraints that its edge count equals to the sanitized edge count of the original network. We set $m = 1$ and examined $\epsilon = 5, 2, 1, 0.5$ for all approaches.

The sanitized CTNs via the GI-based method and DP-ERGM are presented in Figure 7. The density of the sanitization CTNs via the GI-based and DP-ERGM approaches is similar to the original CTN at all the examined $\epsilon$ values.

The visualization of the sanitized networks provides a big-picture comparison between the sanitized vs the original networks in terms of density, clustering, etc. That said, since the nodes in sanitized networks via the GI-based, RR, and RR-debiased approaches match the

---

$^{11}$ We also tried to fit more complicated ERGMs such as those with triangle count, DD, and ESPD as covariates but those models had trouble converging on simulated networks, due to high collinearity with edge count.
original nodes, we can also examine how original relations among the node pairs change after sanitization. Interested readers may refer to Figure S9 in the Supplementary Materials for the comparison among the methods in sensitivity (probability of retaining an edge and equals \( e^\epsilon/(1 + e^\epsilon) \) for RR), specificity (probability of retaining no-edge and equals \( e^\epsilon/(1 + e^\epsilon) \) for RR), and the false discovery rate (FDR, proportion of original edges among all the edges in a sanitized CTN). However, it should be kept in mind that privacy-preserving techniques are to preserve aggregate information or sample- or population-level signals, measures such as high specificity, high sensitivity, or low FDR on the original data are not required and in some sense should not be promoted as they focus on preserving accurate information at the individual level in the original data, which is exactly what privacy protection methods aim not to disclose.

In addition to the visualization in Figure 7, we also calculated some common summary statistics for networks, including edge count, triangle count, degree distribution (DD, alos the number of close contacts per individual in CTNs), and edgewise shared partners distribution (ESPD). For a network with \( n \) nodes, its DD consists of \( d_k \) for \( k = 0, \ldots, n-1 \), where \( d_k \) is the number of nodes that share an edge with exactly \( k \) other nodes; its ESPD consists of \( \text{esp}_k/\# \) of edges for \( k = 1, \ldots, \leq n(n-1)/2 \), where is \( \text{esp}_k \) is the \# of edges whose two nodes are both connected with exactly \( k \) other nodes than themselves. The betweenness centrality measures the centrality of a node in a graph and is defined for node \( i \) as the proportion of the shortest paths that connect nodes \( j \) and \( j' \) while passing through node \( i \) (\( j \neq j' \neq i \)) among all shortest paths that connect nodes \( j \) and \( j' \). The closeness centrality is another measure of the centrality in a network. There are multiple definitions of closeness and we use \( \left( \frac{A_i}{n-1} \right)^2/C_i \), where \( A_i \) is the number of reachable nodes from node \( i \), and \( C_i \) is the sum of distances from node \( i \) to all reachable nodes. If no nodes are connected with node \( i \), then its closeness centrality is 0.

Table 5 presents the edge and triangle counts of the sanitized CTNs. The results are consistent with the observations in Figure 7. The two statistics of the sanitized CTNs via both DP procedures are similar to the original and there is not much of a difference across \( \epsilon \).
Table 5: Average (SD) edge and triangle counts in sanitized CTNs via the GI-based method and DP-ERGM (100 sanitizations)

| $\epsilon$ | GI-based | | | DP-ERGM | | |
|---|---|---|---|---|---|---|
|   | # of edges | # of triangle | # of edges | # of triangle |   |   |
| 0.5 | 34 (5.7) | 6 (3.5) | 38 (6.4) | 13 (9.2) |   |   |
| 1   | 36 (6.7) | 5 (3.8) | 39 (3.1) | 12 (7.4) |   |   |
| 2   | 37 (4.1) | 8 (3.7) | 39 (1.4) | 11 (6.8) |   |   |
| 5   | 37 (2.9) | 9 (2.9) | 39 (0.7) | 11 (7.1) |   |   |

the original edge count is 39 and triangle count is 10.

Figure 8 shows the DD, which is the distribution of close contacts of an individual in a CTN, and Figures 9 depicts the ESPD of the sanitized CTNs with the total variance distance (TVD) in DD between the sanitized and original CTNs. Consistent with the observations in Figure 7 and Table 5, the GI-based sanitization and DP-ERGM produce similar DD and ESPD to the original distributions and the statistics are stable across $\epsilon$.

Figure 10 shows the box plots of the betweenness centrality and closeness centrality of the 100 nodes in the sanitized CTNs vs the original. The distributions of both centrality measures obtained via the GI-based sanitization and DP-ERGM are similar to the original distributions at all the examined $\epsilon$ values.

As expected, RR generates dense networks for $\epsilon \leq 2$ and relatively dense even at $\epsilon = 5$ (Figure S5 in the Supplementary Materials). RR-debiased help to reduce the density level but is still worse than the GI-based and DP-ERGM procedures. Correspondingly, sanitized networks via the RR mechanism have much higher edge and triangle counts than the original and only approach the latter at a very high privacy cost ($\epsilon > 8$; Table S2 in the Supplementary Materials). DD, ESPD, betweenness centrality, and closeness centrality calculated from the CTNs generated via the RR approach also deviate significantly from the original distributions, and only at $\epsilon = 5$ do the distributions start to approach the original (Figures S7 and S8 in the Supplementary Materials). RR-debias improves over RR and its sanitized CTNs are more...
similar to the original, but still worse than the GI-based and DP-ERGM procedures.

To examine the inferential utility of sanitized CTNs, we compare statistical inferences from an ERGM fitted to sanitized networks vs the original. Toward that end, we simulated a data set of \( n = 100 \) locations and applied Eq (7) and (8) to generate the corresponding CTN. We also saved the distances \( \leq \tau_d \) in set \( D_1 \) and the distance \( > \tau_d \) than the threshold in \( D_0 \). We fitted the ERGM in Eq (11) with edge count being the sole covariate and simulated 500 networks from the fitted ERGM. In each simulated network, if there was an edge between two nodes, we sampled a distance from \( D_0 \); otherwise, we sampled a distance from \( D_1 \). This generates a \( n \times n \) matrix of distances, based on which we applied the classical multidimensional scaling to generate a set of 100 best-fitting 2-dimensional coordinates. After obtaining the locations, we construct the original CTN based on the original locations using Eqs (7) and (8). Given the location information and the original CTNs, we followed similar steps as above to generate privacy-preserving CTNs in each approach for \( m = 3 \) for \( \epsilon = 5, 2, 1, 0.5 \) for the GI-based approach and DP-ERGM and \( \epsilon = 24, 18, 15, 5, 2, 1, 0.5 \) for RR and RR-debias. Finally, we fitted the ERGM to the sanitized CTNs with edge count as the covariate to obtain privacy-preserving inference on the corresponding \( \beta \) via the inferential rule in Eq (6) for each approach.

\[ \text{To ensure the CTN generated from these calculated locations is similar to the original CTN in DD, we used Monte Carlo to determine a cutoff on the calculated distance, which turned out to similar, though not exactly the same, to the cutoff used used obtaining the original CTN.} \]
The simulation results are presented in Table 6 for the GI-based approach and DP-ERGM. In summary, both methods are acceptable, especially in CP, for all examined $\epsilon$ and DP-ERGM performs better. The GI-based approach has close-to-0 and smaller bias than DP-ERGM at $\epsilon = 5$ and 2 with smaller RMSE and $\sim 99\%$ coverage. However, it has larger biases for $\epsilon \leq 1$ but still acceptable (5.8% bias relative to the true value 4.89). The inference via DP-ERGM is relatively robust to the examined $\epsilon$ range.

| method  | metric | original† | $\epsilon = 5$ | $\epsilon = 2$ | $\epsilon = 1$ | $\epsilon = 0.5$ |
|---------|--------|-----------|----------------|----------------|----------------|----------------|
| GI-based| bias   | 0.005     | -0.017         | -0.069         | -0.284         |
|         | RMSE   | 0.169     | 0.143          | 0.123          | 0.129          | 0.305          |
|         | CP     | 0.944     | 0.990          | 0.990          | 0.998          | 0.958          |
| DR-ERGM | bias   | -0.021    | -0.026         | -0.031         | -0.051         |
|         | RMSE   | 0.171     | 0.172          | 0.174          | 0.187          | 0.260          |
|         | CP     | 0.942     | 0.954          | 0.954          | 0.952          | 0.944          |

† The original inferences are numerically different for the GI-based method and DP-ERGM because the former is based on CTNs constructed from simulated location data and the latter is based on CTNs simulated from an ERGM model.

The privacy-preserving inferences from the CTNs sanitized by RR and RR-debias are presented in Table S3 in the Supplementary Materials. Since the bias is very and CP is 0 for $\epsilon < 5$, we examined the inference for $\epsilon \in [5, 24]$. RR-debias helps decreasing the bias in the parameter estimate by the RR mechanism, but still suffer from large bias and severe under-coverage for $\epsilon$ as large at 15. When $\epsilon = 24$, RR and RR-debiased are similar in the privacy-preserving inference, but still are as good as the GI-based and ERGM methods at $\epsilon = 0.5$. We expect directly modelling the RR mechanism when obtaining inference from the ERGM model; interested readers may refer to [Karwa et al. (2017)] for the details of the modelling and implementation that is based on Markov Chain Monte Carlo methods.

### 5.3 Summary

Both the GI-based sanitization and the DP-ERGM approach can produce privacy-preserving CTNs that are structurally similar to original CTNs, by various statistical measures in the simulation study. In addition, the utility of the sanitized CTNs is relatively insensitive to $\epsilon$ for the examined range of $[0.5, 5]$, implying that a small $\epsilon$ can be used to provide strong privacy guarantees without sacrificing much of the utility. The sanitized CTNs can be shared with researchers who are interested in learning more about CTNs during the pandemic, without compromising individual privacy at a pre-specified privacy cost. In contrast, the RR mechanism by itself does not generate useful CTNs due to its inherent bias unless the privacy cost is high or CTNs are balanced in its the edge vs no-edge proportions, both of which are not realistic scenarios. The bias can be corrected when running model-based analysis by directly modelling the RR mechanism in inferential procedures or can mitigated, but not removed, by using the RR-debiased procedure. Descriptive statistics and analysis on RR mechanism generated CTNs would lead to misleading results.
6 Discussion

We use three types of data collected during the COVID-19 pandemic to demonstrate how to release and share privacy-preserving data and information and conducted simulation studies and real data applications to examine the statistical utility of the released information. Each data type (surveillance case numbers, case location information, and contact tracing networks) is common during the pandemic and provides critical information to health authorities and policy makers for to make decisions and take actions.

In each case, we apply randomized mechanisms and approaches with formal privacy notions, such as DP and GI, to sanitize and release information aiming at preservation of statistical utility and aggregate information that can be used to infer underlying population parameters, as shown in our simulation studies and real-life application. The approaches do not target at learning individual-level information, which not only conflicts with the goal of privacy protection, but is also unnecessary for the purposes of mining and understanding the population-level information.

DP and its various extensions are state-of-the-art concepts in privacy research and are quickly adopted in practice. Some of the methods we have demonstrated in the study are basic and have been applied commonly for privacy protection, such as the flat sanitizer, some are proposed recently, such as DP-ERGM methods, and others are extensions and proposed by us to deal with a specific COVID-19 data type, such as the GI-based CTN constructions. Our simulation studies suggest that different DP methods may have different utility and vary in their easiness for implementation complicated for a given statistical analysis procedure at the same privacy costs. This is also well documented in the literature – one of the reasons why new DP methods are being constantly proposed to improve on existing methods to achieve either better utility or more straightforward implementation.

When applying DP procedures in practice, especially repeated to the same dataset, one needs to keep track of the privacy loss/spending. As mentioned in Section 1, released data are synthetic data generated at a pre-specified privacy budget in all the data types and examples examined in this study. One the privacy-preserving data are released, users may perform analysis on their own without incurring additional privacy loss per the immunity to post processing property. Also noted is that absolute privacy protection for individuals in a dataset only exists on paper unless the released information is completely random or independent of the dataset. In reality, there is always some loss in privacy when releasing new and useful information; larger privacy loss parameters means sacrifice in privacy in hope for better utility in released information. Choice of a proper privacy loss is a key step when implementing DP procedures.

We hope our study and the examples shed light on privacy-preserving sharing of COVID-19 data to help promote and encourage more data sharing for research use. We will continue to conduct research on this topic and develop methods to deal with more complicated COVID-19 data sharing situations, such as releasing travel trajectories of COVID-19 patients, longitudinal data, and dynamic CTNs, CTNs with nodal attributes, etc.

References

5Lab. Covid-19 news tracker-location-based news about covid-19 in thailand. https://covidtracker.5lab.co/zh-hans?fbclid=IwAR1bAH4qDAZtWkdh2MVwAiFmow91AtRFPg78-
Ahmet Aktay, Shailesh Bavadekar, Gwen Cossoul, John Davis, Damien Desfontaines, Alex Fabrikan, Evgeniy Gabrilovich, Krishna Gadeppalli, Bryant Gipson, Miguel Guevara, et al. Google covid-19 community mobility reports: Anonymization process description (version 1.0). *arXiv preprint arXiv:2004.04145*, 2020.

Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi. Geo-indistinguishability: Differential privacy for location-based systems. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 901–914, 2013.

Apple and Google. Privacy-preserving contact tracing. [https://covid19.apple.com/contacttracing/](https://covid19.apple.com/contacttracing/), 2020. accessed on 5/23/2021.

Adrian Baddeley and Rolf Turner. Spatstat: an r package for analyzing spatial point patterns. *Journal of statistical software*, 12:1–42, 2005.

Claire McKay Bowen and Fang Liu. Comparative study of differentially private data synthesis methods. *Statistical Science*, 35(2):280–307, 2020.

Claire McKay Bowen, Fang Liu, and Bingyue Su. Differentially private data release via statistical election to partition sequentially. *METRON*, 79(1):1–31, 2021.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pages 635–658, 2016.

David Butler, Chris Hicks, James Bell, Carsten Maple, and Jon Crowcroft. Differentially private health tokens for estimating covid-19 risk. *arXiv preprint arXiv:2006.14329*, 2020.

CDC. location history. [https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/keyinfo.html](https://www.cdc.gov/coronavirus/2019-ncov/php/contact-tracing/keyinfo.html) 2021. accessed on 5/11/2021.

Paulien Coppens, Laurence Claeyts, Carina Veeckman, and Jos Pierson. Privacy in location-based social networks: Researching the interrelatedness of scripts and usage. In *Proceedings of the Symposium on Usable Privacy and Security*, 2014.

Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1):1–5, 2013.

Jinshuo Dong, Aaron Roth, and Weijie Su. Gaussian differential privacy. *Journal of the Royal Statistical Society*, 2021.

Mark Dredze, David A Broniatowski, and Karen M Hilyard. Zika vaccine misconceptions: A social media analysis. *Vaccine*, 34(30):3441, 2016.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 429–438, 2013.
Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *arXiv:1603.01887v2*, 2016.

Cynthia Dwork and Adam Smith. Differential privacy for statistics: What we know and what we want to learn. *Journal of Privacy and Confidentiality*, 1(2), 2010.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 486–503, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284, 2006b.

Cynthia Dwork, Guy N Rothblum, and Salil Vadhan. Boosting and differential privacy. In *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 51–60, 2010.

Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. Differential privacy in practice: Expose your epsilons! *Journal of Privacy and Confidentiality*, 9(2), 2019.

Evercita Eugenio and Fang Liu. Construction of differentially private empirical distributions from a low-order marginals set through solving linear equations with l2 regularization. In *Intelligent Computing, Proceedings of the 2021 Computing Conference*, volume 3, pages 949–966. Springer, 2021.

Alex Fabrikant, Andrew Mingbo Dai, Chaitanya Kamath, Charlotte Stanton, Damien Desfontaines, Dennis Kraft, Evgeniy Gabrilovich, Gerardo Flores, Gregory Alexander Wellenius, Ilya Eckstein, et al. Google covid-19 search trends symptoms dataset: Anonymization process description. *arXiv preprint arXiv:2009.01265v1*, 2020.

Diane H Felmlee and Derek A Kreager. The invisible contours of online dating communities: A social network perspective. *Journal of Social Structure*, 18:0_1–27, 2017.

Magdalena Gad-Nowak and Malgorzata P. Grzelak. Covid-19: Poland launches an official tracking app. [https://www.natlawreview.com/article/covid-19-poland-launches-official-tracking-app](https://www.natlawreview.com/article/covid-19-poland-launches-official-tracking-app), 2020. accessed on 5/17/2020.

Quan Geng and Pramod Viswanath. Optimal noise adding mechanisms for approximate differential privacy. *IEEE Transactions on Information Theory*, 62(2):952–969, 2015.

C. J. Geyer and E. A. Thompson. Constrained monte carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society, Series B*, 54:657–699, 1992.

Shirin Ghaffary. What the us can learn from other countries using phones to track covid-19. [https://www.vox.com/recode/2020/4/18/21224178/covid-19-tech-tracking-phones-china-singapore-taiwan-korea-google-apple-contact-tracing-digital](https://www.vox.com/recode/2020/4/18/21224178/covid-19-tech-tracking-phones-china-singapore-taiwan-korea-google-apple-contact-tracing-digital), 2020. accessed on 5/14/2020.
Mark S Handcock, David R Hunter, Carter T Butts, Steven M Goodreau, and Martina Morris. statnet: Software tools for the representation, visualization, analysis and simulation of network data. *Journal of statistical software*, 24(1):1548, 2008.

Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially private histograms through consistency. *Proceedings of the VLDB Endowment*, 3(1-2):1021–1032, 2010.

D. Hunter and M. Handcock. Inference in curved exponential family models for network. *Journal of Computational and Graphical Statistics*, 15:565–583, 2006.

Amol Kapoor, Xue Ben, Luyang Liu, Bryan Perozzi, Matt Barnes, Martin Blais, and Shawn O’Banion. Examining covid-19 forecasting using spatio-temporal graph neural networks. *arXiv preprint arXiv:2007.03113*, 2020.

Vishesh Karwa, Pavel N Krivitsky, and Aleksandra B Slavković. Sharing social network data: differentially private estimation of exponential family random-graph models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):481–500, 2017.

Carsten Keßler and Grant McKenzie. A geoprivacy manifesto. *Transactions in GIS*, 22(1):3–19, 2018.

Jooho Kim and Makarand Hastak. Social network analysis: Characteristics of online social networks after a disaster. *International Journal of Information Management*, 38(1):86–96, 2018.

Bai Li, Vishesh Karwa, Aleksandra Slavković, and Rebecca Carter Steorts. A privacy preserving algorithm to release sparse high-dimensional histograms. *Journal of Privacy and Confidentiality*, 8(1), 2018.

Fang Liu. Generalized Gaussian mechanism for differential privacy. *IEEE Transactions on Knowledge and Data Engineering*, 31(4):747–756, 2018.

Fang Liu. Statistical properties of sanitized results from differentially private laplace mechanism with univariate bounding constraints. *Transactions on Data Privacy*, 12(3):169 – 195, 2019.

Fang Liu. A statistical overview on data privacy. *Notre Dame Journal of Law, Ethics & Public Policy*, 34(2):477–500, 2020.

Fang Liu. Model-based differentially private data synthesis and statistical inference in multiply synthetic differentially private data. *Transactions on Data Privacy*, 15(2), 2022.

Fang Liu, Dong Wang, and Zheng-Quan Xu. Privacy-preserving travel time prediction with uncertainty using gps trace data. *IEEE Transactions on Mobile Computing*, 2021.

Fang Liu, Evercita C. Eugenio, Ick Hoon Jin, and Claire Bowen. Differentially private synthesis and sharing of network data via bayesian exponential random graph models. *Journal of Survey Statistics and Methodology*, 10:2, 2022.
Ashwin Machanavajjhala, Daniel Kifer, John Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In 2008 IEEE 24th international conference on data engineering, pages 277–286, 2008.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS’07), pages 94–103, 2007.

Ilya Mironov. Rényi differential privacy. In 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pages 263–275, 2017.

Zümrüt Müftüoğlu, M Ayyüce Kizrak, and Tülay Yildırm. Differential privacy practice on diagnosis of covid-19 radiology imaging using efficientnet. In 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), pages 1–6, 2020.

National Health Commission of China. Location privacy. https://baijiahao.baidu.com/s?id=1685656325226477831&wfr=spider&for=pc 2021a. accessed on 5/11/2021.

National Health Commission of China. publish location information. https://baijiahao.baidu.com/s?id=1688551020040753847&wfr=spider&for=pc 2021b. accessed on 5/11/2021.

European Parliament. Use of smartphone data to manage covid-19 must respect eu data protection rules. https://www.europarl.europa.eu/news/en/press-room/20200406IPR76604/use-of-smartphone-data-to-manage-covid-19-must-respect-eu-data-protection-rules 2020. accessed on 5/13/2020.

Ramesh Raskar, Sandy Pentland, Kent Larson, and Kevin Esvelt. Safe path. http://safepaths.mit.edu/ 2020. accessed on 5/17/2020.

G. Robins, T. Snijders, P. Wang, M. Handcock, and P. Pattison. Recent development in exponential random graph models for social networks. Social Networks, 29:192–215, 2007.

Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying location privacy. In 2011 IEEE symposium on security and privacy, pages 247–262, 2011.

T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New specification for exponential random graph models. Sociological Methodology, 36:99–153, 2006.

Anwaar Ulhaq and Oliver Burmeister. Covid-19 imaging data privacy by federated learning design: A theoretical framework. arXiv preprint arXiv:2010.06177, 2020.

Rasmus Plenge Waagepetersen. An estimating function approach to inference for inhomogeneous neyman–scott processes. Biometrics, 63(1):252–258, 2007.

Dong Wang and Fang Liu. Privacy risk and preservation in contact tracing of covid-19. Chance, 33(3):49–55, 2020.

Stanley L Warner. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association, 60(309):63–69, 1965.
Gregory A Wellenius, S Vispute, V Espinosa, A Fabrikant, TC Tsai, J Hennessy, A Dai, B Williams, K Gadepalli, A Boulanger, et al. Impacts of social distancing policies on mobility and covid-19 case growth in the us. *Nature Communications*, 12, 2021.

Sarah Wray. South korea to step-up online coronavirus tracking. [https://www.smartcitiesworld.net/news/news/south-korea-to-step-up-online-coronavirus-tracking-5109](https://www.smartcitiesworld.net/news/news/south-korea-to-step-up-online-coronavirus-tracking-5109) 2020. accessed on 5/14/2020.

Xiaokui Xiao, Gabriel Bender, Michael Hay, and Johannes Gehrke. ireduct: Differential privacy with reduced relative errors. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 229–240, 2011.

Yonghui Xiao, James Gardner, and Li Xiong. Dpcube: Releasing differentially private data cubes for health information. In *2012 IEEE 28th International Conference on Data Engineering*, pages 1305–1308, 2012.

Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. Differentially private histogram publication. *The VLDB Journal*, 22(6):797–822, 2013.

Fei Yan, Xing Zhang, Chang Li, Wanjie Li, Shuai Li, and Fuming Sun. Differentially private histogram publishing through fractal dimension for dynamic datasets. In *2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pages 1542–1546, 2018.

Xiaojian Zhang, Rui Chen, Jianliang Xu, Xiaofeng Meng, and Yingtao Xie. Towards accurate histogram publication under differential privacy. In *Proceedings of the 2014 SIAM international conference on data mining*, pages 587–595, 2014.
Supplementary Materials to “Statistical Evaluation of Privacy-preserving Publication and Sharing of Three Types of COVID-19 Pandemic Data: Methods and Case Studies”

UH and UHp for Surveillance Case Number Release

The UH approach forms a hierarchical tree among the data attributes and injects noise to each the node count in each layer of the tree, explores equality constraints between each parent node and its children nodes in the tree to improve the accuracy of the sanitized count of the parent nodes (low-order marginals) and release the final corrected counts from the whole tree. Figure S1 displays a 4-layer hierarchical tree formed in the UH approach on a data set with 3 variables (age group, minority/majority, sex). We refer to the node at the top of the tree as the root (layer 1) and those at the bottom as the leaf nodes (layer 4). The age nodes at layer 3 are parents to the race/ethnicity nodes in layer 3, which are the parent nodes to the sex nodes in layer 4. There is no particular ordering among the three attributes in the example in Figure S1. We can place the attributes in the middle layers of the trees that would enjoy a lower mean squared error (MSE) in their marginal sanitized counts relative to their original counts, compared to the MSE resulting from a simple sum of the directly sanitized counts of the most granular cells as done in the flat sanitizer.

![Hierarchical Tree](image)

Figure S1: A hierarchical tree with three binary attributes

The UH procedure is implemented in 3 steps. First, since each layer is sanitized, the total budget $\epsilon$ should be split among the layers following the sequential composition principle in DP [McSherry and Talwar, 2007]. For illustration purposes, we assume each layer receives $1/l$ of the total $\epsilon$, where $l$ is the height of the tree [14] and $l = 4$ in Example 1. The count $\hat{h}[v]$ in each node $v$ in the tree is sanitized via the Laplace mechanism $\text{Lap}(0, l\epsilon^{-1})$; that is, $\hat{h}[v] = h[v] + e$, where $e \sim \text{Lap}(l\epsilon^{-1})$, where $\hat{h}[v]$ is the sanitized count. In step 2, intermediate node count $z[v]$ for each node $v$ is obtained via Eqn (S1),

$$z[v] = \begin{cases} \hat{h}[v], & \text{if } v \text{ is the leaf node} \\ \frac{k^l-k^{l-1}}{k^{l-1}}\hat{h}[v] + \frac{k^{l-1}-1}{k^{l-1}} \sum_{u \in \text{succ}(v)} z[u], & \text{o.w.} \end{cases}$$

13MSE for a sanitized count $\hat{x}$ is $E_M(\hat{x} - x)^2$, where $x$ is the original count and the expectation is taken over the distribution of the randomized algorithm.

14Other privacy allocation schemes across the layers can be used.
where $\text{succ}(v)$ denotes the set of children nodes to parent node $v$ and $k$ is the number of children per parent node, which is assumed to be the same for each parent ($k = 2$ in example 1). The reason behind Eq (S1) is that for the nodes not from the bottom layer (the non-leaf nodes), a sanitized count comes from two sources (the node being sanitized, and the summation from its children nodes) so Eq (S1) calculates a weighted average of the two. Obviously, $z[v]$ may no longer equal to the sum of the node counts of its children nodes, violating the equality constraints in contingency tables. This inconsistency is corrected via Eq (S2), yielding the final sanitized count $h^*[v]$

$$h^*[v] = \begin{cases} 
  z[v] & \text{if } v \text{ is the root node} \\
  z[v] + k^{-1} \left( h^*[u] - \sum_{w \in \text{succ}(u)} z[w] \right) & \text{o.w.} 
\end{cases}$$

(S2)

where $u$ is the parent node to node $v$, $\text{succ}(u)$ contains the children nodes to parent node $u$, and $h[u] - \sum_{w \in \text{succ}(u)} z[w]$ is the correction term to ensure the equality constraint holds for each parent node in the tree.

We extend the UH approach to sanitizing a proportion tree (Figure S1) in place of a count tree and name it the UHp approach (“p” in the name “UHp” stands for “proportion”), in cases where the total sample size $n$ is public information and can be released directly, or when it is desirable not to alter $n$ from a statistical inferential perspective as $n$ is critical for inferences such as inferential efficiency.

The sanitization process for UHp is similar to UH with a few modifications. First, given the proportion at the top layer is always 1, there is no need to sanitize the node and the total $\epsilon$ is only needed to split into $l - 1$ layers. Second, the Laplace distribution from which the noise is drawn becomes Laplace($0, (l - 1)\epsilon^{-1}/n$) as the global sensitivity for proportion is $1/n$. Third, after obtaining $\tilde{h}[v] = h[v] + e$, where $e \sim \text{Lap} \left( (l - 1)\epsilon^{-1}/n \right)$ for all the non-root node proportions, we normalize the proportions in layer 2 as in $\tilde{h}[v^{(2)}] / \sum_u \tilde{h}[u]$, where $u$ refers to all the nodes in layer 2, so that the layer-2 proportions sum up to 1, honoring the constraint of $\tilde{h}[v] = h[v] = 1$ for the root node. The steps in Eqs (S1) and (S2) after the normalization step remain the same as in the UH approach. After the sanitized proportions are obtained, the corresponding counts can be obtained by multiplying the proportions with the total $n$.

Similar to the flat sanitizer, the sanitized counts or proportions in the UH and the UHp approaches can be negative as the support of the Laplace distribution is $\mathbb{R}$. In addition, the sanitized proportions may be $> 1$. We applied the same methods as used for the flat sanitizer.
to deal with negative counts and in the case of a fixed upper bound such as the proportions add up to 1 and when the total count is fixed.

Simulation study and CDC death count application for UH and UHp

In the simulation study, for both UH and UHp, the tree height is $l = 4$ as there are 3 attributes – $X_1$ is layer 2, $X_2$ in layer 3, and $X_3$ in layer 4 – and $k = 2$ as all attributes are binary. The simulation results are presented in Figure S3, together with the flat Laplace sanitizer and the original results for comparison.

![Figure S3: Privacy-preserving inference from the log-linear model based on sanitized counts via the flat Laplace sanitizer at various $\epsilon$ values in the simulation study ($m = 3$, 500 repeats)](image)

For the application to the CDC COVID-19 death count data, $l = 3$ and $k = 7$ in the hierarchical tree for both UH and UHp. We placed age in layer 2 and race/ethnicity in layer 3 and don’t expect the ordering would affect the results of the analysis we conducted in a statistically meaningful way. The results are presented in Figure S4, together with the flat Laplace sanitizer and the original results for comparison.
Figure S4: Privacy-Preserving results from the Log-linear model fitted on the CDC COVID-19 death data
Examples of sanitized US COVID-19 death counts by the Flat Laplace sanitizer ($m = 3$ and $\epsilon = 0.5$)

Table S1: Sanitized US COVID-19 death counts by age group and race/ethnicity on May 24, 2022 ($m = 3, \epsilon = 0.5$)

| Age (ys) | NH White | NH Black | NH AIAN | NH Asian | NH NHPI | NH Mix | Hispanic | Total |
|----------|----------|----------|---------|----------|---------|--------|----------|-------|
| <17      | 385      | 273      | 19      | 38       | 1        | 25     | 302      | 1056  |
| 18-29    | 2265     | 1450     | 183     | 181      | 33       | 78     | 2018     | 6238  |
| 30-49    | 6665     | 4140     | 570     | 561      | 151      | 160    | 5916     | 18162 |
| 40-49    | 17277    | 8939     | 1024    | 1200     | 267      | 313    | 13979    | 43000 |
| 50-64    | 97407    | 35752    | 3196    | 5310     | 703      | 955    | 43655    | 186979|
| 65-74    | 141417   | 37760    | 2913    | 7431     | 498      | 914    | 38416    | 229350|
| >75      | 380635   | 54588    | 3210    | 16515    | 447      | 1383   | 56700    | 513478|
| Total    | 646051   | 142933   | 11115   | 31236    | 2114     | 3827   | 160986   | 998262|

| Age (ys) | NH White | NH Black | NH AIAN | NH Asian | NH NHPI | NH Mix | Hispanic | Total |
|----------|----------|----------|---------|----------|---------|--------|----------|-------|
| <17      | 387      | 289      | 19      | 34       | 11      | 35     | 304      | 1079  |
| 18-29    | 2273     | 1492     | 183     | 197      | 54      | 86     | 2014     | 6299  |
| 30-39    | 6680     | 4124     | 568     | 571      | 147     | 157    | 5917     | 18145 |
| 40-49    | 17269    | 8930     | 1029    | 1216     | 279     | 314    | 13972    | 43010 |
| 50-64    | 97386    | 35756    | 3200    | 5315     | 723     | 953    | 43647    | 186982|
| 65-74    | 141407   | 37753    | 2890    | 7426     | 513     | 918    | 38420    | 229327|
| >75      | 380591   | 54568    | 3205    | 16518    | 450     | 1382   | 56706    | 513420|
| Total    | 645974   | 142912   | 11095   | 31278    | 2177    | 3846   | 160980   | 998262|

| Age (ys) | NH White | NH Black | NH AIAN | NH Asian | NH NHPI | NH Mix | Hispanic | Total |
|----------|----------|----------|---------|----------|---------|--------|----------|-------|
| <17      | 392      | 284      | 19      | 29       | 10      | 29     | 309      | 1072  |
| 18-29    | 2236     | 1507     | 189     | 196      | 48      | 62     | 2007     | 6243  |
| 30-39    | 6659     | 4146     | 563     | 562      | 150     | 153    | 5903     | 18135 |
| 40-49    | 17260    | 8933     | 1002    | 1208     | 287     | 316    | 13983    | 42987 |
| 50-64    | 97418    | 35743    | 3198    | 5312     | 719     | 964    | 43674    | 187027|
| 65-74    | 141398   | 37767    | 2897    | 7437     | 516     | 903    | 38420    | 229339|
| >75      | 380604   | 54573    | 3192    | 16513    | 460     | 1391   | 56724    | 513458|
| Total    | 645966   | 142953   | 11060   | 31255    | 2190    | 3818   | 161020   | 998262|

Race/ethnicity = 'unknown' is not included in the table.
NH = Non-Hispanic; AIAN = American Indian or Alaska Native; NHPI = Native Hawaiian or Other Pacific Islander;
"Mix" means "more than one race".
The RR mechanism and the RR-debiased procedure

The RR mechanism for sanitizing edges in a network works as follows. Let $p_{ij}$ denote the probability the original edge $e_{ij} = 1$ is retained and $q_{ij}$ be the probability that $e_{ij} = 0$ is retained after sanitization for nodes $i \neq j = 1, \ldots, n$. To satisfy $(\epsilon_{ij})$-DP in the sanitization of the relational information $e_{ij} = 1$, one may set $p_{ij} = q_{ij} = \epsilon_{ij}/(1 + \epsilon_{ij})$. When there is no particular reasons for using different $\epsilon_{ij}$ for different pairs of nodes, one may set $\epsilon_{ij} \equiv \epsilon$ and the probability of edge flipping in the network is

$$p_{ij} = q_{ij} \equiv 1/(1 + \epsilon).$$ \hspace{1cm} (S3)

If all edges are mutually independent, the total cost for sanitizing the whole network is also $\epsilon$ per the parallel composition principle.

Liu et al. (2022) employs a debiasing approach as an attempt to remove bias in sanitized networks via the RR mechanism (with edges $e^*_{ij}$) by synthesizing new networks with edges $\tilde{e}^*_{ij}$ given a RR-sanitized network. Specifically,

$$\tilde{e}^*_{ij} | e^*_{ij} = 1 \sim \text{Bern}(p_1), \quad \text{where } p_1 = \frac{(p + q - 1)q}{(2q - 1)p},$$ \hspace{1cm} (S4)

$$\tilde{e}^*_{ij} | e^*_{ij} = 1 \sim \text{Bern}(p_0), \quad \text{where } p_0 = \frac{q(p + q - 1)}{(1 - p)(2q - 1)},$$ \hspace{1cm} (S5)

where $q = \epsilon/(1 + \epsilon)$ is the probability of retaining an original edge by RR and $p$ is the proportion of all $e^*_{ij} = 1$ in row $i$ of the adjacency matrix of the synthetic network generated by RR (without the diagonal element), and. Synthetic networks via RR-debiased can be summarized and analyzed in the same way as the original network including descriptive statistics, visualization, and inference. For inference, there is no need to explicitly model the RR mechanism or the subsequent debiasing/sanitization process if $m > 1$ sets of synthetic networks are released. The debiasing procedure does not use the information from the original network and thus maintains the privacy guarantees, but at the cost of introducing another layer of variability. The debiased sanitized network is made of edges $Y^*$ drawn from two Bernoulli distributions, depending on whether the synthetic edge $Y^*$ from the DWRR is 1 or 0.

Simulation Studies on RR and RR-debiased

For RR, the probability of flipping an edge per Eq (S3) is $(1 + \epsilon^5)^{-1} = 0.7\%$, $(1 + \epsilon^2)^{-1} = 11.9\%$, $(1 + \epsilon)^{-1} = 26.9\%$ and $(1 + \epsilon^{0.5})^{-1} = 37.5\%$ at $\epsilon = 5, 2, 1, 0.5$, respectively. Though the probability retaining the original relation between nodes $i$ and $j$ is very low at $\epsilon = 5$, but the number of edges is expected to double $(39 * \epsilon^5/(1 + \epsilon^5) + (4950 - 39) * \epsilon^5/(1 + \epsilon^5) = 71.6$ where 39 is the edge count in the original network).

The sanitized CTNs via RR and RR-debiased are presented in Figure S5 with the original CTN presented for comparison. Table S2 presents the # of edges and # triangles of the sanitized CTNs via RR and RR-debias. Figure S6 shows the DD, which is the distribution of close contacts of an individual in a CTN via RR and RR-debias and Figures S7 depicts the ESPD of the sanitized CTNs with the TVD in DD between the sanitized and original CTNs. Figure S8 shows the box plots of the betweenness centrality and closeness centrality of the 100 nodes in the sanitized CTNs via RR and RR-debias vs the original.
Figure S5: Examples of differentially privately sanitized CTNs via RR and RR-debiased.

Table S2: Average (SD) number of edges and number of triangles over 100 repeats

| $\epsilon$ | RR          | RR-debiased       |
|-----------|-------------|------------------|
|           | # of edges  | # of triangle    | # of edges  | # of triangle    |
| 0.5       | 1876 (37.4) | 8802 (528.3)     | 844 (82.4) | 1258 (356.4)     |
| 2         | 619 (20.9)  | 320 (38.1)       | 182 (21.9) | 14 (6.1)         |
| 5         | 72 (5.2)    | 10 (0.9)         | 48 (6.2)   | 6 (2.5)          |
| 8         | 40 (1.3)    | 10 (0.3)         | 40 (1.4)   | 9 (0.7)          |
| 10        | 39 (0.5)    | 10 (0.1)         | 39 (0.4)   | 10 (0)           |

original: # of edges = 39; # of triangle = 10.

Figure S6: Degree distribution in the original and sanitized CTNs via RR and RR-debias.

S7
Figure S7: Edgewise shared partner distribution in sanitized CTNs via RR and RR-debias

(a) RR 
(b) RR-debiased

Figure S8: Box plots of betweeness centrality and closeness centrality of 100 nodes in original and sanitized CTNs via RR and RR-debias

Figure S9: Comparison of GI, RR, RR-debias on sensitivity, specificity and false discovery rate

Figure S9 presents the comparison among GI, RR and RR-debias on sensitivity, specificity, and FDR in sanitized networks.

The simulation results on the privacy-preserving inferences from the ERGM based on the sanitized CTNs via RR and RR-debias are presented in Table S3. We examined $\epsilon = 5, 15, 18, 24$ as the inferential results at $\epsilon < 5$ are worse.

$^{15}$Precisely speaking, $\epsilon_{ij}$ edge DP; see Liu et al. (2022) for details.
Table S3: Privacy-preserving Inference of $\beta$ in the ERGM model based sanitized CTNs via RR and RR-debiased ($m = 3$; 500 repeats)

| method       | metric   | $\epsilon = 5$ | $\epsilon = 15$ | $\epsilon = 18$ | $\epsilon = 24$ |
|--------------|----------|----------------|-----------------|-----------------|-----------------|
|              | bias     | 3.260          | 0.627           | 0.269           | 0.023           |
| RR           | RMSE     | 3.260          | 0.635           | 0.299           | 0.165           |
|              | CP       | 0              | 0.002           | 0.366           | 0.832           |
|              | bias     | 1.500          | 0.305           | 0.147           | 0.008           |
| RR-debiased  | RMSE     | 1.501          | 0.330           | 0.204           | 0.165           |
|              | CP       | 0              | 0.366           | 0.704           | 0.830           |

Original data: bias $= -0.021$, RMSE $= 0.171$, and CP $= 0.942$. 