# Intelligent Pandemic Surveillance via Privacy-Preserving Crowdsensing

Hafiz Asif, Periklis A. Papakonstantinou, Stephanie Shiau, Vivek Singh, and Jaideep Vaidya,
*Rutgers University, Camden, NJ, 08901-8554, USA*

*Intelligently responding to a pandemic like Covid-19 requires sophisticated models over accurate real-time data, which is typically lacking at the start, e.g., due to deficient population testing. In such times, crowdsensing of spatially tagged disease-related symptoms provides an alternative way of acquiring real-time insights about the pandemic. Existing crowdsensing systems aggregate and release data for pre-fixed regions, e.g., counties. However, the insights obtained from such aggregates do not provide useful information about smaller regions—e.g., neighborhoods where outbreaks typically occur—and the aggregate-and-release method is vulnerable to privacy attacks. Therefore, we propose a novel differentially private method to obtain accurate insights from crowdsensed data for any number of regions specified by the users (e.g., researchers and policy makers) without compromising the privacy of the data contributors. Our approach, which has been implemented and deployed, informs the development of the future privacy-preserving intelligent systems for longitudinal and spatial data analytics.*

At the heart of tracking Covid-19 (or other similar infectious diseases) are *spatiotemporal range (SR) queries*. An SR query asks for the number of data points in a given region and a time period, e.g., how many people were diagnosed with Covid-19 in Downtown Manhattan in the past 14 days? These queries provide crucial information such as daily cases, moving averages, and cumulative case counts for sick people. Importantly, they are used to identify and track new and emerging hotspots. Thus, answering SR queries is vital to making informed policy decisions to curb the spread of the disease, e.g., by having smart-lockdowns instead of locking down the country and its economy.

In the early months of the Covid-19 pandemic, testing was deficient, and there was inadequate data about the actual Covid-19 case counts to make any informed decisions. Therefore, many symptom-tracking apps were developed that crowdsensed location-tagged Covid-19 related symptoms to estimate the lacking information (e.g., region-wise case counts). Such apps were adopted across the world[1–4] to help track Covid-19.

However, all such existing apps are rudimentary in their functionality. First, they aggregate and release insights, e.g., the count of symptomatic people, for a fixed set of administrative regions. For instance, the fixed regions for “COVID Near You,” “How We Feel,” and “Facebook” symptom-tracking apps are respectively based on ZIP codes, towns, and counties. Thus, these apps can only answer a limited set of SR queries and allow a user to only access information for their respectively defined prefixed administrative regions that arbitrarily vary in sizes and population density. Hence, they are unable to effectively track the emergence of the virus since the outbreaks are typically more localized and do not adhere to administrative boundaries.

Second, the aggregate-and-release method—used by all the symptom-tracking apps—fails to protect the privacy of all the data-contributors as has been shown numerous times by various attacks.[5–8] Indeed, even if these apps had allowed aggregation over arbitrary regions, they would have suffered from significant privacy problems—in terms of the data used in our experimental evaluation, up to 30% of the symptom reports were individually identifiable, i.e., there were

1541-1672 © 2022 IEEE
Digital Object Identifier 10.1109/MIS.2022.3145691
Date of publication 25 January 2022; date of current version 12 September 2022.

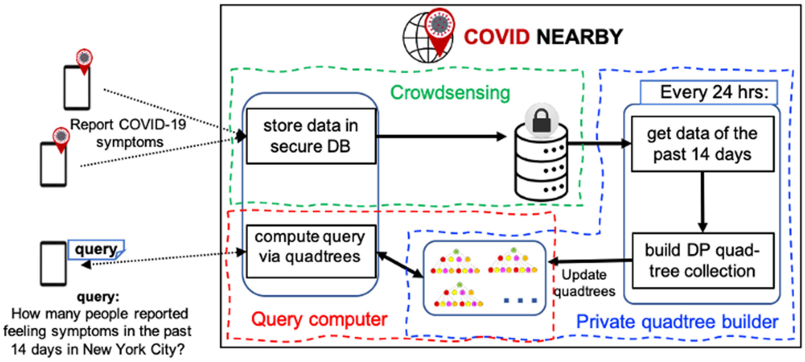
COVID NEARBY
Report COVID-19 symptoms
Crowdsensing
store data in secure DB
Every 24 hrs:
get data of the past 14 days
build DP quad-tree collection
Update quadtrees
Private quadtree builder
query
compute query via quadtrees
Query computer
query:
How many people reported feeling symptoms in the past 14 days in New York City?

**FIGURE 1.** The system has three components. The Crowdsensing component collects and stores user reports. Private quadtree builder dynamically partitions the space via a collection of DP quadtrees. Query computer uses the DP quadtrees to compute SR queries.

no other reports within a radius of 0.5 sq. km. The percentage of reports at risk would increase with more sophisticated queries such as intersection queries. Furthermore, we note that the use of intelligent and AI systems in epidemiology is recent, and so far, privacy-preserving solutions are lacking.[9,10]

Therefore, to address the above-mentioned shortcomings, we propose a crowdsensing system to develop symptom-tracking apps that, compared to the existing systems, can answer any number of SR queries for any user-specified region, all the while guaranteeing differential privacy (DP)[11]—a provable guarantee—for all the data contributors. We have also deployed this system via web and phone apps under the Covid Nearby Project.[12]

Here, the main problem to solve is: how to enable users to ask any number of SR queries (e.g., the number of symptomatic people) for any region of their choice but without adversely impacting privacy. We solve this problem under the provable guarantee of DP. DP provides state-of-the-art data-privacy protection, wherein the risk to a data contributor's privacy is specified via a parameter $\varepsilon > 0$: the higher its value, the higher the risk. Our system guarantees that its answers to the SR queries remain statistically almost identical regardless of whether any single user reported their information or not. Furthermore, by design, the number of queries and choices for the query regions are unbounded. Thus, the main challenge is to answer all the SR queries—for different regions and at different or overlapping times—with DP but without increasing the privacy risk (i.e., for a reasonably small value of $\varepsilon$) or degrading the accuracy of the answers. Note that accuracy is equally important here as one can achieve perfect privacy by giving completely random answers.

To solve the challenge mentioned above, our system creates differentially private spatially indexed hierarchical partitions of the space (e.g., the USA) using temporally partitioned data and computes the DP count of the reports (e.g., the number of symptomatic people) in each partition. We then use these indexed partitions with their corresponding DP counts to compute SR queries. Figure 1 gives the system-level overview.

This enables a privacy-preserving crowdsensing based pandemic surveillance system that:

1) generates insights about the pandemic while guaranteeing privacy, in particular, it guarantees $\varepsilon$-DP over an unbounded number of arbitrarily chosen SR queries;
2) can be used to identify regions, for example within counties, with higher cases/reports (by identifying the child nodes with the highest depth); and
3) is computationally efficient and accurate.

## DIFFERENTIALLY PRIVATE SR QUERIES VIA SPATIOTEMPORAL PARTITIONING

We consider the following setting:

1) The space (of possible locations) is two-dimensional and bounded, with north-south and east-west as two perpendicular axes (e.g., obtained by Mercator).
2) The database contains the reports that:
   i. are from within the USA and tagged with location coordinates in the space;
   ii. contain at least one Covid-19 symptom.
3) The database is stored with a trusted curator who answers SR queries using a DP algorithm.

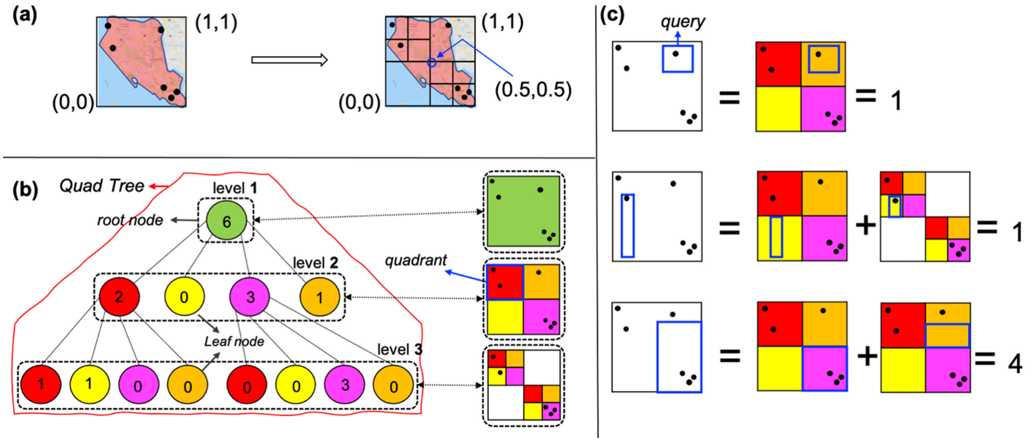

**FIGURE 2.** (a) Partition of space created by a quadtree; black points represent the data. (b) Quadtree (max height = 3, count threshold = 1) for (a). (c) How to compute SR queries via the quadtree's quadrants; the answer to the third query is 4 because a node only stores the count of its quadrant.

4) There is only one report per user (this assumption is relaxed and discussed later).
5) Any SR query's region is an axis parallel rectangle, and for simplifying the exposition, the time period for it at any given day is the past $\mathrm{d}$ days, with $\mathrm{d} = 14$.

## Overview

To create spatially indexed partitions of the space, we use a hybrid of data-agnostic and data-dependent approaches. First, we partition the space without looking at the data, based on administrative units, e.g., counties—let us call them *divisions*. Then, every day for each division, we temporally partition the data reported from the division into $k$ groups and create $k$ different data-dependent partitions of the same division by building $k$ DP quadtrees over it; each of the quadtrees uses the data from one of the $k$ groups and is guaranteed to be ($\varepsilon$/n)-DP (here $k = \lceil d/n \rceil$). We use a covering algorithm to create temporal partitions (every day); it ensures that no report is included in more than $n$ groups created over time. Thus, the overall privacy risk remains at most $\varepsilon$. The details follow.

## Spatial Partitioning

We first partition the space data-agnostically (without looking at the data) using division as a county. Then, on any given day, we partition the division data-dependently by building a quadtree,[13] a hierarchical spatial data structure. For a given rectangular space and the data lying in it, a quadtree recursively, level by level partitions the space into rectangular regions (called *quadrants*) by bisecting their sides. Each quadrant represents a region of the division and holds the DP count of the reports from the region it represents [e.g., see Figure 2(a) and (b)].

To build a quadtree over a division, we use the smallest rectangle bounding the division and use the length $w$ (in Km) of its longest side to compute the max height as $h = \lfloor log_2 w \rfloor$, i.e., the maximum number of levels. This strategy ensures that one of the quadrant's sides will be $1$ km long, even in the smallest partition. Furthermore, since the privacy budget is prefixed, knowing $h$ beforehand allows for a better allocation of the privacy budget, $\varepsilon$ , and significantly improves the accuracy. We set $\varepsilon' = 1$ at the root level, and divide the remaining budget (i.e., $\varepsilon - \varepsilon'$) geometrically among the rest of the levels as it probably reduces the error over queries.[13] To stop partitioning zero or low (symptomatic reports) count subregions and focus on the subregions in the division with a high count, we use a minimum count threshold, $c$ , as a prerequisite for dividing a subregion. This thresholding further improves the accuracy and efficiency of the system. For Covid Nearby, we set $c$=10, and for the data, we use the past *d* days' data reported from the division—recall that *d*=14.

## Temporal Partitioning

To reduce the error in achieving $\varepsilon$-DP, we do not build one $(\varepsilon/d)$-DP quadtree per day over the past $d$ days' data. Instead, we temporally partition the past $d$ days' data into $k$ groups, called *acceptable partition* (described below), and build $k$-many $(\varepsilon/n)$-DP quadtree (one over each group with $k = \lceil d/n \rceil$ ). Note that $\varepsilon$ is reduced (to $\varepsilon/n$ and $\varepsilon/d$) to make overall risk $\varepsilon$, which

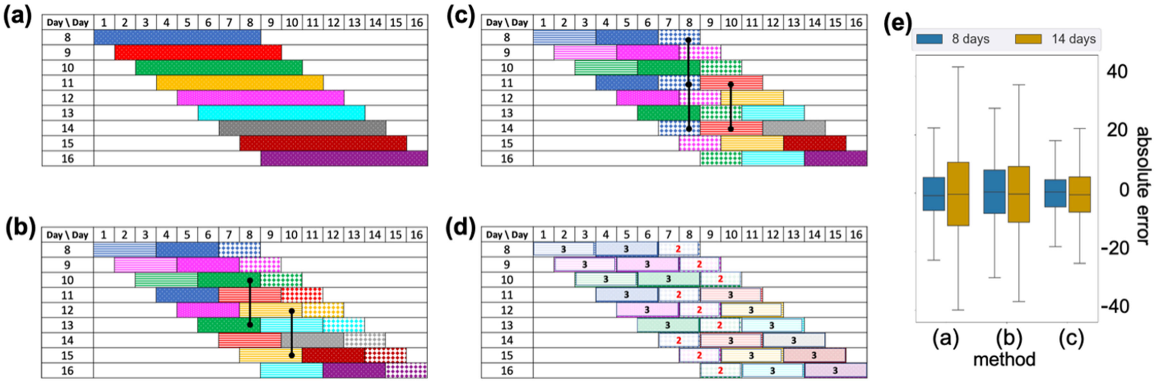

**FIGURE 3.** Three methods of creating acceptable partitions of 8 days, i.e., *d* = 3. For (b)–(d) *n* = 3 Both rows and columns show the progression of time. The colored rectangles, together in each row, show the acceptable partition for the day labeling the row. Rectangles in each row give the groups in the acceptable partition, and each is uniquely identifiable by its color and pattern—therefore, any two rectangles with the same color and pattern across rows refer to the one unique group of days. (a) Naïve approach that groups all eight days into one group (i.e., *n* = *d* = 8), and each is unique. (b) Acceptable partitions for *n* = 3 and *k* = 3. The vertical connections explicitly show, as an example, the same unique group across different partitions. Similarly, (c) Acceptable partitions by the covering algorithm[14] where the groups are reordered in a particular way; compared to (a) and (b), this method produces a lesser number of unique groups. (d) Reordering of the groups in terms of their sizes done via a circular-shift after every *n* = $3$ days. (e) Three methods are compared via boxplot of the noise (generated over 100 iterations) at the root level of a quadtree corresponding to the three methods given in (a)–(c) for the same privacy risk.

follows from the serial composition property of DP[11]; it bounds the privacy risk of a data record used in $N$ independent $\varepsilon$-DP queries by $N\varepsilon$. For instance, building one $\varepsilon'$-DP quadtree (per day) over the past 14 days' data gives an overall privacy risk of 14 $\varepsilon'$ over time because each day's data is used in building 14 quadtrees [Figure 3(a), where for simplicity, we use *d*=8 instead of *d*=14].

For any given $d$ consecutive days, an *acceptable partition* divides the $d$ days into *k*=⌈*d/n*⌉ groups of consecutive days such that:

1) there are $k-1$ groups of size $n$;
2) there is one group of size $r$ , where *d* = *n*(*k*-1) + *r*; and
3) each of the $d$ days is present in exactly one group [e.g., see Figure 3(b) and (c), where each given temporal partition is acceptable, and *d* = *8* and *n* = *3*].

Given an acceptable partition, $P$, of the past $d$ days, creating an acceptable partition of *the data* is straightforward: combine the data from all the days in each group of $P$, which divides the data into $k$ groups.

Since we need to create an acceptable partition of the data every day and build new quadtrees, naïve methods incur a higher privacy risk even when we build only one quadtree per one unique group of data. For instance, one such naïve method incurs $(n+r)\varepsilon$ privacy risk under DP when $r \neq n$, (e.g., see Figure 3(b), where privacy risk can be calculated by multiplying $m$, i.e., the max number of unique groups a day is included in, with $\varepsilon$). Here, we use a covering algorithm[14] for this task; it starts with a given acceptable partition, $P$, (of the days) and shifts $P$ to the right every day with an additional circular-shift after every $n$ days (see Figure 3 (c) and (d); see the article by Asif[14] for details). The covering algorithm guarantees that every day will at max be present in $n$ unique groups from all the acceptable partitions created over time. Thus, giving a privacy risk of $n\varepsilon$. This approach reduces the magnitude of noise added to achieve $\varepsilon$-DP and improves the accuracy [see Figure 3(d)].

For our case, i.e., $d=14$, we specify $k$ by choosing and updating $n$ over time. We use the following empirically supported heuristic, see the article by Asif[14] for this task. Set $n=1$ if the number of reports, $\#R$, from the division is less than 19 (we use DP counts to compute $\#R$). When $\#R$ exceeds 19, we pick $n$ based on $\#R$ and the max height of the quadtree, $h$, for the division. When $h \leq 4$ and $20 \leq \#R \leq 4^4$, we set $n=7$. When $h=6$ and $20 \leq \#R \leq 4^4$, we set $n=3$. For the rest of the cases, we set $n=2$.

## SR Query Computation

To compute an SR query, we find all the divisions that intersect with the query region, then compute the query over the quadtree for each such division and aggregate their results to compute the final answer. To compute a query over a quadtree, we traverse the tree to find all the quadrants that intersect with the query region and sum their counts to compute the result [e.g., see Figure 2(c)]. To improve the accuracy, we use the count of the parent node if all of its children are selected.[13]

Since the DP quadtrees do not store the actual points, we get the count for the whole quadrant, even when the query region partially intersects the quadrant [3rd query in Figure 2(c)]. In such a case, one can improve the estimate by employing uniformity assumption,[13] and giving the count proportional to the area, $A(R \cap Q)$, of the query region $R$ that intersects with a quadrant $Q$, i.e., $c_Q \times A(R \cap Q))/A(Q)$, where $c_Q$ is the count for $Q$ and $A(Q)$ is the area of $Q$. However, in many instances, the actual region from a division makes a small part of a quadrant's region. This is because we build quadtrees over the bounding boxes of the divisions, and in such instances, a quadrant's area can be much larger than the actual area of the region of the division it contains; thus, proportional counts give a lower estimate. We solve this problem by taking a polygonal (shape) approximation of divisions and using the intersection of the polygon with a quadrant as the area of the quadrant and the intersection of the query region with the polygon and the quadrant as the area of the query region in the quadrant to compute the proportional count.

## Data Inclusion Criterion and Privacy

Our approach protects every report with an $\varepsilon$-DP guarantee. However, when a user reports more than once, the user's privacy risk increases linearly with the number of reports the user makes (due to serial composition of DP). To control this risk, we limit a user's reports that we use to build quadtrees; this simple technique works in practice effectively.[15] Let us say $D$ is the database in which we insert selected reports; it will be used to build quadtrees. We restrict the total number of reports (by a user) inserted in $D$, to be $N$ in the following way. At any day, any user $u$ can submit only one report, which we *insert* in $D$ if the following two conditions are met.

1) If, in the past $d$ days, no report from $u$ was inserted in $D$, and
2) The total number of reports by $u$ that were inserted in $D$ is less than $N$.

This insertion mechanism will incur a privacy risk of $N\varepsilon$ for any data contributor. If one wants to limit the privacy risk to $\varepsilon$one can build quadtrees that are $\varepsilon/(nN)$-DP. In the case of Covid-19, where $d = 14$, having $N = 2$, covers a one-month long period for any user, covering most cases of interest.

# EMPIRICAL EVALUATION

Our method has been validated through an empirical evaluation over spatially disaggregated real data of confirmed Covid-19 cases.[16] The original Covid-19 data was given as the aggregate counts (of confirmed Covid-19 cases) at the county level for each day. Therefore, we first disaggregated the data for each county for each day. To do this, we estimated the radius of each county and used it to parameterize the scale of the exponential distribution Q. We prepare as many data points as the count of the county. Then, for each data point, we sample the distance *r* from the center of the county and pick the point's location uniformly on the circle of radius *r*, centered at the county's center coordinate.

The results show that the DP answers, computed via our method, are highly accurate [see Figures 4(a) and (b) and 5(a), (c), and (d)]. This is true even for the arbitrarily picked region within a division [see Figure 4(b)]. Further, our approach yields a much lower error than a baseline approach with the same privacy [see Figure 4(c)]. As noted earlier, the smaller the value of the privacy parameter, $\varepsilon$, the lower the privacy risk. Given the scale and geographic scope of the system, we use $\varepsilon = 6$ following the US Census Bureau's preliminary $\varepsilon$-allocation in 2019 for the 2020 census.[17] We note that the US Census Bureau has since significantly increased $\varepsilon$ and set it to 19.61.

Besides computing SR queries for arbitrary regions—which are used to compute a variety of information to track the pandemic—the DP quadtrees can be combined to compute accurate cumulative case counts over time as well as rank and identify hotspots at the state/county level [see Figure 5(a)–(c)]. The relative error for the cumulative counts in the start is relatively high [see Figure 5(b)] because in the early months of the pandemic (e.g., March–May 2020), the actual counts were very small for most of the states; for this very reason the average relative error over the 25 states with the most cases is always lower than that over all the states.

We use SR queries to compute the moving averages of the new cases (or symptomatic reports) with a very low error [see Figure 5(d)]—one can combine different quadtrees to get a moving average different from 14-days. Even in the case of moving average, if the count is sufficiently high, the error is negligible [see Figure 5(d) and (e)].

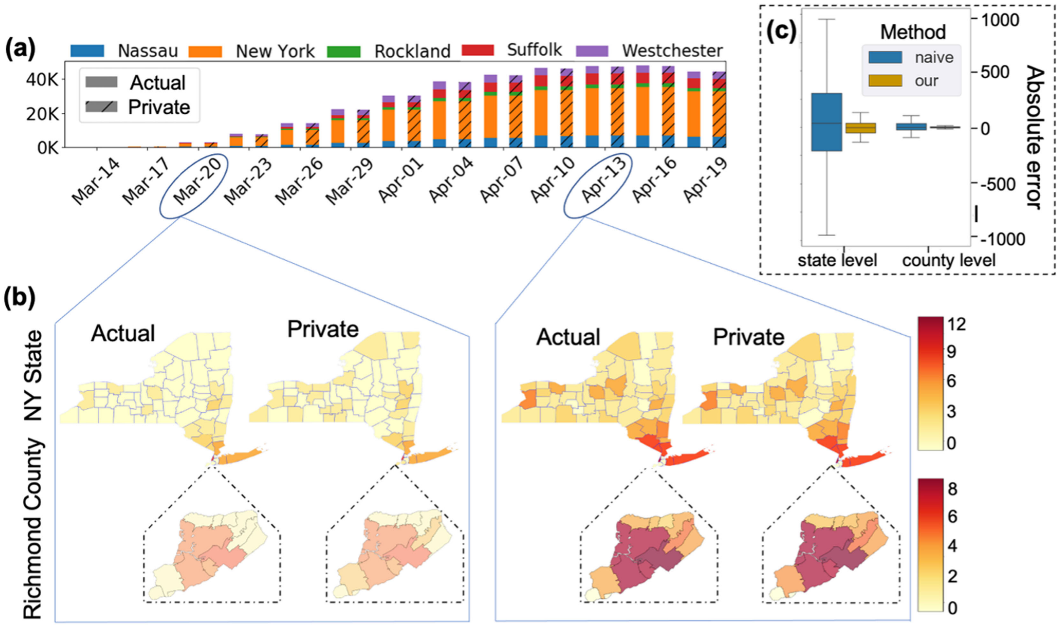

**FIGURE 4.** Private counts are DP answers to SR queries computed by our method from the actual Covid-19 case count. (a) Stacked bar-chart of the case counts of the 5 NY counties with the most Covid-19 cases. For each day: (i) two stacked bars are given, the first for the actual counts, and the second for the private counts; and (ii) each bar gives the total Covid-19 cases for the past 14 days. (b) Heatmaps of the actual and private 14-day case counts (on a log scale) for NY state and Richmond County. (c) Our method to a method based on a naïve spatial partitioning approach (i.e., DP data aggregation over partitions created by a fixed grid base partitioning with a cell-size of 1 km$^2$), which guarantees the same level of privacy. Both the methods are probabilistic, and therefore the boxplots are computed over 100 iterations.

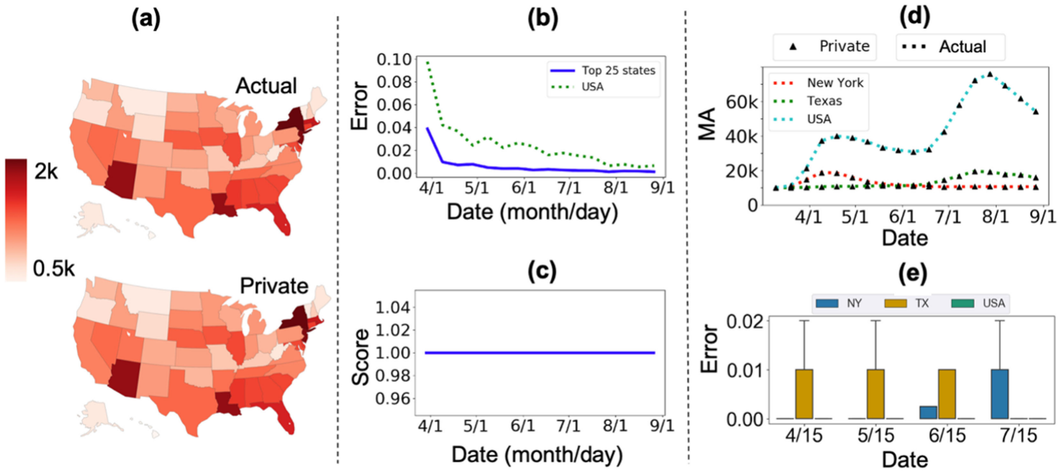

**FIGURE 5.** Private counts refer to the counts computed by our method from the actual Covid-19 case count. (a) Heatmaps of the cumulative counts at the state level, both actual and private, on 7/17/2020. (b) Relative error in cumulative counts using our method for the top 25 states (by case count) and the entire U.S. over the period from 3/20/2020 – 9/1/2020. (c) Kendall's $\tau$ (rank correlation coefficient [18]) of the two ranked lists of states obtained from the private and the actual answers of SR queries for counties; $\tau = 1$ when the ranking is identical. (d) 14 days moving average of both the private and the actual counts for New York, Texas, and the entire U.S. over the same period as (b). Since our method is probabilistic, the private counts shown are the average over 100 iterations. (e) Box plot of the relative error of the moving average over these 100 iterations.

## DISCUSSION

Our system relies on a hybrid of data-agnostic and data-dependent spatial partitioning. Below, we discuss why our hybrid approach is better than any nonhybrid approach. Let one use a data-agnostic scheme alone, e.g., by using a fixed grid to partition the space—we refer to it as the naive approach. To achieve $\varepsilon$-DP, the naïve approach adds the noise (from Laplace distribution of mean zero and scale $1/\varepsilon$) to the aggregate count of each grid cell. However, to achieve the granularity supported by our system, the naïve approach must create grid cells of much smaller sizes, about 1 km $\times$ 1 km. Thus, the naïve approach compared to the hybrid approach, results in a huge number of cells and the data to be stored and processed every day. Moreover, although this approach gives a reasonable estimate for each cell, the answers to the queries that consist of many cells, e.g., for states, counties, or even large enough regions within a county have higher errors than the hybrid approach. This is because most of these cells will contain no repot but the noise introduced by the DP mechanism.

On the other hand, if one uses the quadtree approach alone, only one quadtree will have to be built over the USA. Now, to create the partitions with the level of granularity supported by our system, the max height of the tree would have to be much higher, which will lead to poor accuracy. This is because, at every level, a smaller privacy budget (i.e., the value of $\varepsilon$) will be available to compute the DP count for each partition. Thus, the magnitude of added noise will be higher. Note that while other data-dependent spatial partitioning approaches (e.g., k-d trees) may provide better accuracy when privacy is not considered, they perform worse when privacy has to be taken into account since some privacy budget will now be allocated for creating partitions. This will further reduce the privacy budget for computing DP counts, leading to higher noise, and thus, higher error.

One limitation of our approach—and in general of all privacy approaches—is the inability to limit the privacy risk under continual data updates. For instance, our approach incurs a privacy risk of $\varepsilon$ for a single report. However, when a user reports more than once and each of the reports is used to build a quadtree, then the privacy risk increases linearly with the number of reports. To limit this increase in the privacy risk, we devised a selection criterion to decide which reports by a user should be included; this makes the system usable for Covid-19 for practical purposes.

However, the following general problem remains open: "How to limit the privacy risk and achieve meaningful utility for an arbitrary number of SR queries, *when each user can potentially contribute one report per day*." There is a need to conceptualize a new privacy notion and methods specialized for this setting to solve this problem. While we presented our approach specifically to compute SR queries for 14 day-long period, our system can be used to compute other important insights. For instance, we can identify hotspots within a division (e.g., county in our case) by: 1) identifying highest leaves in the corresponding quadtrees; or 2) partitioning the division as per one's requirement and comparing the counts for these partitions; a similar strategy can be used to identify hotspots in terms of counties and states. Since the system builds and keeps a series of quadtrees over time, they can be used to compute SR queries for time periods other than 14 days. Additionally, dividing any SR query's answer by its time range, $d$, gives the $d$-day moving average for the query's region, e.g., the country, a state, a county, or a region within a county. We note that the series of quadtrees built by our system can be carefully combined to compute other insights which we plan to explicate in future work.

To use our approach in a similar future pandemic/epidemic, one needs to find the corresponding heuristic to select the parameter $n$ (for the covering algorithm). This can be done by generating synthetic data for the new daily cases by, for example, using the SIR model,[19,20] estimating $d$ (which in the case of Covid-19 is 14 days), and then performing a similar evaluation as has been done for Covid-19 in the article by Asif.[14] Furthermore, our approach is general and can be used to privately and accurately crowdsense other health symptoms, data, or adoption behaviors (e.g., vaccination rates) by using the corresponding data and following the approach as outlined in this article.

## CONCLUSION

The proposed privacy-preserving crowdsensing approach enables intelligent pandemic surveillance. It guarantees strong privacy for the data contributors and allows for accurately querying across arbitrary space and time bounds. Since the lack of privacy guarantees has been cited as a leading cause of concern by experts and nongovernmental organizations, the proposed approach can be vital to allaying the concerns of experts and end-users alike for future pandemic crowdsensing efforts. Its support for tracking across administrative boundaries is almost cognizant of the ground realities of the pandemic. Furthermore, the approach is generic and can be applied for reporting spatiotemporal information about other health symptoms or adoption behaviors (e.g., vaccination rates). Overall, this approach paves a way forward for countering pandemics without compromising on individual privacy.

## ACKNOWLEDGMENTS

This work was supported in part by the National Institutes of Health under Award R35GM134927, and in part by the National Science Foundation under Award CNS-2027789. The content is solely the responsibility of the authors and does not necessarily represent the official views of the agencies funding the research. This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Rutgers Arts and Sciences IRB under Application No. Pro2020000711.

## REFERENCES

1. C Menni *et al.*, "Real-time tracking of self-reported symptoms to predict potential covid-19," *Nature Med.*, vol. 26, pp. 1037–1040, 2020.
2. T Sharma and M. Bashir, "Use of apps in the covid-19 response and the loss of privacy protection," *Nature Med.*, vol. 26, pp. 1165–1167, 2020.
3. T. P. Koehlmoos, M. L. Janvrin, J. Korona-Bailey, C. Madsen, and R. Sturdivant, "COVID-19 self-reported symptom tracking programs in the United States: Framework synthesis," *J. Med. Internet Res.*, vol. 22, no. 10, 2020, Art. no. e23297.
4. Apps and covid-19, Privacy International. Accessed: Nov. 20, 2021. [Online]. Available: https://privacyinternational.org/examples/apps-and-covid-19
5. C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed! A survey of attacks on private data. Annual review of statistics and its application," *Annu. Rev.*, vol. 4, pp. 61–84, 2017.
6. J Vaidya, B Shafiq, X Jiang, and L. Ohno-Machado, "Identifying inference attacks against healthcare data repositories," *AMIA Summits Transl. Sci. Proc.*, vol. 2013, 2013, Art. no. 262.
7. N Buescher, S Boukoros, S Bauregger, and S. Katzenbeisser, "Two is not enough: Privacy assessment of aggregation schemes in smart metering," in *Proc. Privacy Enhancing Technol.*, 2017, pp. 198–214.
8. R. Wang, Y. F. Li, X. Wang, H. Tang, and X. Zhou, "Learning your identity and disease from research papers: Information leaks in genome wide association study," in *Proc. 16th ACM Conf. Comput. Commun. Secur.*, 2009. pp. 534–544.
9. A. D. Flouris and J. Duffy, "Applications of artificial intelligence systems in the analysis of epidemiological data," *Eur. J. Epidemiol.*, vol. 21, no. 3, pp. 167–170, 2006.
10. M. V. Marathe and N. Ramakrishnan, "Recent advances in computational epidemiology," *IEEE Intell. Syst.*, vol. 28, no. 4, pp. 96–101, Dec. 2013.
11. C Dwork, F McSherry, K Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proc. 3rd Conf. Theory Cryptogr.*, Berlin, Germany: Springer-Verlag, 2006, pp. 265–284. [Online]. Available: http://dx.doi.org/10.1007/11681878_14
12. COVID nearby, an NSF sponsored initiative by Rutgers University. COVID nearby, 2021. [Online]. Available: https://covidnearby.org
13. G Cormode, C Procopiuc, D Srivastava, E Shen, and T. Yu, "Differentially private spatial decompositions," in *Proc. IEEE 28th Int. Conf. Data Eng.*, 2012, pp. 20–31.
14. H. Asif, "Privacy or utility? How to preserve both in outlier analysis," Ph.D. dissertation, Rutgers Univ.-Graduate School-Newark, Newark, NJ, USA, Chapter 7, 2021.
15. R. J. Wilson, C. Y. Zhang, W. Lam, D. Desfontaines, D. Simmons-Marengo, and B. Gipson, "Differentially private sql with bounded user contribution," in *Proc. Privacy Enhancing Technol.*, 2020, pp. 230–250.
16. E Dong, H Du, and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time," in *The Lancet Infectious Diseases*. Amsterdam, The Netherlands: Elsevier, 2020, vol. 20, no. 5, pp. 533–534.
17. Bureau UC. Memorandum, "Design parameters and global privacy-loss budget," The United States Census Bureau, Suitland-Silver Hill, MD, USA, 2019.
18. G Lebanon and L. J. Cranking, "Combining rankings using conditional probability models on permutations," in *Proc. Int. Conf. Mach. Learn.*, 2002. pp. 363–370.
19. H. W. Hethcote, "Qualitative analyses of communicable disease models," *Math. Biosci.*, vol. 28, no. 3/4, pp. 335–356, 1976.
20. S. Ahmetolan, A. H. Bilge, A. Demirci, A. Peker-Dobie, and O. Ergonul, "What can we estimate from fatality and infectious case data using the susceptible-infected-removed (sir) model? A case study of covid-19 pandemic," *Front. Med.*, vol. 7, 2020, Art. no. 570.

**HAFIZ ASIF** is a Postdoctoral Associate with Rutgers Institute for Data Science, Learning, and Applications, New Brunswick, NJ, USA. His research interests include the areas of privacy, security, machine learning, and algorithmic fairness. He received the Ph.D. degree in information systems from Rutgers University, Camden, NJ, USA. Contact him at hafiz.asif@rutgers.com.

**PERIKLIS A. PAPAKONSTANTINOU** is an Associate Professor with the Department of Management Sciences and Information Systems, the Rutgers Business School, Newark, NJ, USA. His research interests include theory of computing at large and related applications, such as in cryptography. Before Rutgers, he spent six years as an Assistant Professor with Andy Yao's institute, Tsinghua University, Beijing, China. He took up that appointment immediately after completing the Ph.D. degree from University of Toronto, Toronto, ON, Canada. Contact him at periklis.research@gmail.com.

**STEPHANIE SHIAU** is an Assistant Professor with the Department of Biostatistics and Epidemiology, Rutgers School of Public Health. After graduate school, she completed a Postdoctoral Research Fellowship with the Gertrude H. Sergievsky Center, Columbia University. She holds a Certified in Public Health (CPH) credential. Her interdisciplinary research program focuses on the effects of HIV and its treatment over the life course, seeking to identify modifiable factors that influence trajectories of comorbidities in children, adolescents, and adults living with HIV and affected by HIV. She received the Ph.D. degree and an MPH in epidemiology from Columbia University, New York, NY, USA, and the B.A. degree in public health studies from The Johns Hopkins University, Baltimore, MD, USA. Contact her at stephanie.shiau@rutgers.edu.

**VIVEK SINGH** is an Associate Professor with the School of Communication and Information, and the Director of the Behavioral Informatics Laboratory, Rutgers University, Camden, NJ, USA. His research lies at the intersection of computational social science, data science, and multimedia information systems. Before joining Rutgers, he was a Postdoctoral Researcher with the MIT Media Laboratory. He received the Ph.D. degree in information and computer science from the University of California, Irvine, CA, USA. Contact him at v.singh@rutgers.edu.

**JAIDEEP VAIDYA** is a Professor of computer information systems with Rutgers University, Camden, NJ, USA, and is the Director of the Rutgers Institute for Data Science, Learning, and Applications, New Brunswick, NJ. He has authored or coauthored more than 190 papers in international conferences and journals. His research interests include privacy, security, and data management. He is an IEEE Fellow, an ACM Distinguished Scientist, and is the Editor-in-Chief of IEEE Transactions on Dependable and Secure Computing. Contact him at jsvaidya@business.rutgers.edu.