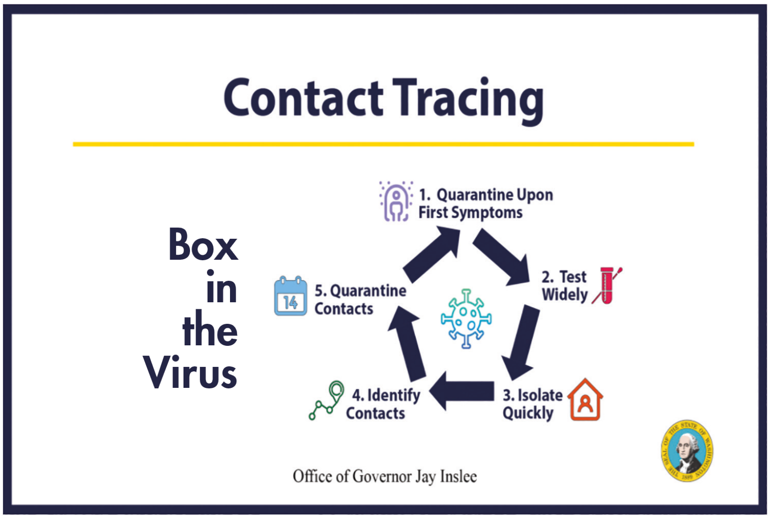
[O Privacy, Where Art Thou?]

Aleksandra Slavkovic and Fang Liu
Column Editors

Privacy Risk and Preservation in Contact Tracing of COVID-19

Dong Wang and Fang Liu



Contact Tracing

Box in the Virus

1. Quarantine Upon First Symptoms

2. Test Widely

3. Isolate Quickly

4. Identify Contacts

5. Quarantine Contacts

14

Office of Governor Jay Inslee

Figure 1. Contact Tracing (courtesy of Washington Governor's Office).

The World Health Organization (WHO) declared the coronavirus or COVID-19 outbreak a public health emergency of international concern on January 30, 2020. Since the first COVID-19 case on November 17, 2019 (according to unpublished government data), the number of cumulative cases worldwide has been around 3.8 million and 345,000 had died from the disease as of May 25, 2020.

A huge amount of data have been and are being collected during the pandemic, and will be in the future. The data, coupled with state-of-art computing and analysis techniques, play a powerful role in the efforts to harness the spread of COVID-19. However, the collected data, including health and medical history, mass surveillance, contact tracing, and social control, often contain personally identifiable information and are at high risk for compromised data privacy.

The EU Parliament has said, "These tools could seriously interfere with people's fundamental rights to a private life and the protection of personal data, and are tantamount to a state of surveillance of individuals." How to balance private protection versus personal data collection and release for monitoring the pandemic and improving public health has attracted much research interest and will be a subject of continuing debate. Government, academia, and industry are already working together to search for effective solutions to this problem.

A variety of types of information has been collected during the COVID-19 pandemic that can lead to privacy concerns. Depending on data type, the approaches and measures taken to mitigate privacy concerns can be different. The privacy issues incurred by collecting and sharing location and contact tracing data have their own importance.

Contact Tracing in the COVID-19 Pandemic

Contact tracing in the COVID-19 pandemic uses digital tools to trace and monitor contacts of infected people during the epidemic to

Table 1—Examples of COVID-19 Contact-Tracing Apps and Software

		technology		
		GPS	Bluetooth	GPS+Bluetooth

Data-collection and information-sharing model	Centralized	Alipay Health Code (China);
		WeChat (China);
		Corona 100m (South Korea);
		CovidTracker (Thailand);
		ProteGo (Poland)
			Tracetogether (Singapore)	Aarogya Setu (India)

	Decentralized	safe paths (U.S.)
		HaMagen (Israel)
			Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) (EU);
			COVID watch (mainly U.S.);
			PACT (U.S.);
			COVIDsafe (Australia)	

alert and inform people who have come into contact with them, and help ensure effective quarantine of contacts to prevent additional transmission. Washington Governor Jay Inslee used Figure 1 when announcing the state's contact tracing plan; it provides a summary of how contact information is used and what it can do.

While contact tracing has proved useful in tracking and slowing down the spread of COVID-19 and plays an important role in fighting the pandemic, major newspapers such as the *Washington Post* and *Forbes*; the Reuters news agency; and government agencies have raised red flags about the high privacy risk associated with this process. The information collected during contact tracing often includes detailed and frequent location data that lead to inferences about the private social life and health status of individuals. Location is known to be highly revealing of people's identity. For example, De Montjoye, et. al. (2013) published a study of 1.5 million individuals over a period of 15 months that found four spatial-temporal mobility data points are enough to identify 95% of the individuals involved.

Countries around the world have developed and deployed contact-tracing software or mobile apps, with different levels of alertness to the privacy issues. Table 1 provides some examples of such apps and software, categorized by the technology used and the degree to which authorities are involved in data collection and information sharing. Contact-tracing apps and software that use GPS data collect users' location data, whereas the Bluetooth-based techniques mostly just requires the relative tempo-spatial proximity among users. In that sense, less private information is collected in the Bluetooth-based approaches than in the GPS-based approaches. In either the GPS- or the Bluetooth-based approaches, the centralized or non-centralized models can be deployed to collect and store data, share information, and alert users regarding potential COVID-19 exposure. However, the two models differ in the levels of anonymity and in the approaches to achieve privacy protection for the data contributors.

In the centralized model, contact-tracing data are collected, integrated, and shared with targeted individuals by some authorities (e.g., health authorities or federal, state, or local governments). In this sense, the centralized model operates like a mass surveillance system; data are collected from everyone, whether healthy or diseased, and the authorities have the unique identifiers for all the individuals and know whom to target with certain information. There is no privacy for the users in terms of information sharing with the authorities and they have to trust the authorities to keep their data safe and private.

In contrast, with the decentralized model, there is no need to collect or store information about everyone through a central server. Location and contact information of those who are not tested or who tested negative are stored and processed locally on their respective devices, and they can choose to check whether they have crossed paths with infected people through public platforms like a website built by authorities that contains COVID-19 hot spot information. Often, the information shared on such a website has already gone through some types of data anonymization or blurring through careful planning or integrating a formal privacy concept.

In summary, the decentralized model offers a higher level of privacy protection on individuals compared to the centralized models. Table 1 also suggests the centralized model is mainly employed by Asian countries, whereas the decentralized model is preferred by the US and European countries. In what follows, we will look into how the centralized and decentralized models work in the GPS and Bluetooth based technology, respectively, in more details.

GPS-Based Contact-Tracing Schemes

GPS-based apps collect time-stamped GPS points from individuals on a 24/7 basis. If the collected GPS data suggest that two people were in close proximity to each other at a certain time, and one of them tests positive for COVID-19 later on, then the other person will either receive notifications from authorities regarding the contact event or find this out by checking publicly posted contact tracing information from authorities themselves, and will be subject to self-quarantine. The way that the authorities collect and share information leads to either the centralized or de-centralized model.

The Alipay Health Code from China is an example of the centralized model. The Alipay Health Code assigns an individual a color QR code (green, yellow, red), representing the individual's health status. A green code indicates the highest of level of healthiness and the individual is allowed to go anywhere unrestricted; red stands for high risk and requires a two-week quarantine, and yellow means a one-week quarantine.

The determination of the color code is often based on the location history of the individual. If the person has been to a COVID-19 hot spot, then there is a non-ignorable chance the person may be infected and is likely to receive a red or yellow code. Each time the individual’s QR code is scanned, the information regarding the current location is sent to servers belonging to some authorities, allowing the authorities to track people’s movements over time. Furthermore, the app often requires users to register with unique identification information, such as national identification number/Social Security number, name, and phone number.

Similarly to China, South Korea developed the Corona 100m (Co100) app as its centralized model. The app uses government-collected location data to alert users when they come within 100 meters of a location recently visited by a COVID-19 patient.

Safe Paths is an MIT-led privacy-preserving platform and an example of the decentralized model. It comprises a smartphone app, PrivateKit, and a web application, SafePlaces. SafePlaces shares anonymized and blurred location histories of infected people, while PrivateKit allows users to match their personal location history with the shared information on SafePlaces.

In other words, healthy individuals keep their own location diaries without having to share with or report to authorities. However, once an individual tests positive for COVID-19, their location history information will be reported to the authorities. Since the location history contains private information, some of which is not even relevant to COVID-19 tracing (e.g., the home location where the individual spends most of their time), the information is often redacted or blurred before being placed on SafePlaces, where users can compare their location diaries with those infected to see if they have ever crossed paths.

Israel developed the Hamagen app (hamagen is Hebrew for shield), which has a similar basis to the Safe Paths platform. It allows local comparison of users’ GPS data with the government epidemiological location database of COVID-19 hot spots.

Figure 2 shows the centralized and the decentralized models in the GPS-based contact-tracing scheme. The centralized model tracks location, contact information, and health status with unique identifiers from both patients and healthy users.

The potential privacy risk in this scheme is obvious. First, it can be tricky to keep the identity of the infected people confidential in some cases when broadcasting their location history within several weeks before a diagnosis, especially when an infected person is one of the few with whom the healthy people interacted

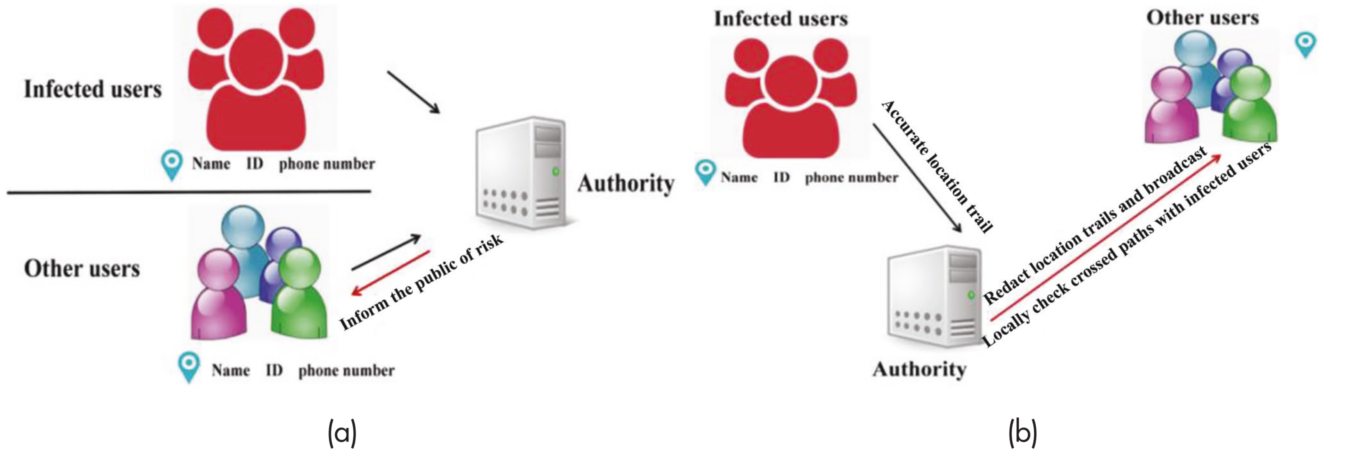


Figure 2. GPS-based contact-tracing schemes: (a) example of GPS-based centralized model; (b) example of GPS-based decentralized mode.

in close proximity recently. Second, it personalizes the alert and notification systems. This level of precision comes at the cost of compromised individual privacy.

Individuals whose data are collected in the centralized model reply on the authorities to keep their data private and safe, but this trust is not always warranted. The decentralized model only collects location information and shares an anonymized version of that information from reported COVID-19 patients. Therefore, it provides a higher level of privacy protection for patients. Furthermore, the decentralized model does not contact-trace healthy people and only shares the information on COVID-19 hot spots in a public forum with no specific targeting of certain individuals.

In other words, everyone can go to the website or platform to see where the hot spots are without having to register their personal information, so the decentralized model presents minimal privacy concerns for healthy people. On the other hand, without personalized alerts, the decentralized model would reply on users' self-initiation and pro-activity to go to the public information-sharing forum and check whether they might have been to any of the infection hot spots recently, and to self-quarantine if that is the case.

Formal notations on privacy guarantee can be incorporated in both the centralized and decentralized models when developing the GPS-based contract tracing apps and software. For example, the *k* - anonymity model introduced by Sweeney in 2002 can be used to collapse detailed location information or pseudo-identifiers to yield at ≥*k* "homogeneous" individuals in each of the cross-tabulations of these attributes. Geo-indistinguishability is a formal location privacy concept proposed by Andrés, et. al., in 2013, that extends the popular differential privacy concept by Dwork, et al., in 2006, and can be used to generate sanitized location information.

Regardless of which formal privacy notation is used to sanitize the data before releasing and sharing information with the public or targeted people, the accuracy of contact tracing will more or less be affected.

Bluetooth-Based Contract-Tracing Scheme

Unlike the GPS-based privacy-preserving scheme, the Bluetooth-based contact-tracing apps do not collect exact location information from their users, so users may feel more private and less anxious about their whereabouts being monitored 24/7. Bluetooth also has higher contact-tracing accuracy than GPS-based apps—Bluetooth signals do not rebound and pass through most soft walls, helping to avoid the false positive when two people in close proximity are regarded as a "contact" event when they are actually separated by walls.

A Bluetooth-based contract-tracing scheme leverages the Bluetooth technology to collect information about whether two people have appeared in the same location within 6 feet of each other at the same time. Each app user generates a time-varying sequence

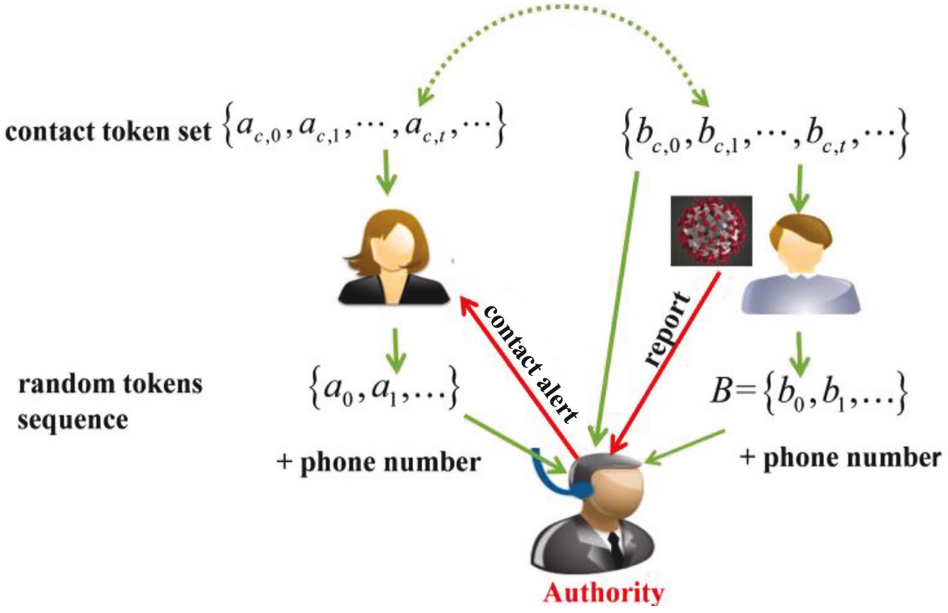


Figure 3. Example of Bluetooth-based centralized model.

of random tokens, which are stored locally on their devices. The time interval between two tokens cannot be so frequent that it causes computational or storage difficulty for the users, and nor so infrequent that it makes the tracing ineffective or incur privacy concerns.

If two users appear within 6 feet of each other at a time *t*, they exchange their tokens at that time, which are stored in their contact token sets. If one user is diagnosed with COVID-19, say within two weeks after the contact event, they will share their contact token sets from the last two weeks with health authorities, who will subsequently develop alert and notification systems to notify people who might have shared a contact event with the infected person. The way that the authorities collect and share information leads to either the centralized or de-centralized model.

Singapore's Bluetooth-based mobile phone app TraceTogether is an example of the centralized approach. Figure 3 shows that in this model, all app users report their tokens (denoted by $\{a_0, a_1, \cdots\}$ and $\{b_0, b_1, \cdots\}$), as well as their phone numbers, to the authorities regardless of their health status. If a person is diagnosed with COVID-19, they update the authorities about their health status and shares their contact tokens (denoted by $\{b_{c,0}, b_{c,1}, \cdots, b_{c,t}, \cdots\}$). The authorities then match each token in the contact token set with its database of tokens, and alert the users with matches through their phone contacts.

The privacy risk for infected individuals is similar to the centralized model in the GPS-based scheme. In addition, since the authorities have each user's phone number, which is a unique identifier, it can be used to link to other databases that might contain sensitive information about the users, if the authorities feel there is a need to do so. Similarly to the centralized model in the GPS-based system, users have no choice but to trust the authorities will keep their information safe and private.

The Covid Watch, Private Automated Contact Tracing (PACT), and COVIDsafe apps and the Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) software are examples of decentralized approaches that leverage Bluetooth technology. The Covid Watch app represents an international effort from more than 400 volunteers around the world (U.S., Canada, Australia, etc.), and sends anonymous privacy-preserving COVID-19 exposure alerts via private and local Bluetooth signals. PACT was developed by MIT, working with partners from around the world, to collect information about not only binary contact events but also the distance and time duration of a contact event.

COVIDsafe is the app used by the Australian government for

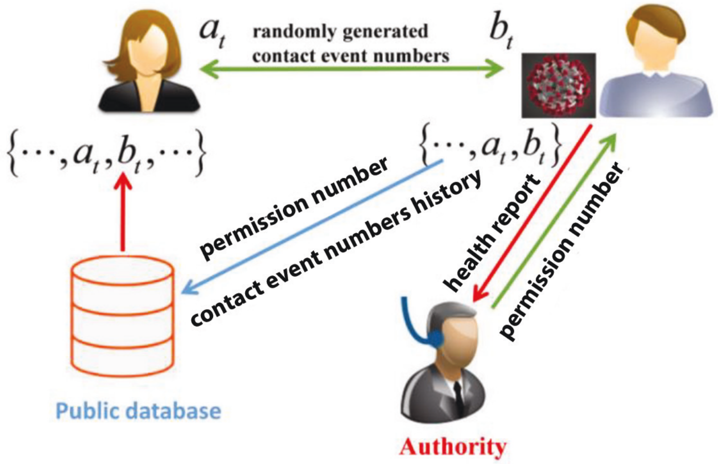


Figure 4. Example of Bluetooth-based decentralized model.

contact tracing. PEPP-PT is a larger software system with many individual components that sends alerts about possible exposures.

In the decentralized model (Figure 4), the authorities only collect the COVID-19 patients' tokens. A person who is diagnosed with COVID-19 receives a permission number from the authorities, which is then shared with a public database, together with their history of contact event numbers. The public database verifies the permission number and updates itself with the contact numbers. Other users can compare their contact event numbers with the publicly posted contact event numbers. A match indicates that they may have been exposed to the virus and need to self-quarantine.

For the Bluetooth-based privacy-preserving contract tracing, anonymity can be enhanced by randomly swapping generated random tokens among users to better prevent linkage to privacy attacks. Cryptography solutions for privacy protection, such as the technology developed by Apple and Google, use secure multi-party computation without relying on a trusted server, or sending anonymous encrypted or random messages, as proposed by Cho, et al.; Hekmati, et al.; and Reichert, et al., in March and April 2020.

Final Remarks

In addition to the technology-based approaches for privacy protection in contact tracing, some general principles for data privacy protection apply in the COVID-19 pandemic data collection and information sharing. For example, data collection and release regarding COVID-19 should be guided by necessity, proportionality, and transparency.

It is often permissible to share anonymized data or aggregated statistics that are associated with low individual re-identification risk. If a risk of non-ignorable re-identification exists or there is a need to reveal individual identity when releasing information, there must be a justification for doing so. Minimizing data collection, limiting access, and retaining data only for the minimum amount of time that is necessary also help to reduce privacy harms due to COVID-19 data processing.

Obtaining consents is also commonly used for privacy protection. The subjects from whom data are collected and shared should receive clear communications from authorities regarding the purposes and usage, and the retention duration, of their data, among other details. Given the unprecedented situation of COVID-19, the consent might have to take a form than is different from the regular consents when it comes to data sharing, especially when an individual feels compelled to share their contact and location history after testing positive.

We believe authorities and researchers should be committed to privacy preservation in the contract tracing of COVID-19, now and in the future. All parties (the public, authorities, academia, and industry) should work together to develop effective policies and technologies to protect the privacy of the people when collecting data about COVID-19 to help curb the global pandemic.

Further Reading

COVID 19 Pademic

WHO declared the coronavirus disease. 2019. *https://bit.ly/2Po6IQx*.

First covid-19 case happened in November, china government records show-report. *https://bit.ly/2Dx2WBG*.

COVID-19 pandemic. *https://en.wikipedia.org/wiki/COVID-19_pandemic*.

Formal Privacy Concepts

Dwork, C., McSherry, F., Nissim, K., and Smith, A. 2006. Calibrating noise to sensitivity in private data analysis. in Theory of cryptography conference, 265–284.

Andrés, M.E., Bordenabe, N.E., Chatzikokolakis, K., and Palamidessi, C. 2013. Geo-indistinguishability: Differential privacy for location-based systems. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, 901–914.

Sweeney, L. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 557–570.

Privacy on Location Data and Contact Tracing

De Montjoye, Y.-A., Hidalgo, C.A., Verleysen, M., and Blondel, V.D. 2013. Unique in the crowd: The privacy bounds of human mobility. *Scientific Reports* 3:1376.

Contact tracing apps can help stop coronavirus. But they can hurt privacy. *https://wapo.st/2EUqLEp*.

Coronavirus Contact-Tracing Apps Miss the Point About Privacy. *https://bit.ly/3guAoY3*.

UK privacy advocates warn over covid-19 contact tracing app. *https://reut.rs/2XvbYpQ*.

Use of smartphone data to manage covid-19 must respect EU data protection rules. *https://bit.ly/2PBOVFP*.

GPS-Based Contact Tracing of COVID-19

What the US can learn from other countries using phones to track covid-19. *https://bit.ly/3k9BBWZ*.

In coronavirus fight, China gives citizens a color code, with red flags. *https://nyti.ms/31pUwoa*.

South Korea to step-up online coron-avirus tracking. *https://bit.ly/3fvfjve*.

Covid-19: Poland launches an official tracking app. *https://bit.ly/39YgLoI*.

Covid-19 news tracker-location-based news about covid-19 in Thailand. *https://bit.ly/2Dvx2pc*.

HaMagen: The Ministry of Health App for Fighting the Spread, *https://bit.ly/2Dvxa8a*.

Sawant, N. 2020. Aarogya setu: Whether we like it or not, the app is here to stay, but it's still riddled with privacy issues that need strong answers. *https://bit.ly/33rwbAU*.

COVID watch. *www.covid-watch.org*.

Safepaths. *http://safepaths.mit.edu*.

Bluetooth-based Contact Tracing of Covid-19

Tracetogether, safer together. *www.tracetogether.gov.sg*.

Pact: Private automated contact tracing. *https://pact.mit.edu*.

Pan-European Privacy-Preserving Proximity Tracing (PEPP-PT) (EU). *https://www.pepp-pt.org*.

Australian Government Department of Health COVIDSafe app. *https://bit.ly/2Pq0i3y*.

Aarogya setu: Whether we like it or not, the app is here to stay, but it's still riddled with privacy issues that need strong answers. *https://bit.ly/3fuaxhI*.

Privacy-preserving contact tracing. *www.apple.com/covid19/contacttracing*.

Cho, H., Ippolito, D., and Yu, Y.W. 2020. Contact tracing mobile apps for COVID-19: Privacy considerations and related trade-offs. *arXiv preprint 2020, arXiv:2003.11511*.

Hekmati, A., Ramachandran, G., and Krishnamachari, B. 2020. CONTAIN: privacy-oriented contact tracing protocols for epidemics. *arXiv preprint, arXiv:2004.05251*.

Reichert, L., Brack, S., and Scheuermann, B. 2020. Privacy-preserving contact tracing of covid-19 patients. *https://eprint.iacr.org/2020/375.pdf*.

About the Authors

Dong Wang is working toward a PhD degree in computer science at the State Key Laboratory of Engineering in Surveying, Mapping, and Remote Sensing of Wuhan University, China. She is currently a visiting PhD student in the Department of Applied and Computational Mathematics and Statistics at the University of Note Dame. Her research interests include data privacy and data mining. She would like to thank the China Scholarships Council program (No. 201906270230) for supporting her work.

Fang Liu is a professor in the Department of Applied and Computational Mathematics and Statistics at the University of Notre Dame. She obtained her PhD in biostatistics from the University of Michigan, Ann Arbor. Her research interests include data privacy and differential privacy, statistical machine learning of Big Data, model regularization, Bayesian methodology, and analysis of missing data. She would like to thank the National Science Foundation (Grant #1717417) for supporting her work on data privacy.