COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Jaideep Vaidya (Rutgers University-Newark), April 14, 2021

Title: RAPID: Privacy-Preserving Crowdsensing of COVID-19 and its Sociological and Epidemiological Implications

Jaideep Vaidya CIC Database Profile

NSF Award #: 2027789

YouTube Recording with Slides

April 2021 CIC Webinar Information

Transcript Editor: Macy Moujabber

Transcript

Jaideep Vaidya:

Slide 1:

Hello everyone. I'm Jaideep Vaidya. I'm the director of the Rutgers Institute of Data Science, Learning, and Applications and this is essentially our response to the pandemic, you know what the institute tried to do to help with respect to this. And this is joint work with my colleagues in the School of Public Health and the School of Communication and Information at Rutgers.

Slide 2:

So, I did want to talk about one of the key problems that we saw essentially all through last year as the pandemic sort of unfolded, and that was the issue with insufficient testing and preparedness. And effectively this was visible all, you know- this was reported on by all of the news media, whether it was conservative or liberal. In fact, the New York Times put this up you know is your state doing enough coronavirus testing and the answer, of course, was no at the start of the pandemic, but for quite a long period of time all the way through the summit as well, and even the Wall Street Journal sort of reported that, hey, we really need to carry out sufficient amount of sort of COVID testing in order to bring the economy back on track.

Slide 3:

So, you know from a technology perspective, technology tried to sort of come to the rescue with respect to this, given the lack of resources. And what happened was we saw quite a few apps like "COVID Near You" or "How We Feel" that came up which were basically symptom tracking apps you would report your symptoms and then that information would be aggregated and reported back. Even the CMU app, which Facebook put out, had this interactive map and dashboard. So, you would put in your information. It would then be aggregated and then released, essentially.

Slide 4:

So, one huge problem that we saw with respect to this was privacy. So essentially, giving in this information, there was a lot of concern, and correctly so, regarding what would happen to your privacy as you reported such information. And in fact, there was a second problem as well- that the apps were basically pre-defining regions. You would get reports for your city, for your state, for the county, but at fixed sort of levels.

Slide 5:

So, what we wanted to look at- we wanted to look at two different things. One was to answer this question of specific areas. If you are interested in your local neighborhood or you are interested in the particular region of interest, what was happening with respect to the symptoms? So, suppose you have this question: how many people reported feeling symptoms in the past eight days in New York City? Or you wanted to look at your local neighborhood or across the state, what have you. So, we wanted an arbitrary number of queries referring in both sizes and regions. I could ask for this, I could ask for this, maybe the whole map over there, something small over there, or what have you. But you quickly realize that once you allow these sorts of queries, you know any number of such queries, the privacy problem actually gets way more magnified as well. So essentially no sort of simple aggregate and release mechanism is ever going to be able to fully protect your privacy. In fact, with a very small number of queries, I can quickly figure out if- you know, particular information regarding you or whether or not you are contributing your data to a survey like this.

Slide 6:

So, what we came up with was actually an app and a framework called "COVID Nearby" and basically what we are providing is a formal guarantee called differential privacy. This is a state-of-the-art model that the U.S. census is using- in fact companies are using as well and it essentially guarantees- I won't go into the match but intuitively speaking, it guarantees that whether or not you put in your information into any such app or any such collection of data, any survey, the- your risk, or the risk to your identity is not magnified. You don't have a larger risk simply because you're part of the survey. It's going to be essentially very similar to- even if you're not sort of participated as part of this. So, what we wanted to use was differential privacy and this is a formal guarantee, it's very nice because it guarantees your privacy against privacy attacks today, but also privacy attacks that may be invented way ahead in the

future which we know nothing about. One problem however is that once you try to do this especially over such spatiotemporal data and / or dynamic data, it is very difficult to provide sort of utility, good answers, while ensuring privacy even if you are doing sort of differential privacy. So, what we did was we came up- and I don't have time to get into this- but we came up with a unique way of representing the data. So we're still we created this app that will let you put in your symptoms and that will be stored in a secure database so we are acting as a trusted curator here, but then there's an intermediate representation that is built over this data, which is then used to answer any number of questions over the data to ensure that no matter- you know, even if you ask like a thousand questions, the privacy risk for anyone does not keep on increasing. It's basically bounded over there.

Slide 7:

And this app has actually been developed, so there were a bunch of challenges as I said from the technical level. It's spatio-temporal data. It keeps changing. You have an arbitrary number of queries and actually if you want people to use it, it really needs to be able to work fast as well. The results have to be usable, and they have to work actually for the specific situation at hand, in this case the COVID pandemic. And surprisingly, you know apart from the technical challenges, one of the biggest challenges was with the rapid development effort, getting it certified by the Play store and the [Apple] App store so that we could have it out there. But we managed to do that and actually the app is out there and if you have an Android phone or iOS you know Apple device you absolutely can get it and put in your data through it with a guarantee of privacy. One thing that I wanted to point out- we have looked at this data and, yes, we've gone through the IRB [Institutional Review Boards] and all of that. The good news is that you can actually get very close to the non-private results using these techniques as well. So, these bars that you see are basically the non- the original data and then the private data. The stack bars sort of give you the idea that they are roughly the same, but just to give you a better view on a particular date, for example, you can see the heat map for the private versus the original, and as you can see basically the relative ranking, not just at the county level but even within counties, it turns out to be exactly the same. So again, I don't have time to go into the details of this, but in terms of hot spot tracking or ranking it's very easy to sort of get exactly similar results, and you don't lose out on accuracy by sort of doing this.

Slide 8:

So, with that, I'll basically stop and basically say, yes, you should use the app and the links are there if you want to use it on your Apple device or your Android device and I'll stop right there. Thanks.