COVID Information Commons (CIC) Research Lightning Talk

Transcript of a Presentation by Murat Kantarcioglu (University of Texas at Dallas), October 16, 2020

Title: RAPID: Collaborative: A Privacy Risk Assessment Framework for Person-Level Data Sharing During Pandemics

Murat Kantarcioglu CIC Database Profile

NSF Award #: 2029661

YouTube Recording with Slides

October 2020 CIC Webinar Information

Transcript Editor: Macy Moujabber

Transcript

Murat Kantarcioglu:

Slide 1

Good- I think now it's good afternoon for some. My name is Murat Kantarcioglu. I'm a professor of computer science at the University of Texas at Dallas. This is our RAPID work on privacy risk assessment for data sharing during pandemics with professor Brad Malin from Vanderbilt Medical School.

Slide 2

So as many of you follow the news, as the COVID-19 progresses there are new questions coming up with respect to whether it's related to race, whether smoking for example increases your risk with COVID-19, bad outcomes, etc. So therefore, in order to answer this kind of questions whether it's relationship of race and COVID-19 mortality or other factors such as smoking high, blood pressure. Clearly, we need to collect more data and of course this data needs to be analyzed.

Slide 3

But unfortunately, the- or fortunately in some sense, the secrecy and privacy requirements may slow this attempt If you look at this Science article published in July, you would see that California didn't want to share- California state officials didn't want to share the COVID-19 data details with some researchers, hindering some certain research.

Slide 3

So, in one sense they're right because it's been known that shared data can be easily re-identified. It's been shown in the past many years ago that if you combine different data sets, for example, if you have a COVID-19 data that reports zip code, age, and gender, there is a higher risk of linking this with water registration lists and identify the people who are in the COVID-19 patient data. And of course, this has been known for a while so it's not a new problem and understanding these risks has been active research over the past twenty years. But still, COVID-19 changed many things.

Slide 4

One thing is that now we have, especially with the social media, we have different re-identification sources. This is an example from *The New York Times* where they profiled many patients who passed away due to COVID-19 in their life stories.

Slide 5

So, if you go back to this attack scenario, now you may have additional source where you know some people have passed away and their information and you can use it to remove the targets and rerun the identification attacks.

Slide 6

So therefore, this RAPID is really trying to understand the unique challenges due to pandemic data sharing, especially re-identification risks. So, one thing we realize is that unlike the past work, the attributes that are needed to be shared such as smoking or race can be changing over time. So, we need to really understand that. Also, case numbers are changing, which means that some privacy risks may be changing over time, so we need to really update the privacy risks over time as new data sources are being shared with others. So therefore, this rapidly changing and uncertain behavior of pandemic data sharing impacts the risk assessment frameworks, and for that reason we are now developing and updating existing models.

Slide 7

So, our goal is to really look into what's the best utility we can achieve given the data privacy goals, which in this case to make sure that re-identification risk is low in other words it's very hard to reidentify people. And also, what's the best possible privacy we can provide given the data utility goals? And again, our goal is to give as accurate information as possible. The second option comes into play because there is this argument that, you know, we should really focus on combating the disease so data privacy may be secondary. Still, our past research showed that given your goals with respect to data sharing different alternatives may have different privacy outcomes. So, understanding that would be important. Of course, the major direction would be- what's the data utility and how to define it and that's our ongoing research.

Slide 8

To conclude, we built our own past work on this analyzing identification risk and we update and develop techniques based on this unique challenge of time-varying and changing information due to pandemic. And soon, hopefully, like in a month or so, we would have a paper on giving general data sharing policies to guide the health officials like what data can be shared with low risk, what are the higher risk options, and its impact. So please follow us if you have any concerns about data sharing and privacy of pandemic data. So, I'll stop here I guess a little bit over time. Thank you.