Centro de Información de COVID (CIC): Charlas científicas de relámpago

Transcripción de una presentación de Murat Kantarcioglu (Universidad de Dallas en Texas), 15 de abril de 2022

Título: Colaboración: Un marco de evaluación de riesgos de privacidad para el intercambio de datos a nivel de persona durante pandemias

Perfíl de Murat Kantarcioglu en la base de datos de CIC

Subvención de La Fundación Nacional de Ciencias (NSF, por sus siglas en inglés) #: 2029661

Grabación de YouTube con diapositivas

Información del seminario web del CIC de abril 2022

Editora de la Transcripción: Saanya Subasinghe

Editora de la Translación: Isabella Graham Martínez

Transcripción

Murat Kantarcioglu:

Diapositiva 1

Muchas gracias por invitarme a volver. Así que cuando dimos la primera charla acabábamos de empezar este proyecto. Ahora vamos a concluirlo pronto. Así que estoy feliz de poder tener, supongo, el registro de lo que es nuestra visión, y lo que hemos hecho. Así que voy a hablar de un trabajo específico que surgió de este proyecto RAPID sobre cómo compartir datos públicos, datos de salud pública, mientras se preserva la privacidad durante la pandemia.

Diapositiva 2

Así que está claro que compartir la vigilancia para una respuesta basada en datos es muy importante. Utilizamos los datos para comprender cómo se produce la transmisión. Los datos se utilizan para estimar diferentes intervenciones, cuál sería su impacto, y también, por supuesto, en muchos casos, y para futuras pandemias, lo necesitaremos para detectar brotes tempranos.

Diapositiva 3

Ahora la pregunta es si podemos compartir directamente estos datos, especialmente si podemos compartir directamente datos individuales a nivel de paciente, lo que sería útil para construir diferentes modelos. Durante la crisis de Coronavirus, también tuvimos algún tipo de crisis de datos en cierto sentido. Las organizaciones eran reacias a compartirlo y les preocupaba la privacidad, con razón. Y esto les requirió analizar cuidadosamente, y de una manera que consume mucho tiempo, qué datos se van a compartir, en qué formato y cómo se pueden hacer públicos para su uso futuro.

Diapositiva 4
Por lo tanto, y esto es y uno de los desafíos de composición, es que a diferencia de la configuración tradicional de intercambio de datos, tenemos un tamaño de conjunto de datos que está cambiando cada día. Porque cada día podemos tener nuevos pacientes que pueden ser diagnosticados y los datos de estos nuevos pacientes pueden necesitar ser compartidos. Y también para fines de regulación, hay desafíos adicionales si cumple con la HIPAA, que es la legislación de privacidad que rige los datos de atención médica. Así que algunas personas se sienten muy cómodas con las reglas de puerto seguro HIPAA, que garantiza cierta forma regular de desinfectar los datos, pero las fechas no están permitidas, y luego esto crea desafíos para algunos de estos usuarios finales. Y, por supuesto, debido a la legislación de emergencia tenemos que hacer esto muy rápido. Tenemos que compartir los datos rápidamente sin realmente relacionados con la privacidad.

Diapositiva 5
Así que, en cierto sentido, desarrollamos un marco en el que podemos adaptarnos a este número de registros, número de registros de pacientes que están cambiando cada día. Y podemos priorizar información específica diferente. Digamos que quieres ser más detallado con respecto a la edad, pero no menos raza, pero tal vez quieres ser más detallado en la raza, etc., mientras que la comprensión de las implicaciones de privacidad.

Diapositiva 6
Así que desarrollamos esta estimación de riesgos - marco de estimación de riesgos de privacidad. Y la primera parte de esto es que analizamos la generalización de los datos. En este trabajo, nos centramos en las herramientas donde compartimos los datos correctos reales que se están dando, pero a un nivel menos específico o más generalizado.

Diapositiva 7
Entonces, ¿qué significa la marginación? Es que, por ejemplo, en nuestro marco en lugar de por razones de privacidad, en lugar de compartir la edad de alguien puede compartir el rango de edad. Como dice, de cinco a diez' o puedes, si quieres proteger la privacidad aún más, puedes compartir un rango más alto y puede subir a la cima donde no compartes ninguna información. Por supuesto, los nodos de la hoja [?] son muy precisos, pero los problemas de privacidad más potencial, por lo que menos protección de la privacidad. Y cuando subimos, menos información pero más protección de la privacidad.

Diapositiva 8
Así que la segunda cosa es que para estimar los riesgos realmente miramos la distribución de la población en diferentes condados y si, especialmente para el riesgo que estimamos en este trabajo, se llama riesgo de re-identificación. En otras palabras, un atacante que sabe algo de información sobre los pacientes - ¿pueden volver a identificar los datos y saber que: 'oh oh, este registro debe pertenecer a Murat' o 'el segundo debe pertenecer a John. Así que para hacer esta estimación miramos los datos del censo y usamos la distribución de la población para identificarla. El siguiente ajuste es que una vez que obtengamos estos datos, los casos de series temporales, como cuántos casos reportados, la métrica de riesgo de privacidad usaremos una específica que describiré en un segundo. Y también con qué frecuencia, a lo que llamamos ciencia de las ventanas ,' con qué frecuencia nos gustaría compartir los registros de los pacientes. Creamos este marco de simulación de Monte Carlo donde seleccionamos aleatoriamente la población, estimamos el riesgo, y lo hacemos miles de veces para estimar esta mirada - en los riesgos. Y aquí vemos algo llamado riesgo PK11, y queremos ser menos del uno por ciento, lo que significa que el porcentaje de registros que caen en el grupo demográfico de tamaño 10 o menor debe ser menor o igual al uno por ciento. En otras palabras, estamos estimando que menos del uno por ciento de la población estaría en un grupo de pacientes menos que el

tamaño - tamaño total 11 o menos que otros 10 en otros registros. Así que dada esta estimación de riesgo, y esto se basa en lo que los CDC están usando, así que tratamos de investigar ese riesgo básicamente utilizado por los CDC. Y miramos la distribución, y en base a estas distribuciones relacionamos los registros de privacidad y las políticas.

Diapositiva 9
Así que en los próximos experimentos que voy a mostrar, utilizamos esta lista PK 11, como he mencionado. Realizamos las simulaciones 1.000 veces y buscamos 96 políticas alternativas. Y hacemos esto en todos los condados y lo hacemos para cada condado por tamaño y número de casos.

Diapositiva 10
Así que lo que obtenemos es que para los condados pequeños cuando comienza la epidemia y tenemos pocos casos, los riesgos de privacidad son mucho más altos que el umbral aceptado que mencionamos. Así que no puedes compartir datos. Pero a medida que avanza el tiempo, incluso en los condados pequeños no podrás compartir mucho, pero en los más grandes, al menos desde este punto de vista de riesgo, podrías tener muchas políticas. Por ejemplo, este diagrama dice que si el recuento está entre 1.000 a 50.000 [personas] rango y llegamos al total de 5.000 casos, seríamos capaces de encontrar entre las 96 políticas que miramos que obtendríamos 31 para satisfacer el riesgo. Y estas políticas se enumeran, algunas de ellas aquí, como cómo grano fino la edad compartida, si tuvimos sexo, nacionalidad, raza, y así sucesivamente.

Diapositiva 11
Además de eso, buscamos un cambio político dinámico. En otras palabras, no nos limitamos a cambiar - a compartir un tipo de datos, pero evolucionamos lo que compartimos todo el tiempo y lo comparamos con las políticas estáticas de los CDC. En el caso del CDC, divide la edad en 0-9 [años], 10-19, etc. Este tipo de intervalos, como intervalos de 10 años. Ha combinado rango y etnia, género, estado de residencia y condado de residencia, y fecha de la primera colección de especímenes. Esa es la política estática del CDC en términos de sensibilización de datos. A este respecto, examinamos si nuestra política dinámica, que se adaptó sobre la base del riesgo, podría funcionar mejor. Especialmente para, nos gusta, hacemos lanzamientos diarios y semanales, básicamente.

Diapositiva 12
Así que no voy a entrar en todos los detalles, pero lo que sucede es que las políticas estáticas en la mayoría de los mismos casos, ya sea un condado pequeño o un condado grande, resultan tener más número de liberaciones, donde se supera el umbral de privacidad de riesgo. Así, por ejemplo, cuando miramos el 95 por ciento de cuantiles para un condado pequeño con un tamaño de población inferior a 1.000 estaríamos teniendo para el período que nos fijamos, tendríamos 22 días que el riesgo está por encima del umbral. Estos son lanzamientos diarios. Pero para una política dinámica teníamos incluso cero. Y por supuesto por un millón, de nuevo, ves el mismo umbral. Por lo tanto, este tipo de muestra que una política sobre los datos liberados y qué formato es puede no ser bueno y tenemos que ajustar realmente a medida que la pandemia evoluciona.

Diapositiva 13
Así que en este estudio lo que tratamos de mostrarle es que nuestro marco dinámico de evaluación de riesgos de privacidad puede dar resultados mucho mejores en términos de estimación de riesgos de privacidad. Y realmente puede adaptarse a los entornos cambiantes que protegen con mejores opciones de privacidad y utilidad. Pero, por supuesto, este trabajo que ahora continuamos solo mira el riesgo de privacidad. No nos fijamos en cuál es la diferente utilidad de estas políticas. En otras palabras, en algunos escenarios en los que la privacidad dada es un riesgo de privacidad aceptable, tenemos 40 diferentes políticas. Pero dadas las

nuevas tareas, qué política es mejor, por ejemplo, para la detección de brotes, o qué política es mejor para entender si el brote está ocurriendo en alguna carrera, por ejemplo. Así que no nos fijamos mucho en ellos.

Diapositiva 14
Así que me detendré aquí. Una vez más, me gustaría agradecer a NSF por apoyarnos. Y este es un trabajo conjunto con Vanderbilt Medical School y también un colega de IBM. Y esto es lo que presenté en muy poco tiempo. Si desea más detalles, se publica en el Journal of the American Medical Informatics Association recientemente. Así que voy a parar aquí y luego hacia el final, cualquier pregunta que voy a responder en vivo en línea gracias.