Nociones de Privacidad

- La privacidad es la capacidad de asegurase así mismo o información sobre uno mismo de forma que solo un sujeto puede decidir hasta que punto puede ser sabida.
- La privacidad ha evolucionado y se ha desarrollado a través de la percepción del individualismo.
- El derecho de cada uno a decidir cuanto, cuando y hasta que punto se puede dar información de uno mismo a los demás.

¿Quienes son los demás? (Adversarios)

- Sociedad: ostros individuos, compañías.
- Estados: Censos de países bajos en 1940 (Clasificación por religión)
- Cuando la sociedad era el estado: escándalo de Crypto AG: Empresa suiza que vendía dispositivos hardware de cifrado para conexiones que venían con una puerta trasera ya que parece ser que la cía estaba detrás de la empresa en cuestión.

La seguridad y la privacidad van de la mano ya que si no hay cierto nivel de seguridad es difícil mantener la privacidad, pero no son la misma cosa. La privacidad generalmente es algo subjetivo dependiendo de la persona o entidad.

Privacidad y Seguridad

La seguridad es una manera de alcanzar la privacidad. La seguridad coincide con la privacidad de que:

- Existen adversarios estratégicos: Alguien que tiene un plan para atacar con el objetivo de averiguar información.
- Muchos principios de seguridad que también se aplican a la privacidad.

También hay bastantes diferencias entre seguridad y privacidad:

- La privacidad transciende el dominio digital
- Modelo de amenazas: Normalmente los agentes débiles y a veces adversarios poderosos.
- No se puede asumir la existencia de terceras partes de confianza.
- Susceptibilidad a que el gobierno te obligue a usar determinadas herramientas o a solicitar cierta información mediante presión social.

Ataques de inferencia

Tratan de averiguar otras cosas a partir de datos disponibles. Normalmente cuando se observan unos datos, se puede revelar sobre alguien mucho más de lo que parece, el ser humano es muy predecible. A partir de unos datos se pueden inferir muchos otros datos. Un ataque de inferencia estadística permite, mediante un predictor estadística, que probabilidades hay de otros atributos diferentes.

Un ejemplo sería el de los likes de facebook, que servían de predictores de cosas privadas. A partir de

los likes se pueden saber cosas que no deberían saberse. Los adversarios observan los likes y a base de un estudio con 58k voluntarios, mediante una aplicación que permitía obtener:

- Información del perfil de facebook
- Lista de likes
- Test Psicométrico

Usando estos datos se creó una matriz en función de los datos obtenidos, mediante la técnica SVD (Singular Value Decomposition) convirtieron los likes en dimensiones del problema que se representaba de una forma más informativa y a partir de ahí se alimentaron los datos a un modelo de predicción que predecía la edad, género, preferencias políticas, etc... El resultado se representa como un AUC (Area Under the Curve) para las variables dicótomas. Se predijeron algunas variables mediante el coeficiente de correlación de Pearson. Como resultado se estableción que los mejores predictores de alta inteligencia eran:

- Thunderstoms
- The colbert report
- Science
- Curly Fries

Y el mejor predictor para poca inteligencia eran:

- Sephora
- I love being a mom
- Harley Davidson
- Lady Antebellum

Caso cambridge analítica

Ofrecía obtener perfiles psicológicos a partir de huellas digitales. Su matriz, ICL, destacaba por influir en elecciones e paises en vías de desarrollo desde los años 90. Los datos de 87 millones de usuarios de facebook fueron adquiridos a través de 270.000 usuarios. Cuando alguien le daba permiso a la aplicación para obtener sus datos, también daba permiso para obtener los datos de amigos. Esta compañía decía que daba 5k de puntos de datos sobre cada persona.

Desanonimizar a través desenlazado de datos

Utilizando bases de datos externas es posible desanimizar a personas con atributos externos. El adversario puede acceder por cualquier clase de medio a una base de datos no anonimizada y cruzando datos es capaz de desanonimizar a cualquiera. Cuanto más dispersos son los datos, más fácil es desanonimizar a alguien. Cuanto más único es un registro, más fácil es desanonimizar.

Desanonimización de Netflix

Netflix sacó un concurso que consistía en buscar herramientas de recomendación para los usuarios. Lo que hicieron fue ofrecer un premio a quien pudiera producir el mejor recomendador. Cada uno de los puntos de datos eran un ID de usuario (anonimizado), un ID de la película, la puntuación que daba

la persona y la fecha en la que se había dado. Se pensaba que esta información era completamente privada y que esta información era segura.

La base de datos era muy dispersa por los miles de atributos que habían. El conjunto de películas que una persona había visto era casi una variable única. Habían una serie de películas que solo las habían visto un pequeño puñado de personas que resultaron ser atributos casi únicos. Lo que se hizo fue medir la similitud entre los registros de datos, con cierta tolerancia entre la puntuación y la fecha. Se empezaron a utilizar 1 y 0 para indicar si una película gustaba o no. Se tomó la base de datos de netflix y se cruzó con a IMDb y se realizó una búsqueda de gente que dio puntuaciones parecidas en las mismas fechas consistentemente.

De 50 personas en la base de datos se pudieron identificar 2 personas cruzando los datos con IMDb. Las consecuencias de esto fueron que los que estaban en la base de datos denunciaron a Netflix al considerarse esto una violación de la privacidad y netflix tuvo que pagar 9 millones de dólares a estos usuarios.

Desanonimización bssada en la localización

Se tomaron datos de trazas GPS de coches en el área de Detroit con un minuto de resolución. Cuando los coches estaban apagados no enviaban datos. Se eliminaron sitios donde no habían visitas por la tarde y se eliminaron coches fuera de áreas residenciales. Con estos datos se vio que era posible localizar la casa de alguien con estos datos. Durante 2 semanas, con los datos de 172 personas con una resolución de 6 segundos, medianto el uso de heurística para separar los vehículos como:

- A las 3 am estaban en su casa
- El lugar en el que los individuos pasaban más tiempo era su casa
- Se uso geolocalización reversa y páginas amarillas para desanonimizar a los usuarios.

Se logró desanonimizar a un 5% de los sujetos. Incluso con un ruido de std=500m se obtuvo alrededor de un 5% de éxito para localizar las direcciones correctas.