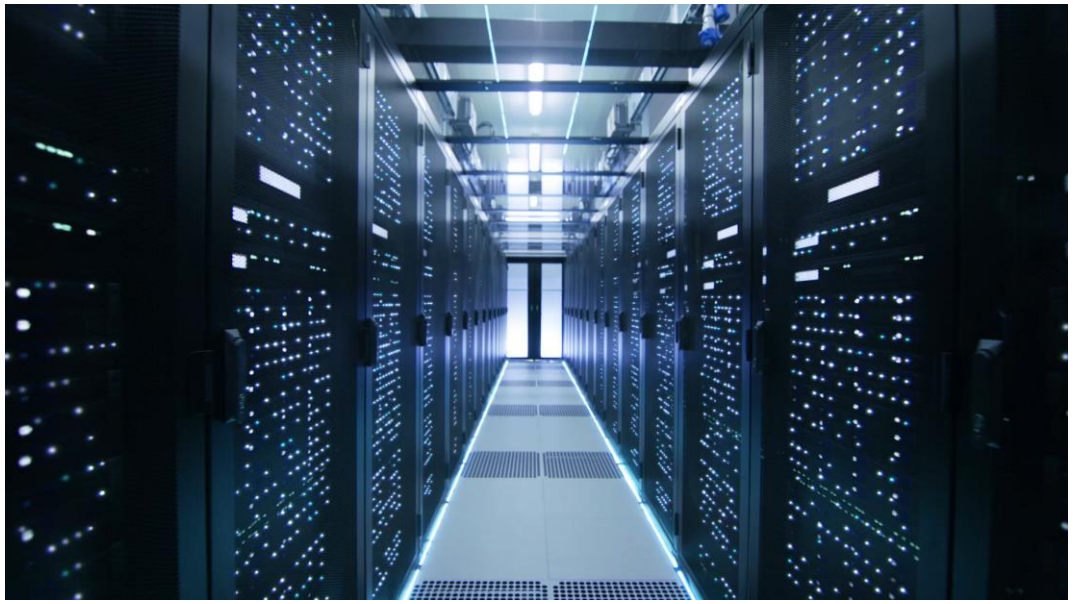

Big Data Privacy

BDM 1043

Submitted by: Aadarsha Chapagain

Big data Privacy

- Big data privacy is all about protecting the privacy of the sensitive information and personally identifiable information PII. (Informatica, n.d.)
- Information privacy is the freedom to have choice of either sharing or not sharing the personal information with other. One of the major privacy issue is identification of personal information which transmitting those information over the internet. (Priyank Jain, 2016)
- More spread of data among the sources means high risk of data breaches.
- The collection of data is most likely to happen more in the future whether we like it or not. (O Reilly, 2012)

Privacy Vs. Productivity

- The Scope of information available to government and business sources has increased with increase in data mining , analytics and computing power and storage. (The Economist, 2010)
- The number of the devices which are connected across the network has ability to communicate, share and access data which contribute to the volume and velocity in Big data (Tene, 2011)
- Everything around us is data now and it directly impacts our economy, productivity and efficiency

But it seems to have some **trade off** between the benefits of the data protection of one’s privacy.

Traditional approach

- De- identification using approaches such as anonymization, encryption and data sharding are some of approaches used to maintain data privacy in the past but it seems those data can also be re-identified and associated with the individuals. So, maintaining Data privacy is really a challenge.
- Clearly applying only, the traditional approach for maintaining the privacy in big data is not enough. So in 90’s concept called “Privacy By Design was developed” to address the privacy issues in ever growing data . The major focus of the concept is that the privacy should be considered as default feature while building the systems not later. The architecture of the system should be designed in the way that treats privacy protection as one of major building block.

7 Foundations Principles(Cavoukian)

1. Proactive not Reactive; Preventative not Remedial
   - preventing privacy breaches before they actually happen.
   - security features built up at the design phase of system
2. Privacy as the Default Setting
   - make sure that privacy is maintained automatically
   - individual are allowed to secure information if they want, but by default the system has to be secure
3. Privacy Embedded into Design
   - privacy must be embedded in a holistic way
   - poor user experience in tradeoff with privacy is not allowed

4. Full Functionality — Positive-Sum, not Zero-Sum
   - No comprise between privacy and functionality
   - A system should be both functional and fully secure.
5. End-to-End Security — Full Lifecycle Protection
   - security is maintained all the way long from start to end.
   - information is secure while is enters and exit the system and even when its destroyed
6. Visibility and Transparency
   - Openness about the use of the information is mandatory
   - transparency in use of information builds up trus

7. Respect for User Privacy
    - Optimization of system for users and all their needs.
    - User privacy is the top priority

Big Privacy (Big Privacy: Protecting Confidentiality in Big Data)

- There has been different approaches for anonymizing the personal data when releasing publicly, but scientist and researcher were able to relate those data to individuals.
- Lately, Netflix released anonymized data of 480,000 users describing their “movie viewing habits”, but Scientist Arvind Narayan and Vitaly Shmatikov linked many customers to an on-line move rating website and were able to identify them.
- It shows that de-identified data are not just enough to maintain privacy.

Methods for Protecting Public Release Data

Aggregation

It involves generalizing the data to a wide scope so that the specific information cannot be tracked.

Example

There may be a individual having specific demographic characters in a city but there might be many of such individuals in a state. So aggregating state level data makes it more difficult to relate the information to individual.

Suppression

The sensitive information from the data can be omitted before release so that analysis biased on those sensitive information becomes difficult.

Data Swapping

Data for selected columns or features could be swapped.

Example

Swap the features such as gender, ethnicity, age for records at risk with other records to demotivate users from matching, since they will be matching on wrong data.

Adding random noise

Data can be obscured by adding randomly selected value. Adding random values to data can discourage the matching and manipulate the values of sensitive variables.

Synthetic data

The main concept of synthetic data is replacing the original data with the data generated using same probability distributions. There are two kinds of synthetic data, partially Synthetic data and Fully synthetic data.

References

(n.d.). Retrieved from Big Privacy: Protecting Confidentiality in Big Data: https://users.cs.duke.edu/~ashwin/pubs/BigPrivacyACMXRDS_final.pdf

Cavoukian, A. (n.d.). *Privacy by design.* Retrieved from The 7 foundation Principle: https://www.ipc.on.ca/wp-content/uploads/resources/7foundationalprinciples.pdf

Informatica. (n.d.). *Big Data and Privacy: What It Is and What You Need to Know*. Retrieved from Informatica: https://www.informatica.com/hk/resources/articles/what-is-big-data-privacy.html

O Reilly. (2012). The Application of big data. In *Big Data Now* (p. 54). O Reilly Media, Inc.

Priyank Jain, M. G. (2016). *Big data privacy: a technological perspective and review.* Retrieved from Springer Open: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0059-y#ref-CR13

Tene, O. (2011, February 01). *Privacy: The new generations*. Retrieved from Oxford Academic: https://doi.org/10.1093/idpl/ipq003

The Economist. (2010, Feb 25). *Data, data everywhere*. Retrieved from The Economist: https://www.economist.com/node/15557443.