
**E-ISSN:** 2706-8927
**P-ISSN:** 2706-8919
www.allstudyjournal.com
IJAAS 2020; 2(2): 294-296
Received: 26-01-2020
Accepted: 28-02-2020

**Abhishek Jaiswal**
Assistant Professor, Shyama Prasad Mukherji College for Women University of Delhi, Delhi, India

**Corresponding Author:**
**Abhishek Jaiswal**
Assistant Professor, Shyama Prasad Mukherji College for Women University of Delhi, Delhi, India

# The game of data and privacy

## Abhishek Jaiswal

**Abstract**
Privacy is a fundamental right and its breach should be punishable offence. But its definition/scope is not uniform globally which gives an opportunity to privacy invaders. Some countries have complex regulation structure for privacy related issues like in U.S. where structure is departmentalized. This makes the system opaque for the citizens and discourages them to exercise their rights. Like any other technology, data analytics domain is swiftly evolving, thus it becomes difficult to consider all the privacy breaching scenarios under the preview of law enforcement and regulation. Even Netflix could not anticipate de- anonymization algorithm and compromised its customers data and paid USD 9 million for lawsuit under VPPA Act 1988. Thus, I would suggest input (training dataset) and output should be encrypted, and algorithm should work on encrypted data for the sequel of such competitions. Ethics is a broader concept than privacy as it deals with the activities which are wrong or right and not necessarily legal or illegal. To be ethically correct, corporates, government, individuals etc. involved in data analytics should follow guidelines, regulations stipulated by various organizations like OECD, concerned/respective regulatory bodies etc. Hence it is unethical to use unregulated software/algorithms such as facial recognition which has low accuracy and higher probability of misuse.

**Keywords:** Data and Privacy

## Privacy Issues
Quality of the solution directly depends on the quality of the problem description. Thus, Einstein once said that he would spend fifty-five minutes to define the problem and only five minutes to find the solution if he had one hour to save the world. In definite science there are globally accepted laws such as law of gravity, law of motion etc. because the problem or phenomenon these laws explain/quantify is universal in nature, thus solutions are also accepted universally. This is unlike with many topics in the domain of social science, psychology etc. Problem of privacy has similar issue.

Privacy as a fundamental right is globally accepted, but its scope/definition is not uniform universally. Thus, it becomes difficult to regulate privacy issues which results in escape for privacy invaders. Parker (1974) [1] asserted that privacy is control over when and by whom the various parts of us can be sensed (heard, touched, smelled, tasted, seen) by others.

On the other hand, Craig and Ludloff (2011) [2] suggested three categories of privacy in the digital age; physical, informational, organizational. Physical privacy is related to protection against infringement possession or space. Informational privacy is regarding financial, medical, online information etc. Organizational privacy involves that government agencies, corporates, business need to keep available information secured. In digital age, data has no border i.e. data can be stored in a cloud or physically in other location, for example Microsoft can provide data stored on its European servers to American investigators without informing the individual in adherence with the U.S. Patriot Act. This would be violation of EU's Data Protection Directive and Safe Harbor agreement with the U.S. Data analytics domain has evolved and constantly evolving which comprises of, 'big data', 'data integration', 'data mining', and 'data matching'. This also creates a constraint on regulators to control privacy issues due to advancement in technology. Data analytics is governed by eleven Australian Privacy Principles mentioned in Privacy Act 1988. The Office of the Australian Information Commissioner (OAIC) administers the protection of privacy of individuals under the privacy act. Spam Act and the DCNRA is enforced by the Australian Communications and Media Authority (ACMA). Similarly, there are other departments responsible for the privacy in their respective department. (Oaic.gov.au, 2019) [3]. The U.S. privacy regulatory model is also departmentalized whereas European Union regulator model has top down approach which has incorporated OECD's eight principles for the protection of

personal data. Irrespective of fragmented regulatory model and scope for privacy across countries, it seems more importance has been given to safety and security against privacy. For example, in the U.S. the Patriot Act, passed into after 9/11 incident allows agencies to search call details, emails, medical, financial and other details. (Craig and Ludloff, 2011) [2]. However, regulations across globe is harsh on breach of personal information privacy when safety and security concern is not involved, for example Netflix data challenge.

**Netflix training data**

In quest to improve customer experience by providing better movie recommendations, Netflix publicly announced a challenge to improve quality (reduce RMSE by 10%) over its existing algorithm 'Cinematch' in 2006. Data of 100 million ratings given by 480,189 users for 17,770 movies rated between October 1998 and December 2005 was released by Netflix as a training dataset. The ratings were on the scale of 1 to 5 (integral) stars. To maintain anonymity, each customer ID was replaced with a randomly-assigned ID. The date of each rating and the title and year of release (not necessarily theatrical, can correspond to the release of DVD) for each movie id were also provided. Netflix also provided probe, quiz and test dataset of 1.4 million ratings each. Quiz dataset was used as validation dataset and test dataset was used for final evaluation of the algorithm. (Kaggle.com, 2019) [4]. Dataset provided was high dimensional sparse data which was randomly selected and sanitized. Usually data having staggeringly high number of explanatory variables, sometimes even larger than the sample size (data points) in the data set is called as high dimension data. On the other hand, when high percentage of attributes/explanatory variables is null or empty i.e. not dense, then it is called as sparse data. Training dataset provided by Netflix was a matrix of 480,189 users (rows) and 17,700 movies (columns). Thus, a 100% density would be 8499.34 million ratings (480,189 * 17,700), however dataset had only 100 million ratings i.e. only 1.18% data points was available. With high number of attributes and extremely low density, this dataset was high dimensional sparse data. Another challenge in this dataset was skewness which created bias because density of the ratings given by the users varied drastically across all movies. In 2009 a team "BellKor's Pragmatic Chaos" won the challenge and reward of USD 1 million. They performed collaborative filtering by using baseline predictor, baseline errors and levered similarities in neighborhood predictor algorithm which reduced RMSE by 10.06%. (Coursera, 2019) [5].

Netflix did provide anonymous data to avoid privacy breach but was unaware about the techniques which can de-anonymous data using the rating from other sources as auxiliary information (Aux). Arvind Narayanan and Vitaly Shmatikov from University of Texas were able to identify some of the Netflix users in the dataset after 16 days of its release. This triggered privacy breach and violation of VPPA Act 1988. Basic functionality of the algorithm was to identify similar Aux in adversary/other database (IMDb in this case) which has more personal information about the user. First step in the algorithm was 'Scoring function' which assigns a score to each record based on how well match is found in Aux. If the highest score is much higher than the second highest score, then return the highest score, else no match. Then the match is checked for its accuracy by using formula $\{(max-max2)/\sigma)\}$ $\geq \phi$, where max is the best match, max2 is the second-best match and $\phi$ is eccentricity threshold (like cosine similarity). Last step is self-testing, in this algorithm removes the best match and re-runs to check there is no match with high probability, i.e. nearest neighbor is too far. Arvind Narayanan asserted that this algorithm is robust on databases published for collaborative filtering even if it is sanitized and perturbated; and if error introduced in the dataset is large then the data utility will be lost. (Narayanan and Shmatikov, 2019)

Netflix dropped the sequel of the competition and settled USD 9 million privacy lawsuit in violation of VPPA Act 1988 for the contest held in 2006. I would suggest CTO that, similar contest can be carried only if the data is encrypted and the algorithm runs on the encrypted data. However, it opens room for other discussion such as, safer network environment for the participants, access control to the data etc.

**Ethical issues and analytics**

If privacy is related to confidentiality element, then ethics is related to integrity element of the information security. Ethics is a broader concept than privacy as it deals with the activities which are ethical or unethical and not necessarily legal or illegal. Thus, privacy breach is also unethical. As per OAIC, an ethical framework helps to categorize ethical issues, standardize guiding questions while using/managing data. (Oaic.gov.au, 2019) [3]. Example, Data Governance Australia has developed 8 Leading Practice Data Principles namely, No-harm rule, Honesty & transparency, Fairness, Choice, Accuracy and access, Accountability, Stewardship, Security (DGA, 2019) [7]. Corporates should appropriately use analytics considering a wide range of stakeholders (consumers, government, businesses, policymakers etc.). Companies indulging in privacy flaws loses consumers trust (Afroz et al, pp. 10-17) [8]. Companies should develop internal policies and have accountability matrix. When dealing with sensitive information, companies should adhere to the regulations, like OECD guidelines suggests having information which is complete, accurate and up-to-date. Companies should anonymize personal information, constantly review and revise analytics and use relevant personal information with their consent. (Schwartz, n.d.).

One of the pervasive concerns raise by academicians, media, Human rights etc is regarding the use of facial recognition models/software. It is widely used by intelligence and police department. This triggered the debate on privacy verses security. Manufacturers of the system and police department suggested that this technology offsets privacy loss by great security benefits. However, opponents claimed that the concern is not only about privacy, but about the accuracy of the algorithm as well, because an incorrect match can result into imprisonment of an innocent. Brey (2004) [9] concluded in 2004 paper that use of facial recognition software should be prohibited due to absence of legislation regarding where, why and who can use such software. Also, accuracy needs to be improved as software yielded many false positives. One of the reasons for false positives might be the bias due to sample database used to train the algorithm. In 2011 study, National Institute of Standards and Technologies (NIST) claimed that algorithms developed in China, Japan and South Korea recognised East

Asian faces more accurately than Caucasians and result was vice versa for algorithms developed in France, Germany, and the U.S. As an impact of such racial bias, African American were twice likely to be arrested than any other race in U.S. (Clare Garvie, 2016) [10]**.** In 2018 study MIT Media Lab suggested that algorithms designed by IBM, Microsoft, and Face++ had error rates up to 35% and performed differently based on the age, gender, and ethnicity of the person. Apart from biasness, there is another concern of misuse of the software by the government officials. Government officials already have access to a wide variety of personal information of the adult population and such software can be misused to track a person's all activities. This will result into complete loss of privacy (Vincent, 2018) [11]. In another incident in Jul 2018, American Civil Liberties Union (ACLU) performed an experiment using Amazon's facial recognition software and found that 28 Congress members were misinterpreted as different people who had criminal record. Amazon responded that ACLU should have used best practice of setting confidence thresholds to 95% considering its use for law enforcement. ACLU replied on this comment that 80% confidence threshold is suggested by Amazon for recognising human faces which was used by ACLU and Amazon should take more responsibility for such software; meanwhile this technology for face surveillance won't be used on law enforcement. (Leswing, 2018) [12].

Considering this, I would suggest CTO that implementation of such software might be useful for identifying one to one match, but for many to one match we should have comprehensive sample size to train the algorithm and can be used as a supporting tool until its accuracy reaches at optimum level.

**References**


1. Parker Richard B. A Definition of Privacy, Rutgers Law Review 1974;27(2):275.
2. Craig T, Ludloff M. Privacy and big data. Beijing: O'Reilly 2011.
3. Oaic.gov.au. Guide to Data Analytics and the Australian Privacy Principles| Office of the Australian Information Commissioner - OAIC. [online] Available at:https://www.oaic.gov.au/agencies-and organisations/guides/guide-to-data-analytics-and-the-australian-privacy-principles 2019.
4. Kaggle.com. Netflix Prize data. [online] Available at: https://www.kaggle.com/netflix-inc/netflix-prize-data 2019.
5. Coursera. Netflix Prize: The Competition - Movie Recommendation on Netflix | Coursera. [online] Available at: https://www.coursera.org/lecture/networks-illustrated/netflix-prize-the-competition-Mx4ze 2019.
6. Narayanan V, Shmatikov. Robust De-anonymization of Large Sparse Datasets, or How to Break Anonymity of the Netflix Prize Dataset. S&P (Oakland) 2008.
7. DGA. Principles - DGA. [online] Available at: http://datagovernanceaus.com.au/leading-practice-data-principles 2019.
8. Afroz S, Islam AC, Santell J, Chapin A, Greenstadt R. How Privacy Flaws Affect Consumer Perception, Third Workshop on Socio-Technical Aspects in Security and Trust, New Orleans, LA 2013, 10-17.
9. Brey P. Ethical Aspects of Face Recognition Systems in Public Places, Journal of Information, Communication & Ethics in Society 2004;2(2)97-109
10. Clare Garvie J. Facial-Recognition Software Might Have a Racial Bias Problem. [online] The Atlantic. Available at: https://www.theatlantic.com/technology/archive/2016/04/the-underlying-bias-of-facial- recognition-systems/476991/ 2016.
11. Vincent J. The tech industry doesn't have a plan for dealing with bias in facial recognition. [online] The Verge. Available at: https://www.theverge.com/2018/7/26/17616290/facial-recognition-ai-bias-benchmark-test [Accessed 27 Mar. 2019].
12. Leswing K. Read Amazon's full response to the ACLU report about its facial recognition software misidentifying members of Congress as previously arrested. [online] Business Insider Australia. Available at: https://www.businessinsider.com.au/amazon-response-to-aclu-facial-recognition-study-congress-member-photos-2018-7 [Accessed 27 Mar. 2019].