Universitetet i Oslo – IN2120 Informasjonssikkerhet– 2022 – midtsemestereksamen

TITTEL

26

Tracking and mass surveillance in the era of Big Data

Kandidatnummer:

julihau

olejfri

Antall ord: ORD

Contents

1 Abstract

How have legislations, attitudes towards and the consequences of tracking and mass surveillance changed over the course of time, taking into consideration the recent global pandemic and how data-tracking was used as a countermeasure.

The practice of mass surveillance was first put to use as a result of the happenings of 9/11 2001 (Sinha, 2013) with the reasoning that it could be used to help prevent future terrorist attacks and the like. It was presented as a necessary evil in order to assist authorities in ensuring the safety of the people, and as such, a new normal emerged. By using massive sets of data collected from its netizens, new systems were created to predict anything from traffic jams to premediated crime (Zuboff, 2015).

The people's willingness to volunteer their data likely stems from the theory that in order to feel a sense of safety from these frightening prospects, a person would rather sacrifice their own privacy than risk becoming a victim themselves (Zuboff, 2019). This would however turn out to be the beginning of something much bigger, namely the understanding of how much you could actually discern about a person by analyzing their data, and how big companies could profit from doing exactly that (Vargo & Hopp, 2020).

There have been a few notable cases of tracking, manipulation and surveillance is years following the start of this new paradigm. One example being the Cambridge Analytica scandal, in which it was brought to light that Facebook had handed over personally identifiable information of more than 87 million users to to Cambridge Analytica. It would later be discovered that this information was used to influence the American presidential election of 2016 (Rathi, 2019). What was made clear following scandals such as this was that this new era carried with it the risk of private companies being able to control the masses (Karas, 2002), even to the point of national leaders being elected by their will.

During the Covid-19 pandemic, multiple forms of mobile applications with the express purpose of tracking the spread of the virus were introduced. Both government-approved applications and ones developed by notorious private companies such as Apple and Google (Kaashoek & Santillana, 2020) were put to use on an international level (Munzert, et al., 2021). This begs the question; "Where does user privacy begin and end?" How can we be sure that the data we are giving away, with the intent of benefitting the public good, is used in a manner that is effective or at all useful?" Could this be the start of an era where mass data-collection and surveillance are an everyday phenomenon? What data will be standard procedure to collect from users of mobile

applications, and how could this information be used in the future? It would seem that it all boils down to the question; “In what ways are we allowing both authorities and private companies to influence our lives by sharing our data?”.

2 Hovedkapittel

Due to the increased popularity of personal profiles on social media sites and the like, we find ourselves in a period of mass personalization (Mulvenna, et al., 2000). Many websites and applications like Facebook, Twitter and Tumblr can potentially create a comprehensive list of personal data such as demographic, behavioral, psychographic and lifestyle-data simply by using its intended functions (Russel, 2013) (Arya, et al., 2019). In addition to this mined data, many applications such as WhatsApp, Messenger and TikTok include a statement in their policies, informing us that by using their applications, the user is consenting to their communication being put through textual analysis and searched for keywords. These keywords are then used to personalize what advertisements are shown to the user and other users who display similar textual patterns (Vargo & Hopp, 2020).

According to (Trang, 2017), advertising strategies have been changed drastically as a result of the development of social media. Due to their 1.39 million active monthly users, Facebook has become one of the most common online marketing tools with 92% of social marketing companies having chosen to use the platform due to their comparatively low prices and massive user base.

In addition to this traditional form of advertisement, Facebook teamed up with IBM on May 6th, 2015 with the purpose of creating better advertising on what was then, and continues to be, the world’s largest social network. By combining Facebook’s targeting technology and IBM’s services for marketers, they created a new form of advertisement system where users’ interactions on collaborating marketers’ websites resulted in advertisements of viewed products or services being generated and put on their Facebook feed. An example of this would be if a user visited company X’s website and looked at a certain product. This viewing data would then be sent by the company to Facebook, and an ad of that product featuring an affiliate link to company X’s website would then later appear on the user’s Facebook feed (Trang, 2017).

While this practice has been shown to have very high success rates compared to the more traditional methods of advertising, it has brought with it the consequence that selling these massive amounts of data to third parties has become a big business in the 21st century (Palos-Sanchez, et al., 2019).

2.1 Manipulation and Recruitment

We have already discussed why using or selling user data for advertising purposes may be one of the main reasons for websites and applications to collect it, but taking into account the questions we posed earlier, there is another aspect to consider. Namely user data being used by potentially malicious parties wishing to manipulate or recruit the user.

As previously mentioned, one notorious example of users can be manipulated as a result of their data being shared is the Cambridge Analytica scandal.

Cambridge Analytica was officially an advertising company with both a commercial and a political wing. The political wing was claimed to "combine the predictive data analytics, behavioral sciences, and innovative ad tech into one award winning approach". The method in practice involved making unique advertisements for many different categories of users which were grouped based on their analyzed traits. These analyzed traits were the results of a regression model that in the words of Professor David Sumpter: "Takes the data we already know about a person, and uses them to predict something we don't know about him or her" (Sumpter, 2018). The idea behind it all was that different types of people will react more strongly to different forms of advertisement, be it appealing to the viewers sentimentality or by inciting feelings of anger with the opposition. By exposing Facebook users to these tailored ads, they intended to yield better results compared to the more traditional blanket advertisement where the same ad is shown to everyone.

The effectiveness of this approach has been both exaggerated and underplayed by both the media, CEO Alexander Nix, and the whistleblower and ex-employee Christopher Wylie, but the fact remains that Cambridge Analytica, to an unknown extent, helped republican politician Ted Cruz get the second highest number of votes in the republican primary election of 2016 (Rathi, 2019).

Despite the arguments of both sides regarding how effective or ineffective these kinds of regression models are, users of platforms such as Facebook are in fact having their data used in what has been called both "an ethical grey area" and "an attempt to manipulate voters by latching onto their vulnerabilities" (McCausland & Schnecter, 2018).

In addition to this form of manipulation, we can also refer to the research done by (Alava, et al., 2017) which speaks of a relatively recent, but growing, phenomenon often referred to as "incitement to radicalization towards violent extremism" or more simply put "violent radicalization". The situation we are faced with according to their research is that this boom of social media has created new and easier ways for radical groups to identify and approach users that they believe could become allies to their cause. Taking into account that these groups, given sufficient funds could purchase potentially massive amounts of user information to aid them in their recruitment, puts to question whether the premise that the gathering and availability of personal data is in the best interest of public safety.

Taking what we have discussed about the subject into account, we can say that by using applications and services like Facebook, Twitter and Snapchat, the user is either knowingly or unknowingly offering their data to the company and whomever they choose to share that data with. This data may be used against the user in the form of targeted advertising, manipulation or potentially tracking and surveillance depending on what data the application chooses to collect. One could argue that a common denominator that allowed, and still allows, these happenings to take place is the culture of mass data collection which, as previously mentioned, has created a thriving market for the commerce and utilization of user data.

2.2 The COVID-19 pandemic

The question of whether or not the handling and management of user data is a violation of user privacy has been the subject of many studies in this era of Big Data (Buchanan, et al., 2006). Something that brought more attention to the issue was however the introduction of the COVID-19 related tracking of citizens.

Since the beginning of the virus-outbreak in 2020, the pandemic caused many systematic changes in organizational structures and society as a whole. Surveillance and control in the form of collecting data about location and interactions for its citizens was advised worldwide, which caused fear for the thought of a future where pandemic countermeasures could systematically invade user privacy (Ribeiro-Navarrete, et al., 2021).

After the recognition of the second wave of the COVID-19-pandemic, increasingly more countries chose to implement large-scale technology-based tracking measures in order to monitor and prevent the spread of the virus across the population. Despite the governments' efforts, these technology-based tracking applications ultimately experienced a global failure to convince the public due to these applications' perceived risk of infringing individual liberties and were met with very low levels of trust from the population (Georgieva, et al., 2021).

The findings in the article by (Hargittau, et al., 2020) describe the public distrust towards contact-tracing apps as being the product of a decade of media-covered security breaches and other violations of users' privacy expectations with one notable example being ”Edward Snowden's revelations of mass surveillance”, which is thought to have single-handedly caused a significant decrease of trust in the federal government. When trust in an institution decreases, people are more likely to turn to a competitor, but with research showing a steady decline of trust in technology companies in light of happenings such as the Cambridge Analytica scandal (Selinger & Hartzog, 2016), these non-governmental contract-tracking apps are distrusted as well.

This situation of distrust towards both governmental and privately owned contact-tracing apps has lead to a dilemma in which the government can chose to use other means in order to track the contagion of the virus, but at the cost of using their power to commandeer location and interaction information from cellular service providers.

(skriv om denne her https://www.tandfonline.com/doi/full/10.1080/0960085X.2020.1802358)

2.3 How powerful is really more data for large corporations?

The Central Limit Theorem (CLT) is "the most important theorem of probability" (Devore, 2018, p. 298). It origins from the observation that "averaging produces a distribution more bell-shaped than the one being sampled" (Devore, 2018, p. 298). In fact, the mean μ will be approximately normal for large n. As a rule of thumb, "if n ¿ 30, the Central Limit Theorem can be used" (Devore, 2018, p. 302). So if we wish to calculate a probability for the mean in order to create confidence intervals or carry out hypothesis testing, "we need only 'pretend' that the sample mean is normal, standardize it, and use the normal table" (Devore, 2018, p. 299). And this by only requiring a manageable sample size.

Another branch under data science is machine learning, a science that is about building models algorithmically based on the sample. In statistics, known models are tested, and their performance in relation to the sample is studied. This is not necessary in machine learning where the power of computers are leveraged. It is interesting to see how impactful this approach has become. A quick news search on AI will paint a story of AI supremacy. "AI 'outperforms' doctors diagnosing breast cancer"; "AI system outperforms humans in designing floorplans for microchips"; "AI Outperforms Humans in Question Answering"... It has been able to vastly expand on data science from what statistics has offered. Different from statistics, machine learning arguably thrives more on large sample size. In a publication titled "Do We Need More Training Data?" the authors write (Zhu et. al, 2016) "Based on our analysis, we conjecture that the greatest gains in detection performance will continue to derive from improved representations and learning algorithms that can make efficient use of large datasets," (p. 76) and highlight importance on correctly tuned models over sample size. The research was done on object recognition systems. For earlier conducted research in the field of Natural Language Processing, different evidence was found. "At least for the problem of confusable disambiguation, none of the learners tested is close to asymptoting in performance at the training corpus size commonly employed by the field" (Banko, Michele and Brill, 2001).

The term *confusion set disambiguation* is explained to be ”the problem of choosing the correct use of a word, given a set of words with which it is commonly confused.” The total corpus size in this experiment reached one billion, which is a large data-set indeed and might explain their findings.

2.4 Is surveillance and tracking moral?

Johnson, Robert and Adam Cureton, ”Kant’s Moral Philosophy”, The Stanford Encyclopedia of Philosophy (Fall 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), URL = https://plato.stanford.edu/archives/fall2022/entries/kant-moral/˙

2.5 Governmental data collection, tracking and surveillance

Paraphrasing the words of (Cate, 2008), The United States government has long sought data about individuals for a wide variety of important public purposes, but due to the then time and cost-heavy processes yielding mostly useless data, this collecting of data was protected under The Supreme Courts concept of “practical obscurity”.

The concept stems from the case “U.S. Department of Justice v. Reporters Committee for Freedom of Press” of 1989, and in practice means that while some information is technically accessible to the public, the time and effort needed to acquire this information makes it somewhat protected from mass access. The example of this used in the Supreme court case was that whilst the rap sheets and court records of a person were technically available to any person with the time to dig through a county courthouse’s paper records in search of them, the high cost and low chance of acquiring them meant that the information had little chance of being compiled by the public (Reporters Committee for Freedom of the Press, 2009).

While this concept was deemed adequate protection of privacy back in 1989, with new technology, the situation has changed. Despite court records and arrest reports being available only as physical copies in most U.S. counties, corporate data brokers are free to collect this information on individuals from all over the country which they can then sell to government agencies, insurance companies and other third parties belonging to the private sector. A notable example of such a company is ChoicePoint which offers aggregate reports of a person’s line of credit collected from multiple public and private databases (Personal Finance, 2019). These reports can in some cases determine whether a person is considered qualified for a job, professional license, or insurance (Testa-Avila, 2008).

While the government does not acquire this data directly, the Supreme Court has refused to restrict the government's access to data held by third parties which they could do by extending the Fourth Amendment that was originally meant to ensure the right to privacy (Cate, 2008). As it currently stands, companies have common authority over their business records and can therefore consent to a government search or acquisition of their databases even if their users oppose it (Kerr, 2021).

This enforces that by using applications, websites or services, the user allows both private sector parties and authorities access to their personal data which could be used for both surveillance and tracking.

2.6 DELKAPITTEL

While this practice of businesses benefitting from or purchasing user data may be profitable, it can in some cases be a violation of user privacy. The handling and management of both public and private information has therefore been the subject of many studies in this era of Big Data (Buchanan, et al., 2006). . . . (skriv mer om dette)

3 Conclusion

4 References

Devore, J. L., & Berk, K. L. (2018) Modern Mathematical Statistics with Application. New York Dordrecht Heidelberg London: Springer.

Zhu, X., Vondrick, C., Fowlkes, C.C. et al. Do We Need More Training Data?. Int J Comput Vis 119, 76–92 (2016). https://doi.org/10.1007/s11263-015-0812-2.

Banko, Michele and Eric Brill. “Scaling to Very Very Large Corpora for Natural Language Disambiguation.” ACL (2001).

Alava, S., Frau-Meigs, D. & Hassan, G., 2017. Youth and Violent Extremism on Social Media: Mapping the Research. s.l.:United NAtions Educational, Scientific and Cultural Organization.

Arya, V., Sethi, D. & Paul, J., 2019. Does digital footprint act as a digital asset? – Enhancing brand experience through remarketing. s.l.:International Journal of Information Management.

Buchanan, T., Paine, C., Joinson, A. N. & Reips, U., 2006. Development of measures of online privacy concern and protection for use on the Internet. s.l.:Wiley Online Library.

Cate, F. H., 2008. Government data mining: The Need for a Legal Framework. s.l.:Harvard Civil Rights-Civil Liberties Law Review.

Etternavn1, F. & Etternavn2, F., 2022. Et publisert tidskrift. Tidskrift 1, Oktober, pp. 1-12.

Jøsang, A., 2021. Informasjonssikkerhet - teori og praksis. s.l.:Universitetsforlaget.

Karas, S., 2002. Enhancing the privacy discourse: consumer information gathering as surveillance.. s.l.: J. Tech. L. & Pol'y..

Kerr, O., 2021. Buying Data and the Fourth Amendment. s.l.:Lawfare Institute.

Kaashoek, J. & Santillana, M., 2020. COVID-19 positive cases, evidence on the time evolution of the epidemic or an indicator of local testing capabilities? A case study in the United States.. s.l.:SSRN.

McCausland, P. & Schnecter, A., 2018. Trump-Linked Consultants Harvested Data from Millions on Facebook.. s.l.:NBCUniversal News Group.

Mulvenna, M. D., Anand, S. S. & Büchner, A. G., 2000. Personalization on the Net using Web mining. s.l.:Communications of the ACM.

Munzert, S. et al., 2021. Tracking and promoting the usage of a COVID-19 contact tracing app. s.l.:Nature Human Behaviour.

Palos-Sanchez, P., Saura, J. R. & Martin-Velicia, F., 2019. A study of the effects of programmatic advertising on users' concerns about privacy overtime. s.l.:Elsevier.

Personal Finance, 2019. What is a ChoicPoint Rrport?. [Internett] Available at: https://personal-finance.extension.org/what-is-a-choicepoint-report/ [Funnen 25 October 2022].

Rathi, R., 2019. Effect of Cambridge Analytica's Facebook ads on the 2016 US Presidential Election. s.l.:Towards Data Science.

Reporters Committee for Freedom of the Press, 2009. Out of sight, out of bounds: Article from the Spring 2009 issue of The News Media & The Law. [Internett] Available at: https://www.rcfp.org/journals/the-news-media-and-the-law-spring-2009/out-sight-out-bounds/ [Funnen 25 October 2022].

Russel, M. A., 2013. Mining the Social Web. 2. red. s.l.:O'Reilly Media.

Sinha, G. A., 2013. NSA surveillance since 9/11 and the human right to privacy. s.l.:s.n.

Sumpter, D. J. T., 2018. David J. T. Outnumbered: from Facebook and Google to Fake News and Filter-Bubbles — the Algorithms That Control Our Lives.. s.l.:Bloomsbury Sigma.

Testa-Avila, E., 2008. Practical Obscurity in the Digital Age: Public Records in the Private Sector. s.l.:UC Berkeley.

Trang, P. T., 2017. Personalized ads on Facebook: An effective marketing tool for online marketers. s.l.:Elsevier.

Vargo, C. J. & Hopp, T., 2020. Fear, Anger, and Political Advertisement Engagement: A Computational Case Study of Russian-Linked Facebook and Instagram Content. s.l.:Journalism & Mass Communication Quarterly.

Zuboff, S., 2015. Big other: surveillance capitalism and the prospects of an information civilization. s.l.:Journal of Information Technology.

Zuboff, S., 2019. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. s.l.:Profile Books.