Does my research expose individuals? A look into privacy violations in open data publishing

Aakash Sharma
aakash.sharma@uit.no

Keywords: Ethics, Open data publishing, Privacy

1. Introduction

Scientific research has been responsible for changing views of the society and direction of human evolution. The limited focus and lack of collaboration across borders slowed down growth in terms of science. Science has seen tremendous growth in terms of collaboration and research outcomes in the last two decades. Especially with the proliferation of the internet and sharing of research data, new perspectives and understandings have helped science and thus society in general to understand natural phenomena. The speed of progress may have baffled many scientists. Casting doubts on the veracity and reliability of research is not uncommon [1]. Science relies on rigor, reproducibility and transparency. As more and more scientific fields rely on collecting and processing huge amounts of data, it is often considered that scientists are cherry picking the results to support their hypothesis. Automated analytic tools which are capable of finding correlations and novel patterns in big-data have raised doubts about the validation of hypotheses [2].

Open access (OA) has gaining traction in the scientific community [3, 4, 5]. Open access allows anyone to read research publications over the internet. The recommendations [4] ensure that publications from public funded research should be made available in a *machine readable* format online. Unlike restricted by jour-

nals, these articles are available free of charge online. In theory, it encourages quick recognition, appreciation in terms of citations [6] and possibly future collaborations [7, 8]. It reinforces trust among peers and encourage validation from the community as the research output is available for discussion and criticism. However, the advantages of open access have been overstated, according to its critics[8]. The goal of this article is not to discuss OA and its effects, however the risk of privacy leaks when researchers make their raw data available publicly. Through this article, we briefly explain which datasets are problematic and why. In the later part, we present our recommendations, avoiding such leaks.

2. Open Data Publishing

Similar to making scientific publications available online for anyone to read, open access (OA) also encourages scientists to publish their experiment data online. This has been argued as a method to contribute to the community, not just the results, but the opportunity to collaborate and groundwork for future of science. Particularly, Anderson [9] argue that the availability of enormous amounts of data has led to a different method of exploration in science. He further states that with enough computing power of the cloud, and petabyte scale of multidimensional data forces scientists to view data mathematically. Even though before knowing what is relevant in a particular heap of information, scientists can just assume that with better data and tools, it will be possible to model novel behaviors. With heavily relying on statistics, companies like Google are able to perform machine translation without actually understanding the semantics of a language. This deeply rooted reliance on statistical methods with petabyte scale data is able to replace scientific research. And with successful products relying on machine learning models, it is very well demonstrated. Anderson further argues that scientists are trained to recognize that the correlation is not causation. And that one must attempt to understand the underlying mechanisms with help of a model and test it with experiments. The models can then be confirmed or falsified using experiments and observations. Further in the article, Anderson accepts the statistical

methods as the new progression in scientific methods. He argues that correlation with petabytes of data make it enough to be accepted. Anderson concludes by saying that correlation supersedes causation and even without models, unified theory or explanations science can progress.

Pigluicci [10] criticized Anderson's [9] view of using *petabytes* of data to show a correlation as science. Unlike traditional scientific methodologies which relied on 'hypothesize, model and test', data deluge and enormous computing power has helped scientists find correlation in human behaviors through the use of enormous amounts of data. The use of cloud computing and machine learning methods enables scientists to find correlation in Petabytes of data which Pigluicci dismisses as 'Science'. He argue that finding patterns in data is not science, however, finding explanations for those patterns is. He further dismisses Anderson's [9] arguments of his example of renowned scientist in the field of biology by questioning his contribution to understanding of newly identified species. He further argues the that without explanation, data based research is just a statistical blip or noise. He acknowledges the contribution of the newly developed methods to sift through huge amounts of data as part of the process. However, without a scientific explanation of 'why' things happen in a certain way, hypothesis testing, it cannot be regarded as 'science'.

2.1. Relying Too Much on Data?

Whether one agrees with Anderson's view of science or Pigluicci's, it is certain that science today is relying on collecting huge amounts of data. Without even knowing the relation or model, data is being collected and processed to find any underlying correlation with an observation. When investigating successful scientists who publish quite significant amount of papers every year, Ioannidis et al. [11] found out that many of the scientists published papers related to a large cohort study. These studies were epidemiological in nature and collected various aspects of society and its population in them. Similar to the methods described by Anderson [9], these studies also produce multi-faceted multi-dimensional data.

Depending upon the access, many scientists make use of the analytical tools and statistical methods to find data to support their hypothesis. Additionally, these exploratory analyses are also performed on the data to find correlations. The data here plays an important role in validating the research and hence, it becomes crucial for peer reviewers to look at the data. Additionally, arguments from Bailey [1] encourage scientists to be more transparent in research. To achieve transparency, scientists are encouraged to publish not just their findings, but raw data along with it.

França & Monserrat [12] focuses on key points as solutions for the reproducibility crisis in science. Their argument is that progression in science depends on previously learned knowledge. They reiterate the importance of fully understanding previous scientific works and reproducing results. They identify misuse of statistics in various scientific work and hence iterate the necessity to include raw data to check the validity. Often statistical methods are misunderstood and misinterpreted, which can be backtracked with the help of raw data. For example, the practice of using p-values does not convey reliability in the test results [13]. Another aspect reported by França & Monserrat is that scientists choose to publish selected results from an analysis. Often, they do not even provide raw data which might have other significant findings which have been overlooked. This leads to misleading readers, limited reproducibility and increases the producing knowledge based on previously acquired results. They suggest making raw data available along with research protocols to mitigate some issues of the reproducibility crisis. With so much information overload and rapid pace of research output, it is necessary to provide mechanisms to fully understand research methodologies along with the data. It will not be possible with just reading research publications and cherry picked results. Further criticizing the growing pace of data production, França & Monserrat argue that we need to find efficient ways of ingesting prior knowledge. They recommend collaboration as a foundation of expressing and understanding science. Our peers will know something that we missed upon. They

argue equal focus is needed on producing better literature along with supporting data which can be peer reviewed with statistical methods.

We understand that science is going through reproducibility crisis [12] and it has been acknowledged by many researchers in their works. While making science interesting and accessible for everyone is one part of the issue, the other part dealing with reproducibility requires changes in our approaches. As indicated in earlier parts of this section, we need to write better about our approach, methods, statistical calculations and provide raw data for peers to review and confirm. Just publishing out raw data along with research output will not solve the reproducibility crisis. We need to confront this problem at many ends. Sharing raw data is not going to solve it. However, it is often seen as one of the easiest to achieve practice. Sharing data is publicized as a good research practice, but we need to be responsible too. Many institutions around the world *emphasize* on making public funded research more accessible. It might not be immediately obvious that extra time & effort is needed to make data shareable. Often cited is multifold benefits of sharing your data. These include error checking, building trust, citations [6], and potential future collaborations. Additional arguments include speeding up innovation and involving citizens and society [4].

2.2. Data Involving Humans

We previously discussed how important it is to publish data alongside research output. In many fields of science, there can be many hindrances in publishing data alongside the research output. There might be legal challenges such as in case of patient's data. In other cases such as geology, there might be trade secrets hidden along with the raw data. Sometimes, cooperating entities make data available to partners through a common portal. The scientists working in participating groups have access to it and share knowledge and collaborate on publications. There are methods to make it available only with participating/ cooperating research groups or consortium. For example, in case of geological sciences, data might indicate valuable information about a location which the company or the research agency

has acquired through substantial efforts and means. It is in the interest of a research group to safeguard its exclusive access to the data for a fixed period of time. In research, this is safeguarded by regional and national laws. Similarly, for research involving human, in cases where their personally identifiable information is collected, regional ethics [14] apply. These laws describe protection mechanisms for safeguarding individual's privacy. National research councils [15] also describe what information is considered private and sensitive. Even though, people might have a different perception towards what is sensitive information. For example, allergies are considered sensitive information by national data protection agency [16]. However, an individual might not consider it sensitive. Even in case of a data protection agency, there are advices available on how misunderstood anonymization is. De Montjoye et al. [17] discuss in detail how certain attributes, even if anonymized can be used to identify an individual. This is based on the fact that certain attributes pertaining to an individual can be so unique to an individual. Data protection agency [16] considers DNA as private and sensitive for this reason.

This article's scope is limited to the data which involves human, especially in epidemiological studies or small attitude studies based on a small cohort of people. Many epidemiology based research collect data over a population typically over a period of time. These large funded studies have a good amount of procedural guidelines and reviews to ensure good research practices. And the process of data collection, processing and storage is reviewed well before the actual start of studies. The sharing of data in these large studies is accompanied by substantive agreements which cover sharing results and shared responsibilities of the researchers. Since, the data involves tens of thousands of people, it is of regional and national interest to make sure that there are no violations. However, similar resources and well established procedures might not be available when conducting a smaller cohort study.

2.3. Auxiliary Information

In the work of Narayanan & Shmatikov [18], they describe auxiliary information as any information that can reveal bits of information about the targeted individual. This information can be obtained by knowing about the person, data from internet, social media profiles, and in some cases national census data [19]. Auxiliary information about individuals was not a problem before the advent of the internet. We tend to share more and more information about us online [20, 21]. It is computationally much easier to find about a person without their knowledge. We are well aware of internet companies tracking our activities online. While the success of targeting might be questionable, there is a high chance auxiliary information improves the accuracy in targeting. For example, based on internet usage pattern, let us assume that an internet company has identified ethnicity of individuals in a certain block to some accuracy. Combining this information with census data [19] which publishes statistics on ethnicity per block level exposes those individuals. Targeting based on ethnicity is not something that is legal or acceptable to individuals. However, to avoid such risks we need to limit information. Later in this article (Section 5), we discuss methods to avoid disclosing information which can be a privacy risk.

3. Incidents

We now look at two incidents in which open data led to re-identification and ended up in withdrawal of the dataset. Similar to Banobi et al. [22] arguments about rebuttals, dataset withdrawals also affect the course of science. In order to avoid them, it is better to be proactive in addressing risks before publishing a dataset.

3.1. Facebook Profile Dataset

We now look at one of the incidents in research where published data led to privacy violation of hundreds of college students. In 2008 [23], a group of researchers published data collected from Facebook accounts of an entire cohort.

In the cohort, researchers collected data from college students over 4 years of their college life, including their social relationships evident from their online facebook activities. The collected data was released in four stages, one each year. According to the declaration by the researchers, precautionary steps were taken to cleanse the data to ensure protection of identity and privacy of the participating students. Even though their procedures were approved by supervising institutional ethics committee, the protection was not enough. As with any other research data, the data published was machine readable to allow other researchers to reuse it. As we will see in multiple examples, compliance not necessarily protects from all attacks. It only minimizes the risk. Even though the cleansing of the data was done in a good faith, it was not enough to protect it from re-identification. Though the ”terms and conditions” for using the dataset included a clause for researchers to prohibit re-identification, it could not prevent it. Within few days of the data release, the college was identified and hence the students. It was only a matter of time when individuals can be linked to their data using auxiliary information (see Section 2.3) available on their social media profiles.

3.2. Netflix Movie Review Dataset

De-anonmyzation is not enough [18]. Narayanan & Shmatikov [18] demonstrated how anonmyzed user ids in a sparse dataset can be used to de-anonmyze individuals. WIth some extra knowledge about the participating subject/user, it is possible to identify an individual. The attacks on the Netflix database demonstrated by Narayanan and Shmatikov demonstrate how vulnerable public datasets are. They also present model for de-anonymizing *sparse* datasets. Sparse here means that even though the number of participants in a dataset can be a large number, the relative data points per individual are very few. This happens in most cases where researchers try to understand very specific theme about individuals with their limited surveys. Even in case of large cohorts, if the published data is a subset of the overall data, it will be susceptible to de-anonymization attacks. The algorithm presented in their work builds up on auxiliary information about individuals and slowly identifying the data points which can identify an individ-

ual from the dataset. The dataset used by Narayanan and Shmatikov was provided by Netflix as part of their contest to improve movie recommendations. Similar to the case presented by Zimmer [23], Netflix said that they have stripped off any customer information from their dataset. According to Netflix's privacy policy it was sufficient. Additionally, they only provided a small subset of their total data $\tilde{1}0\%$, so they claimed that it would not be an issue. However, the method published is still effective to de-anonymize individuals based on their movie review date and ratings. Surprisingly, it only took a day to identify the college the students belonged to. Even though the names were omitted, the attributes such as courses taken in the database led to re-identification of the university and hence the partial identification of the individuals. Similarly, in cases where there was only one student from a country, this led to direct identification. This can further lead to identification of social groups and further identification of individuals as a chain reaction. The researchers argued that the data was already made available by students on their profile. However, the ease at which the university and the individuals were identified was troublesome. This further led to withdrawal of the dataset.

4. What Do The Guidelines Say?

The examples of privacy leakages highlighted in the works of Zimmer [23] and Narayanan & Shmatikov [18] are not the only examples of privacy leakage in public datasets. These examples show that it is not just the academia which has less rigorous processes to de-anonymize data. Even in case of industry [18], practices are insufficient to protect from such attacks. Simple anonymizing of data or removing individual's information is not enough protection against de-anonymization attacks.

Institutional guidelines [24] about how to deal with research data, including publication cover various aspects of the research data lifecycle. The lifecycle consists of the methods to gather research data, to store process and publish. The

recommendations emphasize on having a research data plan which covers various aspects and ethical considerations while collecting, analyzing, processing and publishing data. The researcher is encouraged to make data available for other researchers along with internationally recognized licenses and standardized metadata to assist reuse. In cases where personal data is collected, a researcher is required to seek approval of the methods and plan for information security in accordance with the regional law. In the guidelines [24], one of the statements reads

"The researcher shall make the research data openly available for future use by all relevant users, providing this is not prevented by any legal, ethical, security, or commercial reasons."

While an academic institution provides guidance and assistance in fulfilling these requirements, it is up to the local or regional ethics body to approve a data management plan. A data management plan outlines the standardized methods and procedures defined to be compliant with ethical practices.

The national data protection agency [16] provides a set of guidelines and tools for data anonymization. The guidelines explain the differences and challenges in data anonymization. Pseudonymisation is replacing directly identifiable parameters with with pseudonyms. For example, replacing "David" in the field *name* with "subject A" in a dataset, while keeping all other parameters unchanged is not anonymization. Based on the auxiliary information available about "David" in the dataset and in the public domain, it is theoretically possible to identify "David". This is what happened in the two incidents [23, 18] we discussed earlier. This shall not be confused with anonymization when dealing with data. Similarly, one must not think that encryption is a better solution over anonymization. Encryption only encrypts the dataset in a way that only authorized person/organization can have access to it. The key is only shared with a limited set of identified individuals or teams. However, once decrypted, the data can be copied and distributed without any control. Even though, the guidelines available at institute [24], regional [14] and national level [16, 25] warn against that. It is not always possible to protect

unauthorized copying of the decrypted data, which might take place physically or remotely. Hence, anonymization is highly recommended at institutional and national levels to protect the privacy of individuals. In European Horizon 2020 program [4], it is encouraged to publish data. However, the guidelines provide an exception if you cannot guarantee the protection of individual's privacy. So, privacy is preferred over openness in the guidelines.

5. Recommendations

In earlier sections, we described the case for risking individual's privacy based on publishing dataset. Fiesler & Hallinan [26] work on understanding perceived data sharing and privacy violations gives an insight to sustain research. The work on cost-benefit analyses, role of privacy policies and trust for platforms can provide researchers and research data hubs guidance for increasing trust. Unlike social networks whose benefits are fairly evident to a user, participation in a research study might be difficult for people. Similar to the findings of Fiesler & Hallinan, many users might ignore privacy violations as they have in recent times. However, it is responsibility of the researcher and the organizations to minimize those risks for an uninformed user. We will now describe two of our recommendations for ensuring that individual's privacy is not risked while releasing a dataset. The recommendations are not a replacement for the guidelines for publishing sensitive data but rather complimenting them.

5.1. National Research Data Archives

Many institutions maintain their own archives for research data. Traditionally, affiliated researchers at Universities deposit their research data in those archives or might choose to self-archive. However, the practices or safeguards at the institutional level might not be as stringent as a national research data archive. The national archives have an application process where a researcher seeking to obtain access discloses information about the project and goals. For example, Tromsø study data can be obtained through a standardized application process [27]. Additionally for sensitive data, national archives can provide secure platforms [28] for

data analysis without giving a free access to data. While this may add additional cost to the project, in the long term it is possible to support initiatives that ensure the privacy of everyone involved in research data. Also, national archives can act as collaborative platforms for individuals or groups when working on similar research goals.

5.2. Differential Privacy

Dwork et al. [29, 30, 31] introduced the concept of Differential Privacy. It is a statistical technique that is designed to maximize accuracy in statistical queries while reducing the privacy individual information in the database. It works by adding a certain degree of noise to the result. The noise prohibits individual targeting while attempting to provide accurate information for statistical measures. The ratio of the noise added to different queries is calculated based on the exposure of individual information. It has been widely adopted at government agencies [19] as one of the important techniques to protect privacy of individuals. Abowd et al. [19] highlight the importance of the use of differential privacy in the United States Census. He compares computer scientists to social scientists in approaching accuracy and privacy. While computer scientists settle for lower accuracy to save privacy, social scientists might prefer accuracy over privacy. Abowd et al. [32] argue that statistician and open research data publish too accurately even though it is not needed at a policy level. Statistics need not to be too accurate as they account for risks in the calculation. Abowd et al. endorses using differential privacy when publishing data as it retains statistical accuracy while limiting exposure of individual's data.

6. Conclusion

In open access (OA) and open data publishing, publishing raw data is often seen as a low-hanging fruit. We explained how careless open data publishing can lead to privacy leaks in public datasets using examples from academia & industry. We analyzed data publication guidelines available at institutional and regional

level. In order to minimize privacy leaks, we provide our recommendations which are complimentary to the existing guidelines. Individuals have largely ignored privacy breaches covered in media. However, their attitude towards privacy might change over time. As researchers continue increasing society's participation in research, safeguarding the privacy will go a long way in terms of enabling trust between researchers and society. After all, it is an ethical and moral responsibility of a researcher.

References

[1] R. Bailey, "Broken science," vol. 47, no. 9, p. 18, 2016.

[2] R. Kitchin, "Big data and human geography: Opportunities, challenges and risks," *Dialogues in human geography*, vol. 3, no. 3, pp. 262–267, 2013.

[3] RCN, "Open access to researech data." `https://www.forskningsradet.no/en/Article/Open_access_to_research_data/1240958527698?lang=en`, 2015.

[4] E. COMMISSION, "Guidelines to the rules on open access to scientific publications and open access to research data in horizon 2020." `https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf`, 2017.

[5] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, *et al.*, "The fair guiding principles for scientific data management and stewardship," *Scientific data*, vol. 3, 2016.

[6] B. E. Bierer, M. Crosas, and H. H. Pierce, "Data authorship as an incentive to data sharing," 2017.

[7] J. Willinsky, *The access principle: The case for open access to research and scholarship*, vol. 559. MIT press Cambridge, MA, 2006.

[8] G. Eysenbach, "Citation advantage of open access articles," *PLoS biology*, vol. 4, no. 5, p. e157, 2006.

[9] C. Anderson, "The end of theory: The data deluge makes the scientific method obsolete," *Wired magazine*, vol. 16, no. 7, pp. 16–07, 2008.

[10] M. Pigliucci, "The end of theory in science?," *EMBO reports*, vol. 10, no. 6, pp. 534–534, 2009.

[11] J. P. Ioannidis, R. Klavans, and K. W. Boyack, "Thousands of scientists publish a paper every five days," 2018.

[12] T. F. França and J. M. Monserrat, "To read more papers, or to read papers better? a crucial point for the reproducibility crisis," *BioEssays*, vol. 41, no. 1, p. 1800206, 2019.

[13] A. Gelman, "Statistics and research integrity," *European Science Editing*, vol. 41, no. 1, pp. 13–14, 2015.

[14] REC, "Regional committees for medical and health research ethics (rec)." `https://helseforskning.etikkom.no/`, 2019.

[15] RCN, "Research council of norway." `https://www.forskningsradet.no/`, 2019.

[16] "Datatilsynet." `https://www.datatilsynet.no/en/`, 2019.

[17] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, p. 1376, 2013.

[18] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 ieee symposium on security and privacy*, pp. 111–125, IEEE, 2008.

[19] J. M. Abowd, "The us census bureau adopts differential privacy," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2867–2867, ACM, 2018.

[20] R. Belk, "You are what you can access: Sharing and collaborative consumption online," *Journal of business research*, vol. 67, no. 8, pp. 1595–1600, 2014.

[21] B. Debatin, J. P. Lovejoy, A.-K. Horn, and B. N. Hughes, "Facebook and online privacy: Attitudes, behaviors, and unintended consequences," *Journal of computer-mediated communication*, vol. 15, no. 1, pp. 83–108, 2009.

[22] J. A. Banobi, T. A. Branch, and R. Hilborn, "Do rebuttals affect future science?," *Ecosphere*, vol. 2, no. 3, pp. 1–11, 2011.

[23] M. Zimmer, "but the data is already public: on the ethics of research in facebook," *Ethics and information technology*, vol. 12, no. 4, pp. 313–325, 2010.

[24] UiT, "Principles and guidelines for research data management at uit." `https://uit.no/Content/532111/cache=20172206150343/Policy_forskningsdata_UiT_090317.pdf`, 2017.

[25] NSD, "Norwegian center for research data." `https://nsd.no/nsd/english/pvo.html`, 2019.

[26] C. Fiesler and B. Hallinan, "We are the product: Public reactions to online data sharing and privacy controversies in the media," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 53, ACM, 2018.

[27] UiT, "How to apply for data access." https://en.uit.no/prosjekter/prosjekt?p_document_id=71247.

[28] U. of Oslo, "Tsd service for sensitive data." https://www.uio.no/english/services/it/research/sensitive-data/about/index.html, 2019.

[29] C. Dwork and J. Lei, "Differential privacy and robust statistics.," in *STOC*, vol. 9, pp. 371–380, 2009.

[30] C. Dwork, G. N. Rothblum, and S. Vadhan, "Boosting and differential privacy," in *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pp. 51–60, IEEE, 2010.

[31] C. Dwork, "Differential privacy," *Encyclopedia of Cryptography and Security*, pp. 338–340, 2011.

[32] J. M. Abowd, I. M. Schmutte, W. N. Sexton, and L. Vilhuber, "Why the economics profession cannot cede the discussion of privacy protection to computer scientists," Presented at the Allied Social Science Association Meeting 2019, 2019.