A Letter to the Journal of Statistics Education - A Call for Review of "OkCupid Data for Introductory Statistics and Data Science Courses" by Albert Y. Kim and Adriana Escobedo-Land

Tiffany Xiao [1] , Yifan Ma [1]

1 Statistical & Data Sciences Department, 1 Chapin Way, Northampton, MA, 01063

As Big Data continues to rise in popularity, so does an increased need for protection against potential misuses of data. We are a group of undergraduate Statistical and Data Science major students from Smith College that are actively engaged in ethical discussions concerning the use of data in our society. It can be challenging to predict future trends and technologies in data science that could cause concerns. However, we believe that some essential protections and procedures should be in place to help prevent misuses of data. In particular, we are writing to you to address our concerns with the paper "OkCupid Data for Introductory Statistics and Data Science Courses" by Albert Y. Kim and Adriana Escobedo-Land that was published in your journal [1]. In light of ethical concerns surrounding the paper, herein we describe the background of how the data set was found to contain identifiable information. We communicated this to the authors, who correspondingly corrected the manuscript.

In our opinion, there is no doubt that the dataset presented in the paper holds pedagogical values as well as research values. One aspect of the educational value of the dataset is the fact that the context of possible analysis could better drives students' interests. The research value of the data lies within the self-reported nature of the dataset, which usually is the private property of corporations and could be hard to obtain for researchers in universities. Another context in which the pedagogical value of the dataset remains is where students could use this as a case study in discussions of the ethical implications of such data, even practice anonymization skills to the data. However, we do believe that for the dataset to be used for pedagogical purposes, further anonymizations to the dataset were necessary.

Some ways that datasets like this one could be better anonymized in the future include removing unimportant variables that have identification power disproportionate to their value to research. For example, in the case of the OKCupid dataset associated with the paper, the time the data was collected could be removed, since this fact is not particularly essential but can be used for identification. Other sources of concern for this dataset are the variables that reveal geographical and temporal information on individuals. Another method could be to introduce noise, though this has some inherent issues, especially for a dataset as rich as this one. Introducing noise to the essays, for example, would involve some words being changed, moved, or added/removed. However, if the dataset is being used to learn text-mining or in text-mining adjacent research, this might significantly impact the usability of the data. So, another possibility would be removing the data from the web and make it available through secure data transfers to students/researchers upon request.

Lastly, we recognize that many papers and datasets may pass requirements in the eyes of current legislation. However, many of the current laws concerning misuse of technology are outdated. For example, an article focused on a mathematician who had scraped data from OKCupid to find his ideal match suggested that the mathematician

violated the Computer Fraud and Abuse Act (CFAA) [2]. The CFAA is a 1984 bill created to protect against computer fraud that was inspired by the dated 1983 movie War Games [3]. In 2016, a group of researchers again scraped data from OkCupid and publicly published it without proper anonymization, causing outrage from the public for a clear violation of privacy [4]. The researchers repeatedly argued that they were not violating any laws and thus were allowed to publish the data. There is an evident lack of legislation that adequately protects individual rights against the misuse of their data in the United States; thus we cannot rely on current legislation to adequately address our ethical concerns.

With a lack of proper legal protections for individuals' data, it is our responsibility as members of the data science community to protect individuals' data from misuse. One way we can do this is by carefully considering the requirements we place when reviewing papers and datasets for approval. In the case of the paper and its associated dataset by Kim and Escobedo-Land, there was a clear need for further anonymization before it can be deemed ready for release to the public. We appreciate that you have considered our recommendations and corrected the work, while also acknowledging its high potential for educational use and the work that the authors put into creating it. Furthermore, we hope the points we outlined will help you create a future framework for reviewing work in the context of the protection of individual privacy.

References

1. Kim AY, Escobedo-Land A. OkCupid data for introductory statistics and data science courses. Journal of Statistics Education. Taylor & Francis; 2015;23: null. doi:10.1080/10691898.2015.11889737

2. Penenberg AL. Did the mathematician who hacked okcupid violate federal computer laws? [Internet]. Pando. 2014. Available: `https://pando.com/2014/01/22/did-the-mathematician-who-hacked-okcupid-violate-federal-computer-laws/`

3. Pollaro G. Disloyal computer use and the computer fraud and abuse act: Narrowing the scope. Duke Law & Technology Review 1-12. 2010; Available: `https://scholarship.law.duke.edu/dltr/vol9/iss1/11/`

4. Hackett R. Researchers caused an uproar by publishing 70,000 okcupid users' data [Internet]. Fortune. Fortune; 2016. Available: `https://fortune.com/2016/05/18/okcupid-data-research/`