This FAQ was written doing the media circus in approximately May 2016. I have since updated it slightly with some recent updates since then.

FAQ about the OKCupid dataset

Emil O. W. Kirkegaard

Julius D. Bjerrekær

There are many articles about this research on newspapers, but many of them contain false information. The misinformation spreads because journalists copy each other instead of fact-checking for themselves. To counter this, we provide this FAQ.

1. Was any hacking involved?

No hacking – the use of computer skills to breach security holes in the IT infrastructure – was used to get the data. Everything that was done could be done by a chimp using copy/paste and a large spreadsheet.

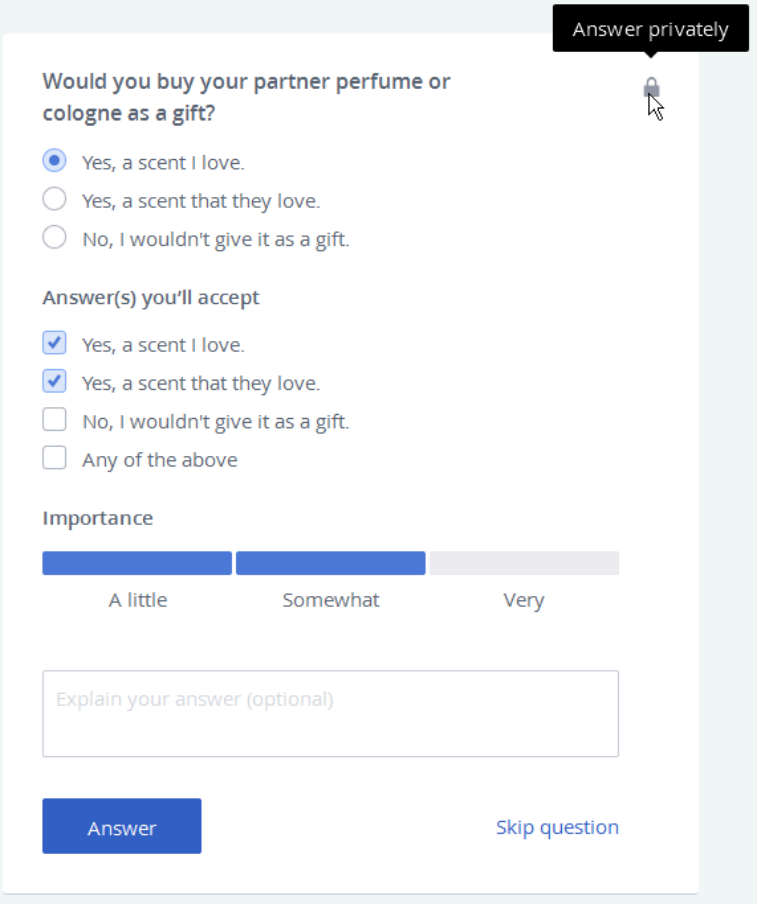
2. Are the data public?

This depends on the definition used, but in our opinion *yes*. The profile information of many users can be freely seen from Google. This includes pictures, age, gender, sexual identity and the profile text. To see users' answers to questions, however, one must have answered the same question. This means that one must be logged in with a user that has answered that question.

OKCupid itself clearly states in their terms of service that the information may be public:

> **Privacy**
> *You should appreciate that all information submitted on the Website might potentially be publicly accessible. Important and private information should be protected by you.* We are not responsible for protecting, nor are we liable for failing to protect, the privacy of electronic mail or other information transferred through the Internet or any other network that you may utilize. See Humor Rainbow's privacy policy for more information regarding privacy. The privacy policy is incorporated into and a part of these Terms of Use. [our emphasis]

Furthermore, when users answer a question, they get the option to answer the question privately as seen below:



Most users *do not choose to answer privately*. We did not and *could not* scrape the private answers because they are not possible to see for others.

In our opinion, because the requirement is solely a free user and that one has answered the question oneself, the answers are also considered public. Furthermore, there is no reasonable expectation of privacy of these answers for the users because 1) they agreed to that the information might be publicly accessible, 2) and they did not use the option to answer privately when given the choice.

One person compared the situation to going to a swinger club and taking pictures. This is not comparable because there is a clear expectation of privacy in that case. People already know that stuff that is done on the Internet stays there. It is more akin to going to a regular pub and taking pictures. This is explicitly considered legal in most places since there is no expectation of privacy there.

3. Why did you publish the usernames?

There were two reasons to do this.

First, we forgot to scrape some of the information such as the profile text (a critical oversight!) and if one has the usernames, one can do this at a later point provided the user is still there. Second, the usernames themselves are an interesting topic of research. Usernames play a crucial part in a person's presentation and so are not randomly chosen. One can thus

research what predicts choice of username. For instance, do people who include "hot" in their username see themselves as more attractive? Many users use animals in their names, are people who chose the same animal more similar than people who don't?

A third reason to publish the usernames, which was however not considered before the release, is that if one does not have the usernames, the dataset cannot be combined with other datasets scraped from the same site, as this would likely result in duplicated users to an unknown extent. Many such other datasets already exist and many will undoubtedly be collected in the near future (see Section 8).

It is possible that the usernames will be removed in a future version of the dataset as one may argue that the three scientific goals above do not outweigh the privacy concern from the usernames being available.

Note that usernames are pseudonyms – not people's real names (unless people used their own names which seems to be extremely rare).

3.1. Is it a special problem that the data will be available for all eternity?

Some have argued that while scraping usernames themselves is not problematic, it is a problem in this case because the data will be available for eternity (or so). In contrast, while all the data in the dataset are visible on the site, it will no longer be visible once the user deletes their user. In this way, the user is thus in control of the information.

This is another instance of a well-debated topic: the right to be forgotten. In 2014 an EU court vindicated a law that makes it possible for persons to require that websites remove stories about them, thus making the internet “forget them”. While in some cases, this can be ethically defensible (e.g. in cases of wrongful convictions where employers and others can find the media reports about the guilty conviction but not the overturning). However, as predicted, many uses of these laws are used by people trying to hide their dubious past. For instance, one doctor had Google remove 50 links to stories about his prior botched operations.

Furthermore, Internet content is routinely being stored for eternity by the Internet Archive. One can access old versions of websites using the Wayback Machine. This includes comments made by persons, often in their own name, that they cannot take down themselves.

Thus, it is not a special problem for this dataset that the data will be available for the foreseeable future.

4. Why did you not scrape the profile pictures?

The privacy concerns with profile pictures are more prominent for pictures. While they are also anonymous in the sense that they do not contain the real name of the user, one can possibly find the user's real name by searching on Facebook using the profile information

(age, gender, maybe first name, university affiliation and so on).

The above task is arduous and would be unlikely to be done for many users. However, with the advent of automatic facial recognition software, it would be possible to automatically attempt to link all the users in the dataset with Facebook and hence de-anonymize them.

On the other hand, profile pictures are an extremely rich resource for scientific studies. One can use automatic rating software to assess e.g. attractiveness which can then be examined in relationship to other profile information.

Finally, there is the technical profile of file size. Many profiles contain multiple photos and downloading them for a large number of users will take up a lot of local storage.

On balance, it is therefore a difficult ethical question whether one should release profile pictures if one could scrape them. *The released dataset does not contain profile pictures*.

5. Was the paper published?

No, the paper was submitted for review. As per the journal policy, all datasets used for papers must be publicly released on submission. This allows reviewers and everybody else to verify the analyses and do other analyses immediately from submission. This is done because research shows that “available on demand” policies do not work (1, 2).

The paper was published on 3rd November 2016, about half a year after the media circus.

6. Was this research related to Aarhus University / Aalborg University?

The research was carried out in private and has no direct connection to our universities. They are not responsible for any of our actions and neither can they claim credit for or be blamed for our actions because we did not learn anything related to this project at the universities. Neither did they provide any guidance or resources for this research project and we did not ask for any.

6.1. But if the research was private, why did you state your university affiliations?

It is common practice to state one's university affiliation if one is a student and we are both Master's students. So, we are affiliated with the universities because we are students there, but the universities had nothing to do with this particular piece of research. Do note that these relationships are not transitive, so one cannot conclude that the universities were involved in the research just because we were students at the universities.

6.2. Did you get IRB approval for this study?

IRB is only for university based research. This was private research so there is no such thing as IRB.

7. Was Oliver Nordberg part of the project?

Yes and no. Oliver helped write the scraper but did not take part in writing the paper, compile the dataset from the scraper output, code the variables or do the research presented in the release paper. Given the negative media attention, he would prefer not to be mentioned.

8. Is the dataset copyrighted? Was the scraping legal?

We thought this was an obvious case of public data scraping so that it would not be a legal problem. The only legal advice came from asking a friend who took some classes at a business school. His judgment was that it was probably legal.

The scraping was done in Denmark, not the US, but OKCupid is owned in the US. It's not clear to us what this implies for this case.

The US copyright law explicitly does not cover databases. The questions/answers themselves are presumably not copyrightable due to being unoriginal and very short.

Scraping datasets is a common procedure used by many people. For instance, flight ticket meta-search engines search the airlines' websites to find prices for rides, car sellers scrape Auction houses to find cheap offers and estimate the current prices for models, housing search platforms scrape real estate websites. In the realm of science, Facebook, Twitter, ResearchGate, academia.edu, Mendeley, and so on have already been scraped many times and the datasets examined by researchers. ResearchGate itself scrapes the internet and uses the information to automatically create user profiles for scientists.

With regards to OKCupid, many people have scraped this website before but no one else, as far as we know, have made the data publicly available. These other scrapings resulted in fairly minimal uproar when they were covered in the media:

> 2014, Jan. How a Math Genius Hacked OkCupid to Find True Love
>
> 2015, Sep. Women on OKCupid Don't Seem to Think Their Jobs Are Much of a Selling Point

One can find details about other programmers scraping the same site:

> 2011, Aug. Scraping OkCupid: Will Bot For Love
>
> 2012, July. Data mining OKCupid
>
> 2014, Feb. OKCupid hacker scripts, data, etc This dataset was seemingly publicly available, but is not currently available.
>
> 2015, Mar. Hacking OKCupid for More Efficient Dating
>
> 2016. May. OkCupid, 2.8million profiles out of 3.2million tested, open to the internet by default!

(many more!)

Finally, there are many available online tools to scrape this website (one here) and people asking about scraping it. Many people approached us privately during this conflict stating their support. A few stated that they had themselves scraped data from the site beforehand.

One site apparently even had OKCupid datasets for sale to interested buyers. OKCupid itself may have been selling or inadvertently leaking data to third parties, which also happens because the site does not use proper site security. It also appears that OKCupid gave some data to a researcher in 2013.

Scraping this kind of site is a perfectly normal thing to do for people with computer science skills and does not normally result in such outcry.

The Danish data protection agency sent Emil a letter asking questions about the case. However, they never filed any charges and dropped the case on 12th July 2017 (their final letter is here, in Danish).

9. Should one need consent to scrape data?

Some have claimed that consent is needed to participate in a study. While universities frequently have rules about this, this research was not done at a university (See Section 6) and is therefore not covered by university rules.

There are laws that mandate consent in research. For instance, the Nuremberg Code (written after WW2 to prosecute Nazi or their doctors who carried out involuntary experiments on inmates of concentration camps) has clauses about consent. This law was not known to us before someone pointed it out during the media attention. However, it is clear that it does not apply to this study because the code is about "*permissible medical experiments*" and this study was neither medical nor an experiment. It was an observational study similar to going into a bar and counting the characteristics of people by observation (such studies have been done too).

Furthermore, OKCupid itself performs experiments on users without asking for consent. These are detailed on their own blog and in their book. They specifically used this approach to get more users as detailed by their own CEO in a talk he gave in 2012.

Update 2019

A list of works based on OKCupid datasets. Some of these are not based on the version we released, but others. One can find these by searching for "okcupid" in Google Scholar limited to 2017 and foward or from before 2016. Both of these avoid the media articles published in 2016, which pollute the results.

- 2012

    - study https://experts.illinois.edu/en/publications/not-just-a-wink-and-smile-an-analysis-of-user-defined-success-in-
- 2013
    - study http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.671.3092
- 2014
    - book https://www.google.com/books?hl=da&lr=&id=-xEcAwAAQBAJ&oi=fnd&pg=PT10&dq=%22okcupid%22&ots=ydM182knjA&sig=w_EseOSD65w1Jljp5XAGG_YOdc4
    - data release, data from 2014 https://github.com/wetchler/okcupid
- 2015
    - data release paper https://amstat.tandfonline.com/doi/abs/10.1080/10691898.2015.11889737 https://github.com/rudeboybert/okcupiddata
    - Study https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4533555/
- 2018
    - student thesis http://fau.digital.flvc.org/islandora/object/fau%3A41255
    - student thesis http://fau.digital.flvc.org/islandora/object/fau%3A41246
    - public scraper https://github.com/stevendevan/okcscrape3
- Other
    - code to automate okcupid behavior https://github.com/tranhungt/okcupidjs
    - code to send/receive messages https://github.com/lehrblogger/OkCupid-Message-Downloader
    - code to send/receive messages https://github.com/tarikjn/OkCupid

Data science report:

- https://www.kimanalytics.com/single-post/2017/08/05/OkCupid-Analysis

OKC

Harvard's Privacy Meltdown - The Chronicle of Higher Education

Hacking OKCupid for More Efficient Dating

Data mining OKCupid | andrewmatteson.com

Scraping OkCupid: Will Bot For Love - Stephen Lee Fischer

yakamo.org

Because it's Friday: Religion and reading level

OkCupid: Finding your Valentine with R

The limits of racial prejudice

Infochimps Climbs the Tree of Success with New CEO Jim Kaskade | SiliconANGLE

A study of user behavior on an online dating site