Ethics and privacy in social media data analysis

Max Pellert

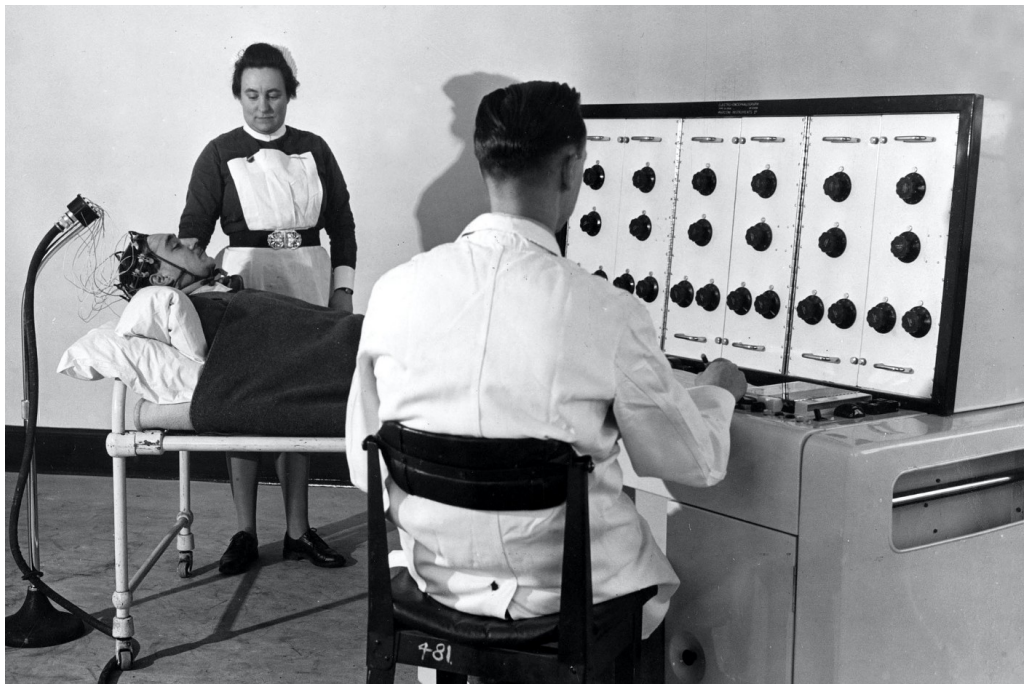
University of Konstanz

Social Media Data Analysis

Outline

1. Ethics in social media research
2. Privacy issues of digital traces
3. Discussing a recent case

When social research can do harm



Milgram's authority experiment (1961)

Zimbardo's prison experiment (1971)

Research ethics

- **Beneficence and No harm principle:** The purpose of research is to discover new information that would be helpful to society. The purpose of research should never be to hurt anyone or find out information at the expense of other people.

- Researchers often seek to evaluate benefits and harms of their research to indicate that **benefits greatly outweigh potential harms**.

- **Institutional Review Boards (IRB)** for ethics: University processes to safeguard ethical principles in academic research. Scientists submit research designs and the board reviews them and makes a statement about whether the design respects the ethics regulations of the university.

First do no harm: An exploration of researchers' ethics of conduct in Big Data behavioral studies. Favaretto et al (2020)

Ethics Committee at the University of Konstanz

Universität Konstanz

Ethics Committee

University > *Administration and organisation* > *University bodies and committees* > *University bodies for scientific integrity* > ***Ethics Committee***

- **Rectorate** +
- **Senate**
- **University Council** +
- **University bodies and committees** −
  - Overview of the committee structures
  - Rectorate, Senate, University Council
  - University of Konstanz committees

Ethical aspects of research projects

The Ethics Committee advises researchers whose projects involve experiments on humans that might affect their health, dignity or personal rights.

https://www.uni-konstanz.de/en/university/administration-and-organisation/university-bodies-and-committees/university-bodies-for-scientific-integrity/ethics-committee/

Example reference: the Declaration of Helsinki

WMA DECLARATION OF HELSINKI – ETHICAL PRINCIPLES FOR MEDICAL RESEARCH INVOLVING HUMAN SUBJECTS

PDF

Adopted by the 18th WMA General Assembly, Helsinki, Finland, June 1964
and amended by the:
29th WMA General Assembly, Tokyo, Japan, October 1975
35th WMA General Assembly, Venice, Italy, October 1983
41st WMA General Assembly, Hong Kong, September 1989
48th WMA General Assembly, Somerset West, Republic of South Africa, October 1996
52nd WMA General Assembly, Edinburgh, Scotland, October 2000
53rd WMA General Assembly, Washington DC, USA, October 2002 (Note of Clarification added)
55th WMA General Assembly, Tokyo, Japan, October 2004 (Note of Clarification added)
59th WMA General Assembly, Seoul, Republic of Korea, October 2008
64th WMA General Assembly, Fortaleza, Brazil, October 2013

6th September 2022

Policy Types

Declaration

Archived Versions

» DoH-Jun1964

» DoH-Oct1975

» DoH-Oct1983

https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

Ethics of social media research

Beyond the No harm principle:

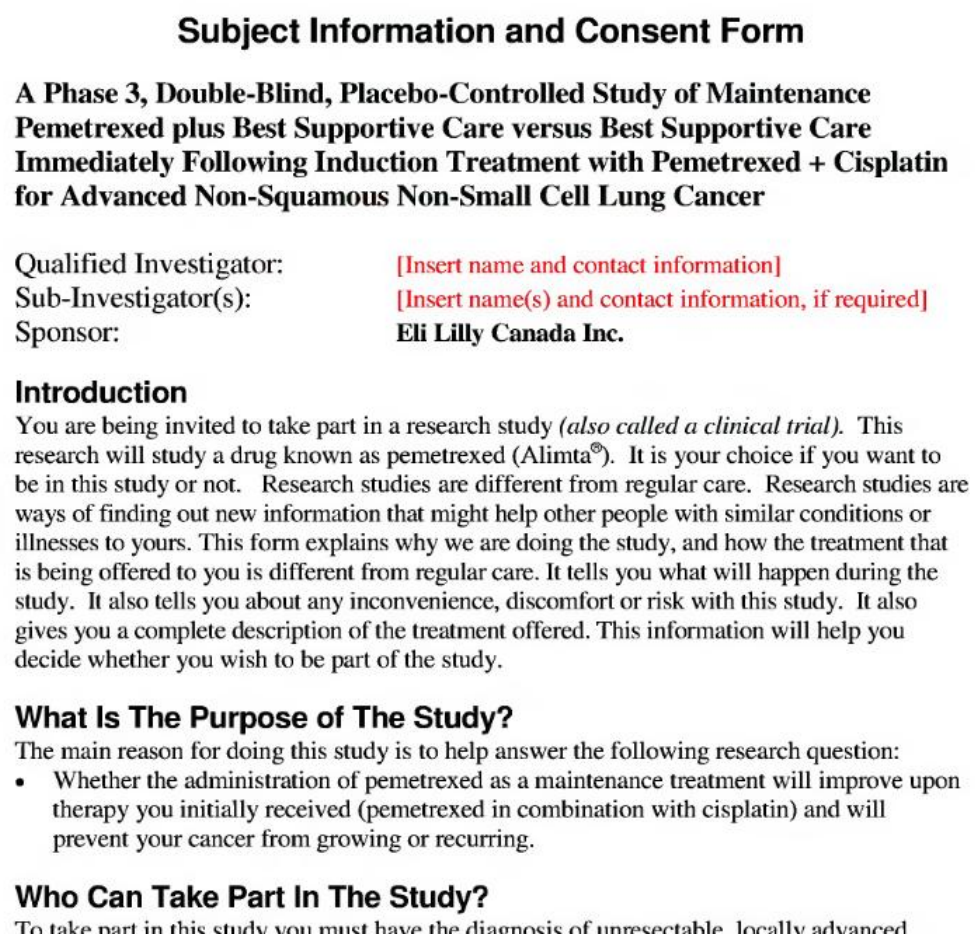
- Respecting privacy expectations of **data subjects**: Data are people
- Downstream consequences of technological development
- Right to information in the digital society

There is no *magic rule* or *solution* in research ethics: Every research design has to be evaluated in detail and its risks and benefits considered carefully

How to do that is evolving over time, building on examples that expose research ethics issues. Here we are going to see a few that have shaped current research ethics in the analysis of social media data and other digital traces

Internet Research: Ethical Guidelines 3.0. Association of Internet Researchers (2019)

Preventing harm: Informed consent



Subject Information and Consent Form

A Phase 3, Double-Blind, Placebo-Controlled Study of Maintenance Pemetrexed plus Best Supportive Care versus Best Supportive Care Immediately Following Induction Treatment with Pemetrexed + Cisplatin for Advanced Non-Squamous Non-Small Cell Lung Cancer

Qualified Investigator: [Insert name and contact information]
Sub-Investigator(s): [Insert name(s) and contact information, if required]
Sponsor: **Eli Lilly Canada Inc.**

Introduction

You are being invited to take part in a research study *(also called a clinical trial)*. This research will study a drug known as pemetrexed (Alimta®). It is your choice if you want to be in this study or not. Research studies are different from regular care. Research studies are ways of finding out new information that might help other people with similar conditions or illnesses to yours. This form explains why we are doing the study, and how the treatment that is being offered to you is different from regular care. It tells you what will happen during the study. It also tells you about any inconvenience, discomfort or risk with this study. It also gives you a complete description of the treatment offered. This information will help you decide whether you wish to be part of the study.

What Is The Purpose of The Study?

The main reason for doing this study is to help answer the following research question:

- Whether the administration of pemetrexed as a maintenance treatment will improve upon therapy you initially received (pemetrexed in combination with cisplatin) and will prevent your cancer from growing or recurring.

Who Can Take Part In The Study?

To take part in this study you must have the diagnosis of unresectable, locally advanced

- A way to verify that harm is reduced is to check that research subjects consent to participate in the experiment
- Gathering that consent in an informed way can be challenging
  - Long forms with jargon
  - Unconscious participants
  - Non-invasive or in vivo research

Users' Views of Ethics in Social Media Research: Informed Consent, Anonymity, and Harm. Williams et al, 2017

When can you say that consent was informed?

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock
+ See all authors and affiliations

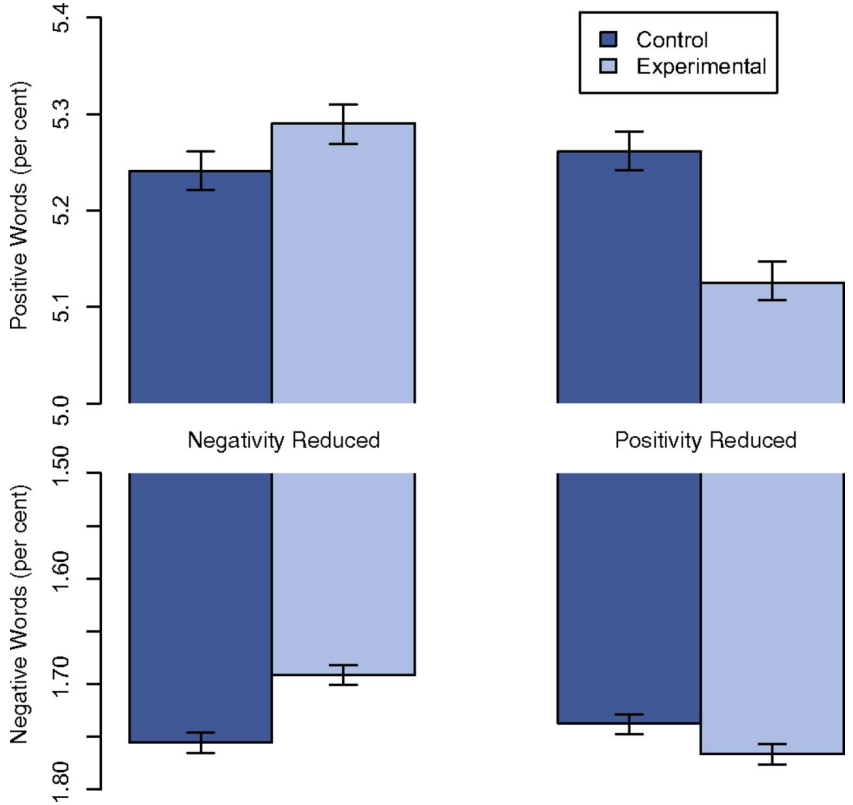
PNAS June 17, 2014 111 (24) 8788-8790; first published June 2, 2014; https://doi.org/10.1073/pnas.1320040111

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

This article has Corrections. Please see:

Editorial Expression of Concern: Experimental evidence of massivescale emotional contagion through social networks - July 03, 2014

Correction for Kramer et al., Experimental evidence of massive-scale emotional contagion through social networks - July 03, 2014



Experimental evidence of massive-scale emotional contagion through social networks. Kramer et al, PNAS (2014)

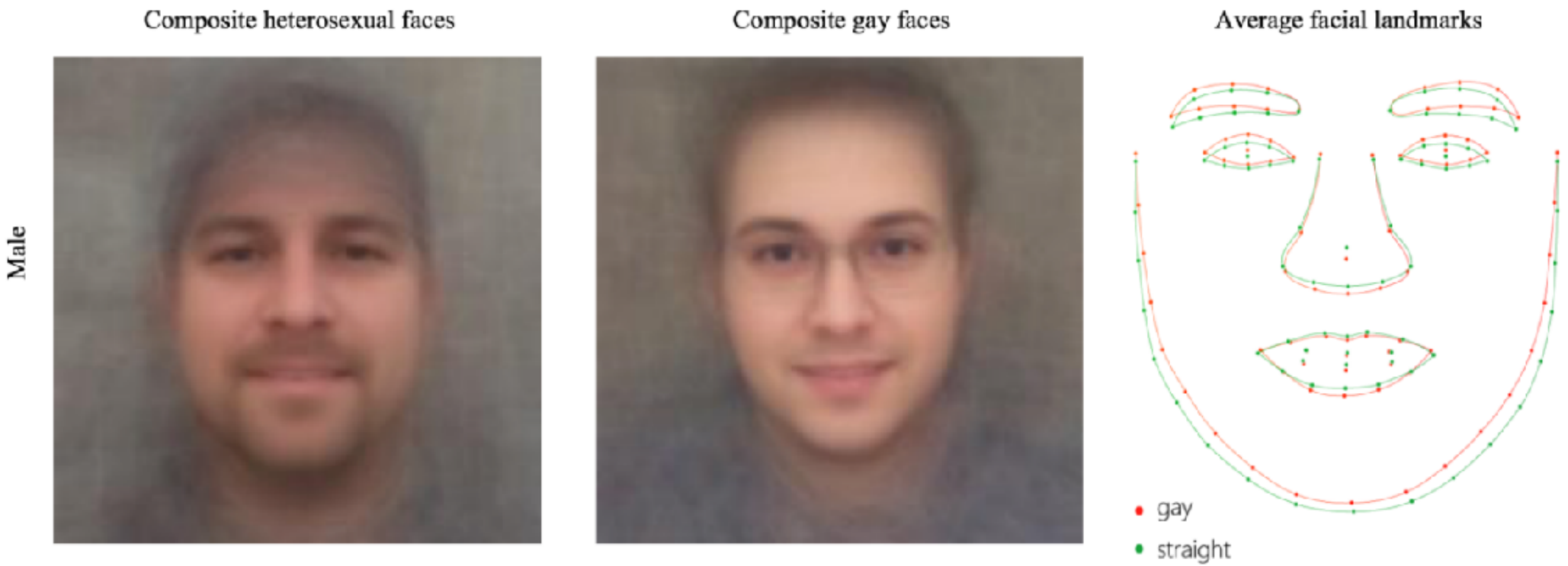
When can you say that consent was informed?

> Posts were determined to be positive or negative if they contained at least one positive or negative word, as defined by Linguistic Inquiry and Word Count software (LIWC2007) (9) word counting system, which correlates with self-reported and physiological measures of well-being, and has been used in prior research on emotional expression (7, 8, 10). LIWC was adapted to run on the Hadoop Map/Reduce system (11) and in the News Feed filtering system, such that no text was seen by the researchers. As such, it was consistent with Facebook's Data Use Policy, to which all users agree prior to creating an account on Facebook, constituting informed consent for this research. Both experiments had a control condition, in which a similar

- "Blanket consent" without explicit purpose is not considered informed
- The GDPR regulates this and similar clauses are not legal in the EU any more

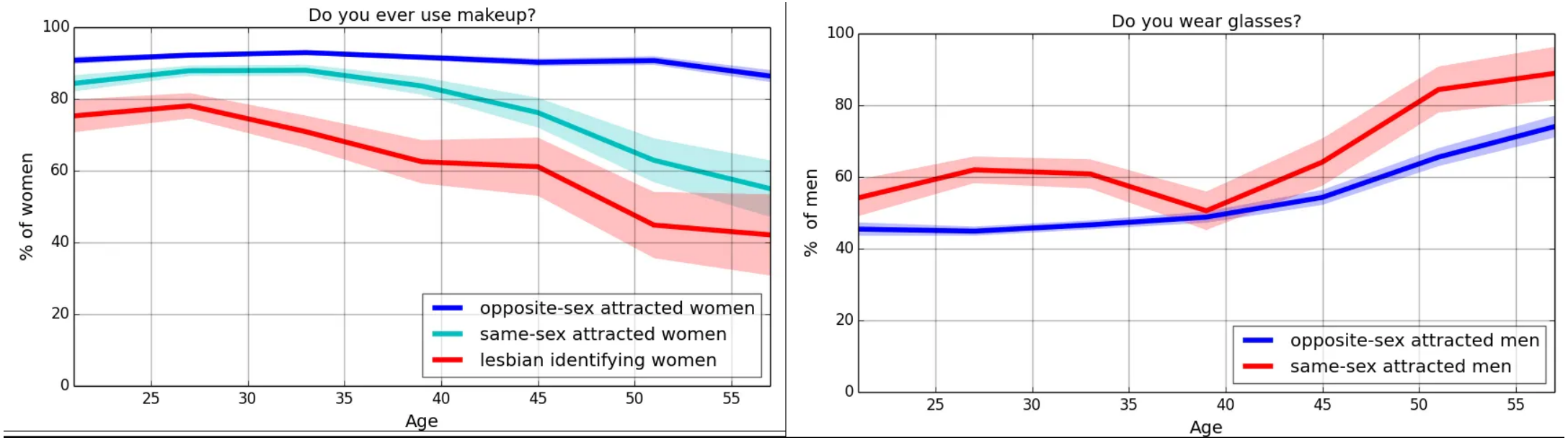
Experimental evidence of massive-scale emotional contagion through social networks. Kramer et al, PNAS (2014)

Use expectations: when public is not enough



Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images. Kosinski and Wang (2018)

Downstream consequences: coding stereotypes



Do algorithms reveal sexual orientation or just expose our stereotypes? Blaise Aguera y Arcas, Alexander Todorov and Margaret Mitchell (2018)

Downstream consequences: coding stereotypes

Do algorithms reveal sexual orientation or just expose our stereotypes? Blaise Aguera y Arcas, Alexander Todorov and Margaret Mitchell (2018)

A note on "ethics approval"

This study was peer reviewed and published in the *Journal of Personality and Social Psychology*, the leading academic journal in psychology. In addition, before it was sent for a formal peer review, the manuscript was reviewed by over a dozen experts in the fields of sexuality, psychology, and artificial intelligence. The research has been approved by Stanford's Internal Review Board.

- IRB only provide statements that a research project design complies with the regulations of the institution regarding human subjects research
- IRB are not "ethics approvals" or "ethics certificates", but they are proof that researchers have followed certain due process to consider ethics issues
- IRB focus on direct harm in experiments (e.g. health adverse effects) and do not consider other downstream social or technical risks

Authors' note: Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. M. Kosinski & Y. Wang (last update 2022)

Downstream consequences: technology misuse

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

Responses to Critiques on Machine Learning of Criminality Perceptions (Addendum of arXiv:1611.04135), Wu and Zhang (2017)

Terms of Service vs Right to Information

HOME / INDUCTEES / AARON SWARTZ

INTERNET HALL of FAME INNOVATOR

Aaron Swartz

Posthumous Recipient

Aaron Swartz was a computer programming prodigy and activist who played an instrumental role in the campaign for a free and open Internet and used technology to fight social, corporate and political injustices.

Federal Court Rules 'Big Data' Discrimination Studies Do Not Violate Federal Anti-Hacking Law. Americal Civil Liberties Union (2020)

Outline

1. Ethics in social media research
2. *Privacy issues of digital traces*
3. Discussing a recent case

Personal data and the GDPR

'Personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

- Examples of identifiers: IP addresses, full names, phone numbers, addresses
- Requirement to get data subject consent for processing and for **sharing with other legal entities** (e.g. through an API)
- The GDPR has numerous exceptions that apply to social media data analysis:
  - When data are manifestly made public by the data subject (Art 9.2e)
  - When processing does not require identification (Art 11)
  - For purposes in the public interest, scientific or historical research (Art 5.1b and Art 89)

What is personally identifiable information (PII)?

The New York Times

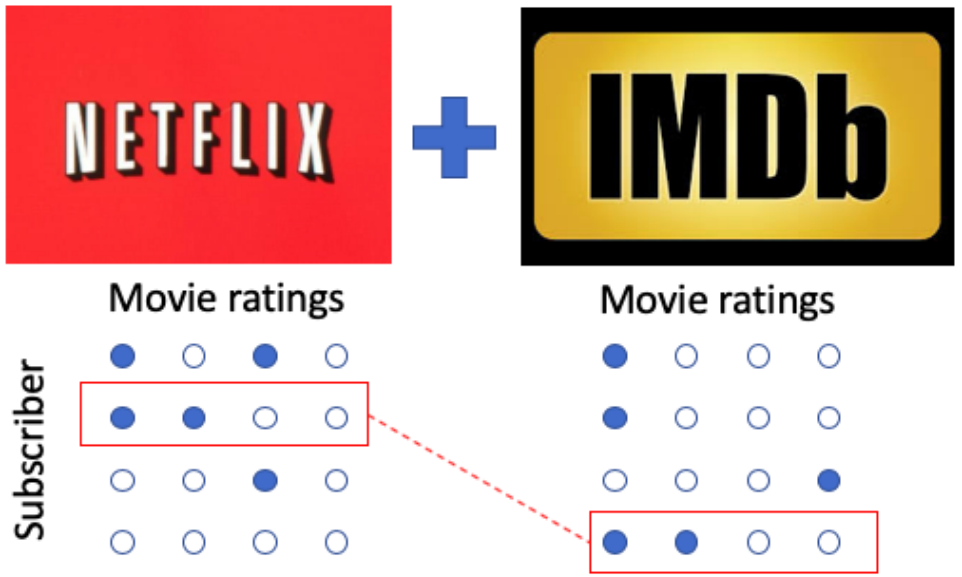
A Face Is Exposed for AOL Searcher No. 4417749

By Michael Barbaro and Tom Zeller Jr.

Aug. 9, 2006

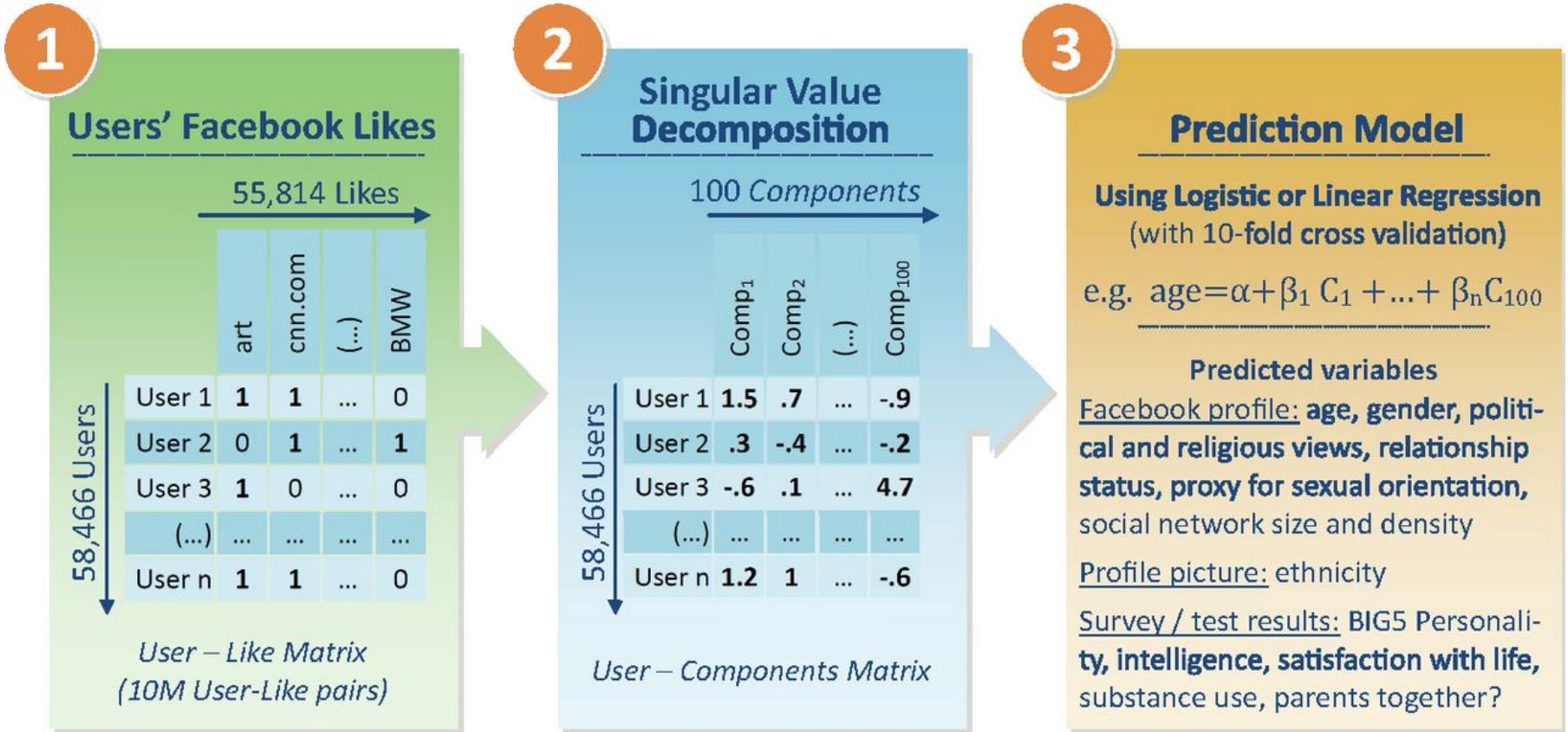
Buried in a list of 20 million Web search queries collected by AOL and recently released on the Internet is user No. 4417749. The number was assigned by the company to protect the searcher's anonymity, but it was not much of a shield.

NYT article on AOL data deanonymization



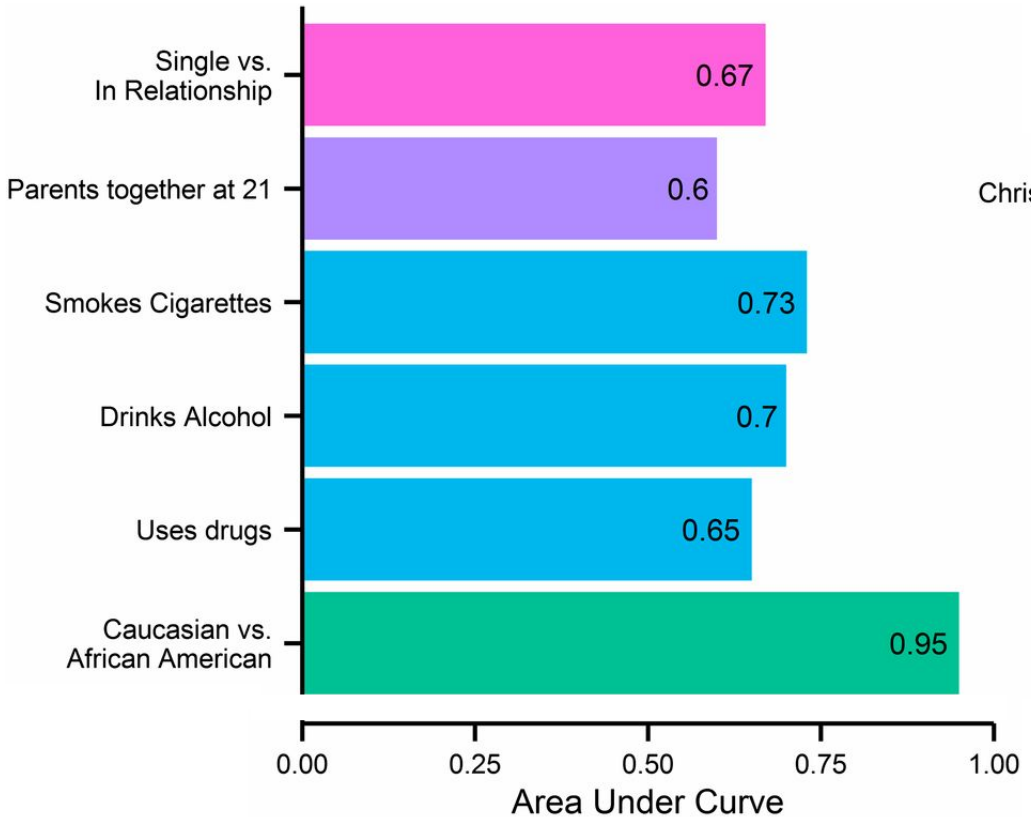
Artificial Intelligence magazine on the Netflix challenge deanonymization

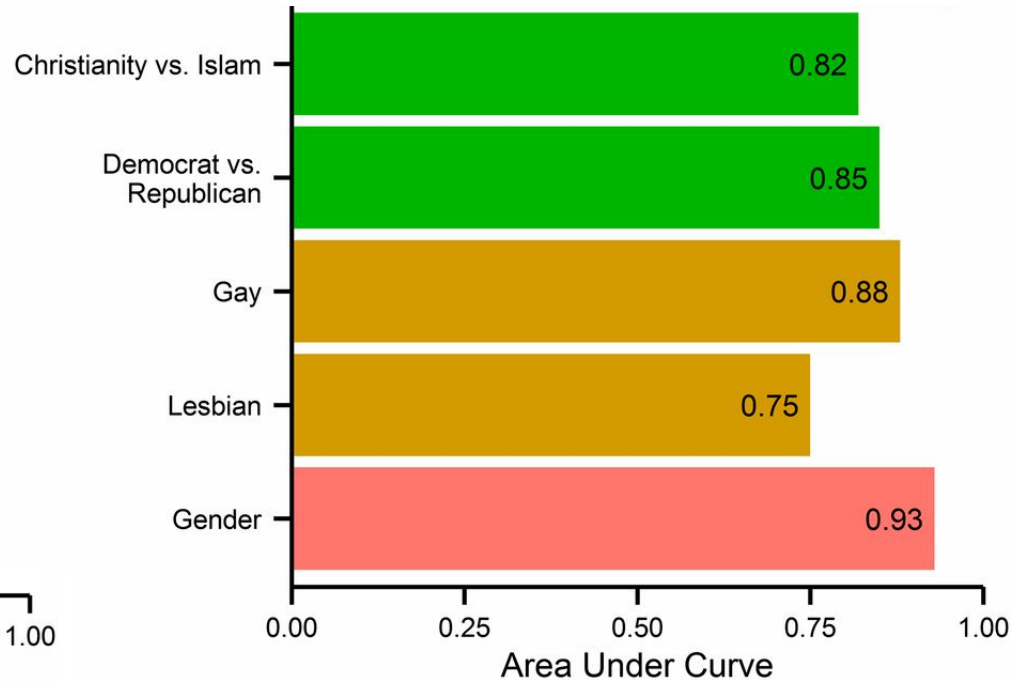
Facebook likes predict private attributes



Private traits and attributes are predictable from digital records of human behavior (Kosinski et al, 2013)

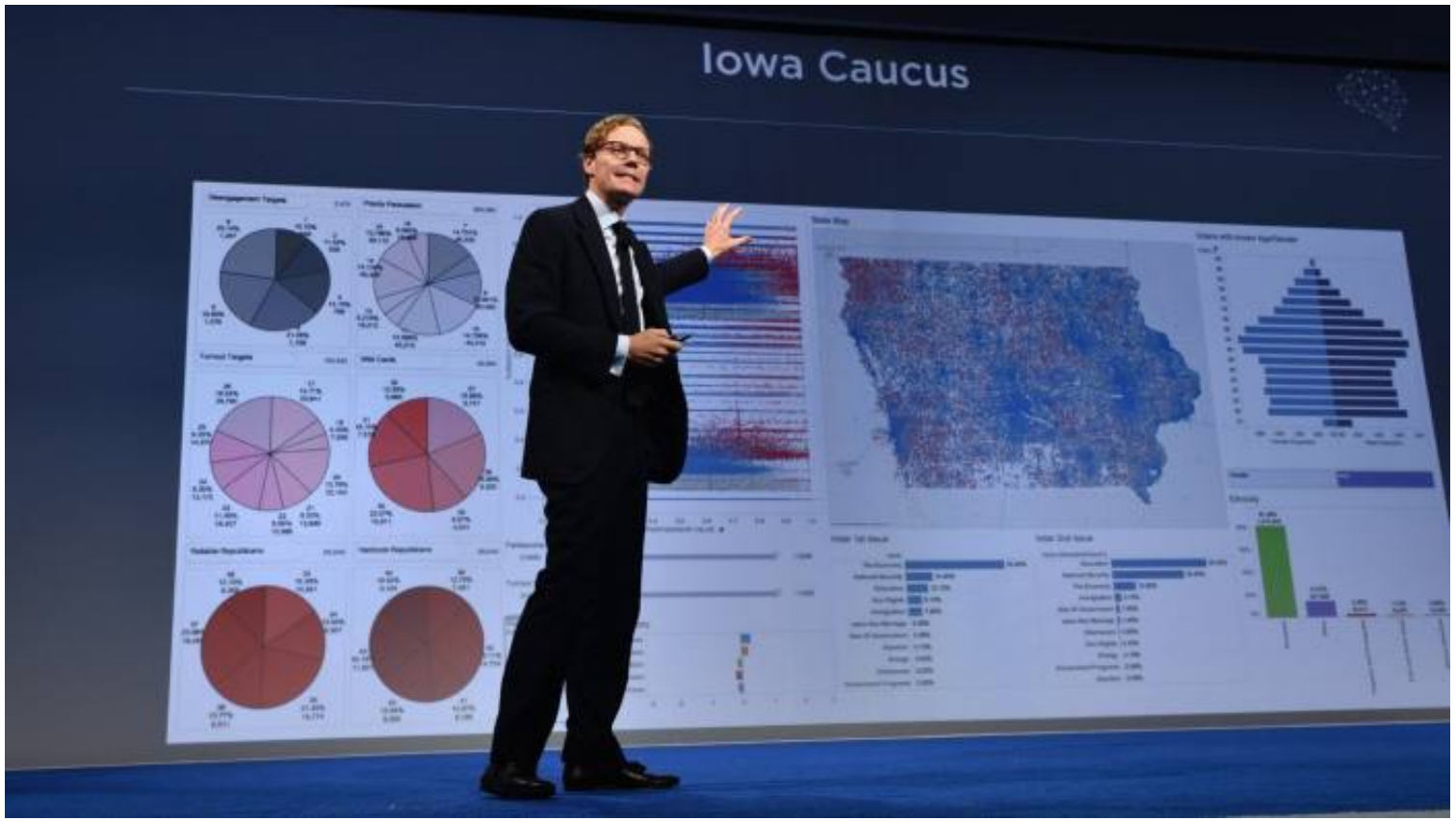
Facebook likes predict private attributes





Private traits and attributes are predictable from digital records of human behavior (Kosinski et al, 2013)

The Cambridge Analytica Scandal



https://www.ft.com/content/e325e3d0-2c3b-11e8-a34a-7e7563b0b0f4

Privacy Risks in Online Platforms

Informational self-determination: The protection of the individual against unlimited collection, storage, use and disclosure of their personal data, the right for the individual to determine the use of their personal data (habeas data).

One approach: Individualized solutions to privacy risks:

> ”Providing users with transparency and control over their information, leading to an individually controlled balance between the promises and perils of the Digital Age.” *(Kosinski et al, 2013)*

Can we guarantee individual control over online privacy?

The Problem of Shadow Profiles

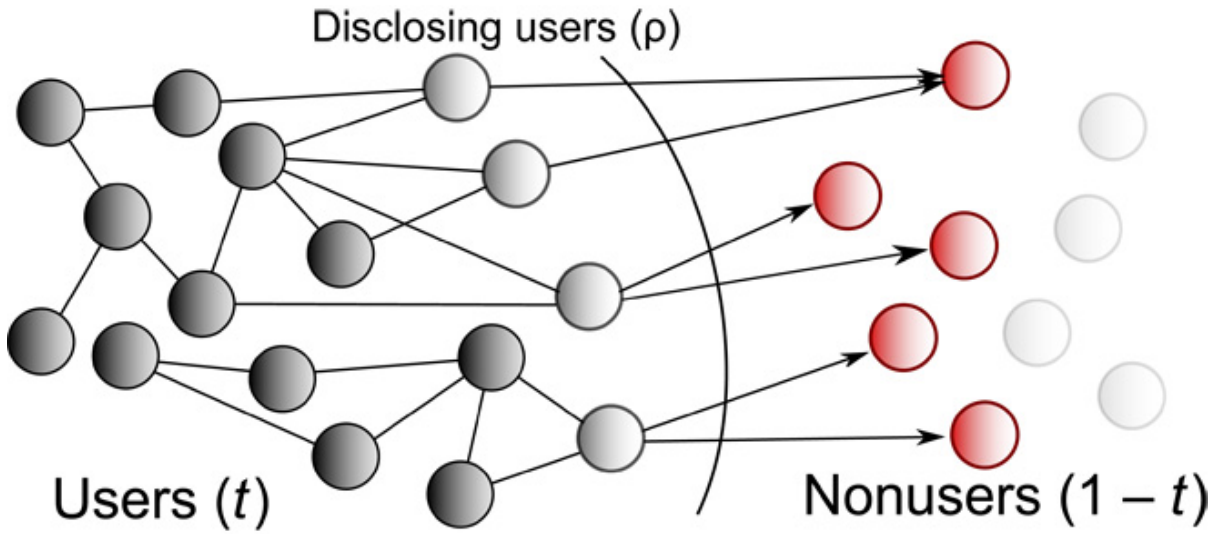
Facebook Shadow Profile:
A file that Facebook keeps on you containing data it pulls from looking at the information that your friends voluntarily provide. *(Digital Trends, 2013)*

- 2011: Europe-vs-Facebook files complaint on shadow profiles
- 2013: Bug reveals private information of 6 million users
- Some users notice their shadow profiles with mobile numbers
- Apr 2014: 1B users of Facebook for mobile share contact lists
- Aug 2016: Whatsapp and Facebook share information

Not a problem unique to Facebook: also in Twitter, Google, Amazon...

Auditing Shadow Profiles

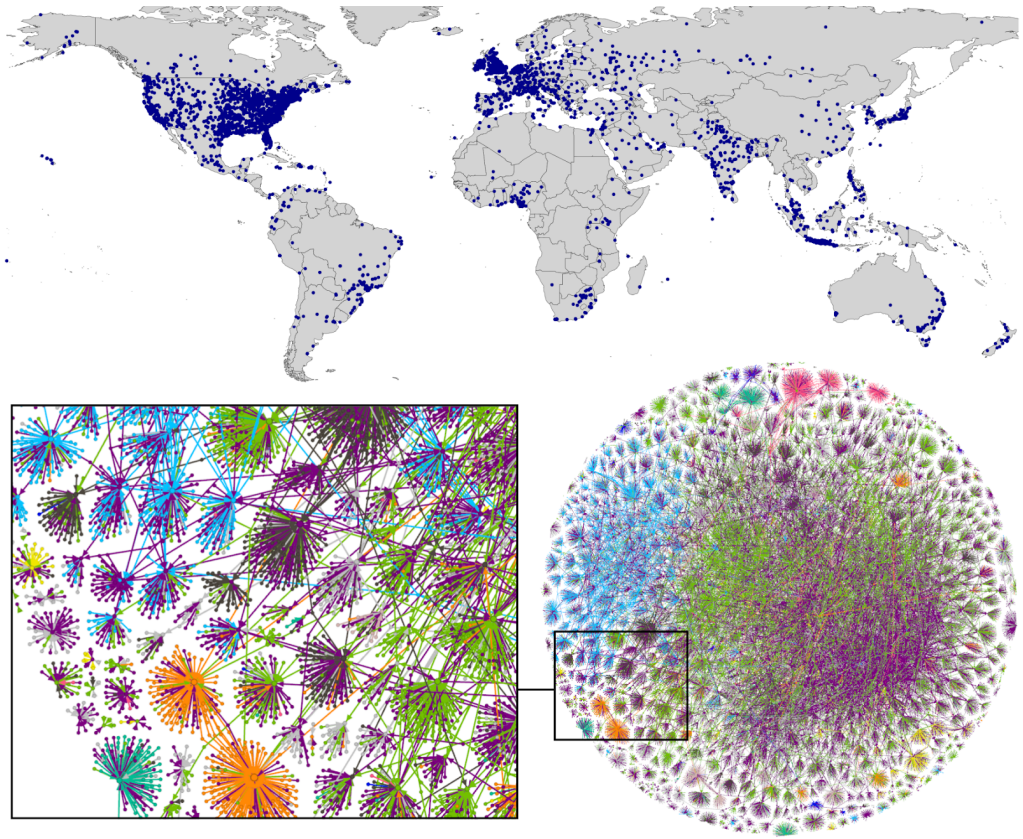
Shadow profile hypothesis: The data shared by the users of an online platform predicts of personal information of non-users



Approach: historical audit to evaluate how social networking sites could have predicted information of individuals who were not users yet

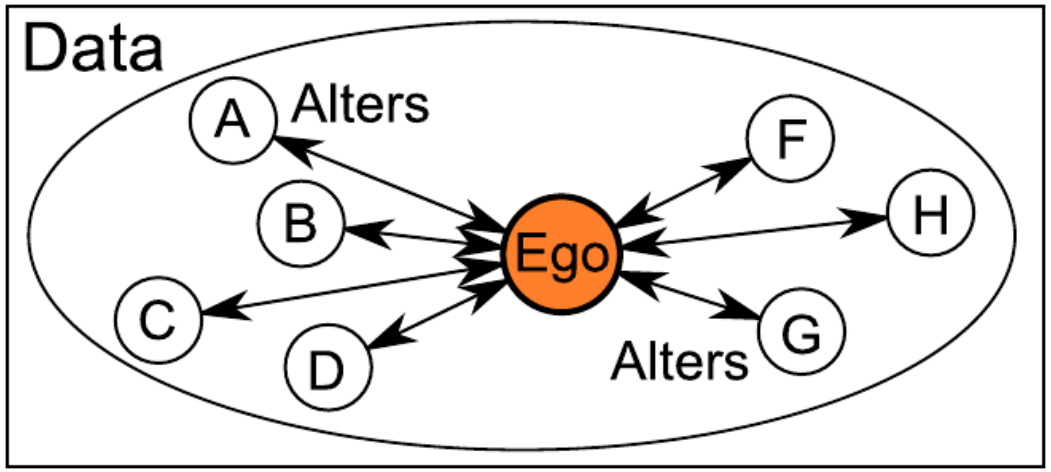
One Plus One Makes Three (for Social Networks). Emöke-Ágnes Horvát, Michael Hanselmann, Fred Hamprecht, Katharina Zweig. Plos ONE (2012)

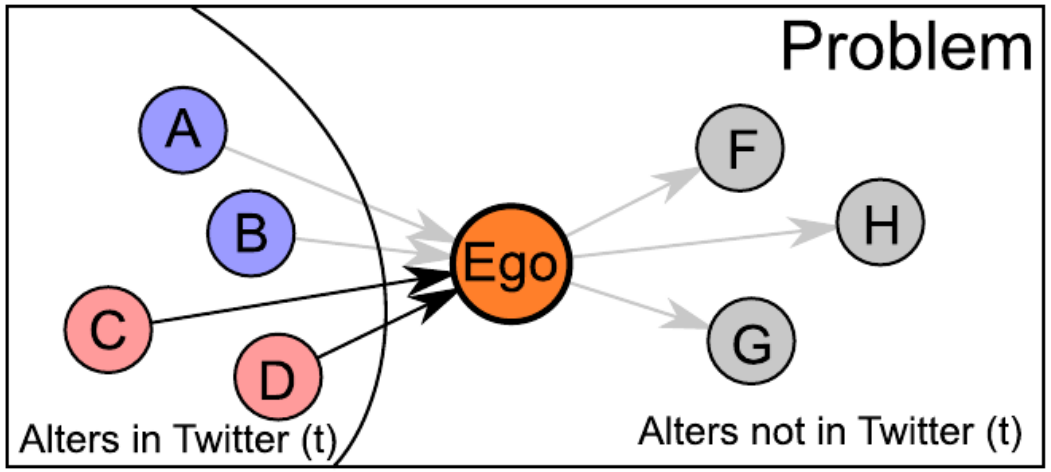
Shadow profiles case: Location in Twitter



- Random sample of 1,017 users
  - excl. bots, mass media
- Ego network of reply links
  - 68,447 alters
- Timelines: 157M tweets
- Location from profile text + Google Maps API
- Tweet metadata to identify users sharing contact lists

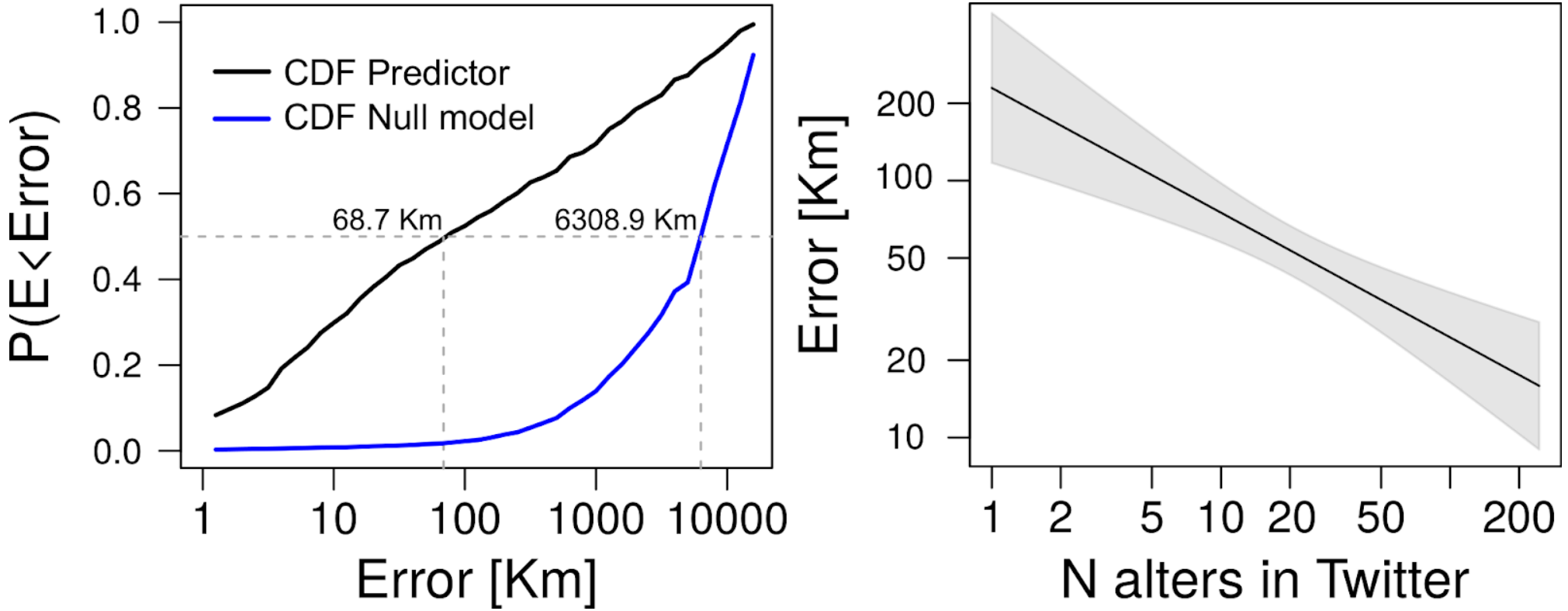
Twitter shadow profile test





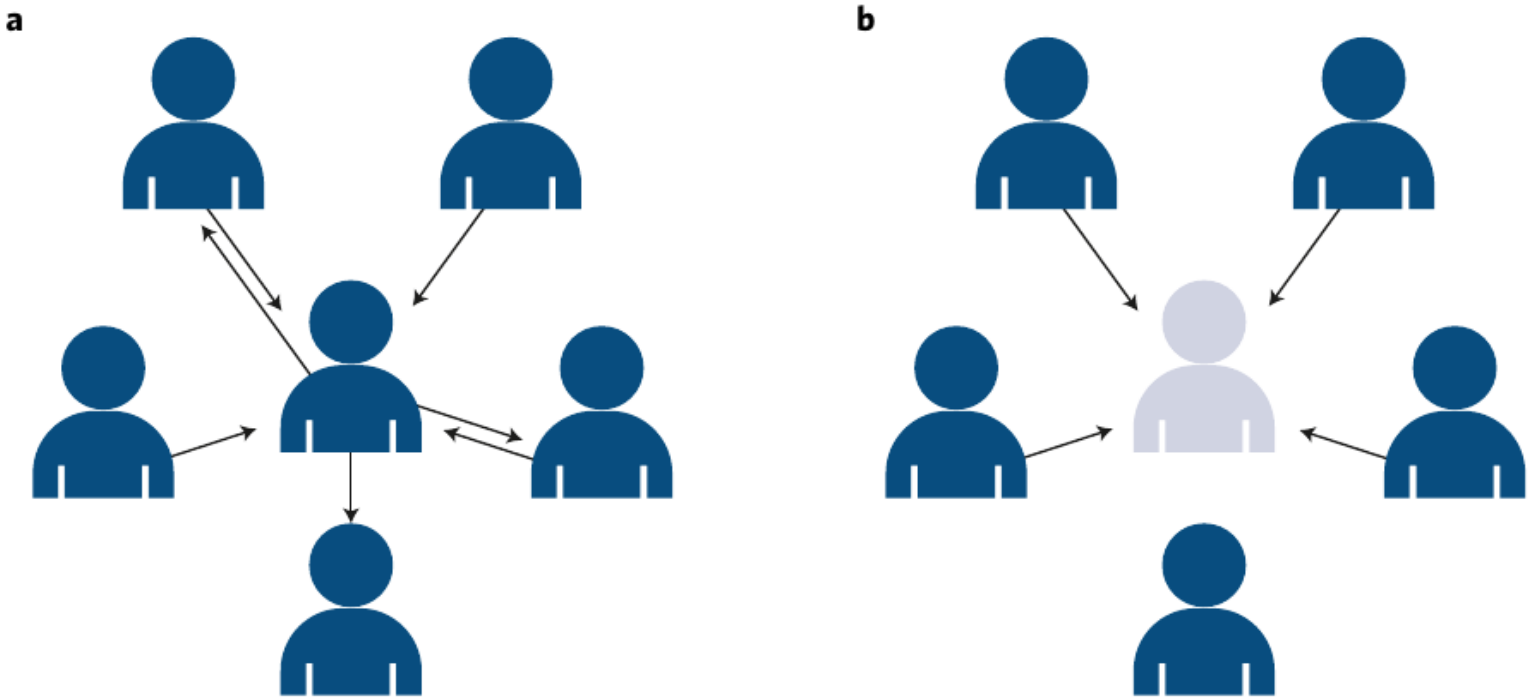
- Data represents a bidirectional social network
- Joined date of users shows network growth
- Tweet metadata contains posting app: disclosing users are only those who post at least once from Twitter mobile apps (shared contact lists)
- Predicting location of users who didn't join yet based on disclosing friends already on Twitter

Location prediction and heterogeneity



- Location prediction of non-users greatly outperforms null model
- Individuals with more alters sharing contact lists in Twitter have lower error

Shadow profiles after leaving a social network



Privacy beyond the individual. David Garcia. Nature Human Behavior (2019)

Information flow reveals prediction limits in online social activity. James Bagrow, Xipei Liu & Lewis Mitchell. Nature Human Behavior (2019)

Complex privacy: Online privacy as a collective phenomenon

- The decision of individuals to share data is mediated by the decisions of others
- Privacy externalities of the contract between a user and a platform
- Complex problems, collective solutions: International data cooperatives

Online privacy as a collective phenomenon. E. Sarigol, D. Garcia, F. Schweitzer. Second ACM Conference on Online Social Networks (COSN) (2014)
Leaking privacy and shadow profiles in online social networks. D. Garcia, Science Advances, 3 (8) e1701172 (2017)
Collective Aspects of Privacy in the Twitter Social Network. D. Garcia, M. Goel, A. Agrawal, P. Kumaraguru, EPJ Data Science, 7(3) (2018)
Privacy beyond the individual. D. Garcia. Nature Human Behavior (2019)

Summary: Ethics and privacy in SMDA

- Weigh benefit and risk before deciding to start research
- Public data is not automatically fair game: data are people
- Privacy is not binary: Is my research in the interest of the data subjects?
- Ethical data sharing: do share research data, but carefully
- Informed consent is not ticking a box. Do debriefing
- Consider downstream consequences: how can my science be misused?
- Society has a right to information: critically consider Terms of Service

Ten simple rules for responsible big data research. Zook et al. (2017)

Outline

1. Ethics in social media research
2. Privacy issues of digital traces
3. *Discussing a recent case*

A recent case: Koko and GPT-3

Rob Morris
@RobertRMorris

We provided mental health support to about 4,000 people — using GPT-3. Here's what happened 👇

8:50 PM · Jan 6, 2023 · **8.7M** Views

1,232 Retweets **3,088** Quotes **6,076** Likes **3,437** Bookmarks

Rob Morris @RobertRMorris · Jan 6

To run the experiment, we used @koko — a nonprofit that offers peer support to millions of people...

21 49 287 510.4K

A recent case: Koko and GPT-3

Rob Morris @RobertRMorris · Jan 6

On Koko, people can ask for help, or help others. What happens if GPT-3 helps as well?

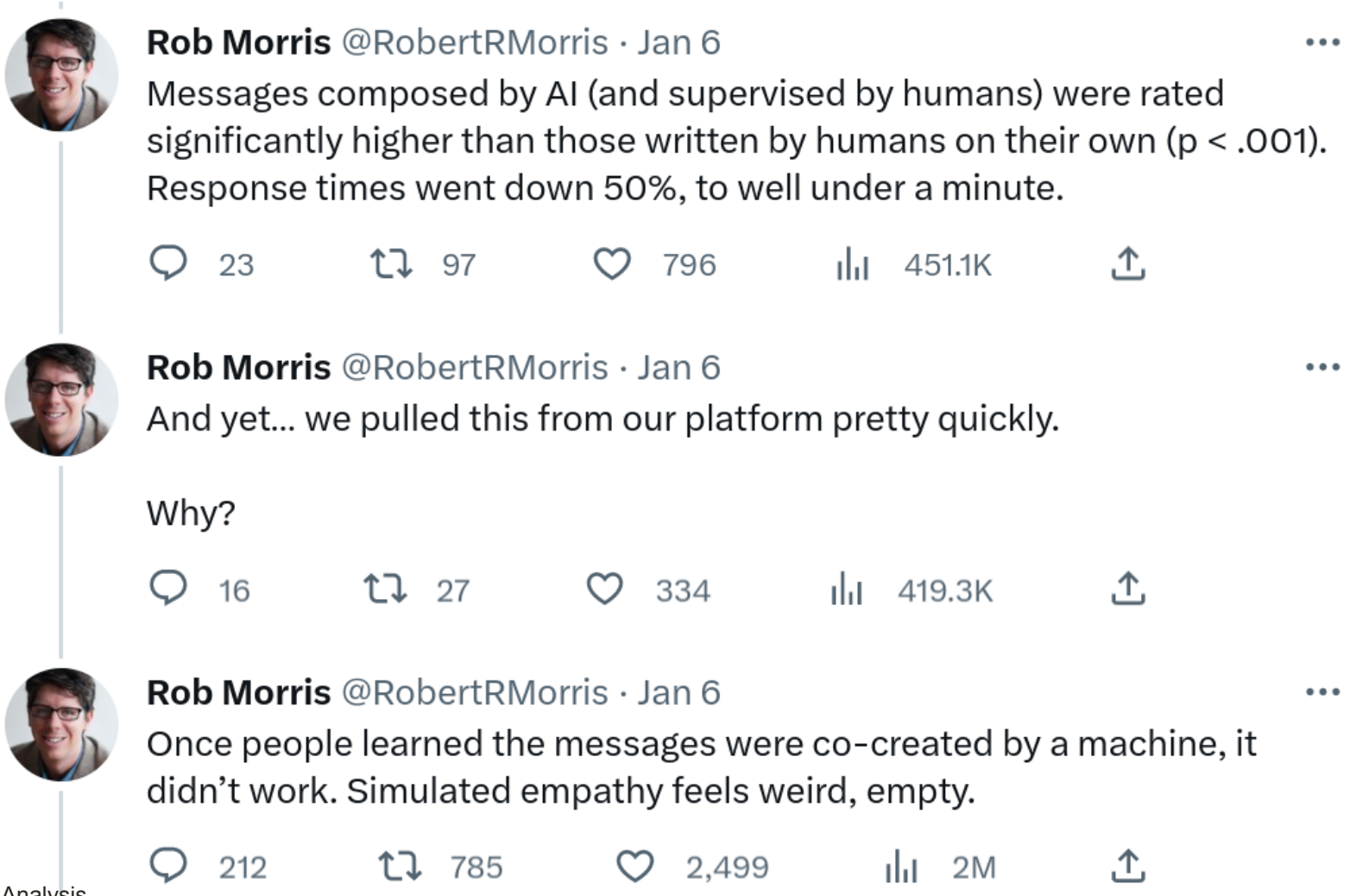
6 17 214 499.7K

Rob Morris @RobertRMorris · Jan 6

We used a ‘co-pilot’ approach, with humans supervising the AI as needed. We did this on about 30,000 messages...

8 27 278 503.3K

A recent case: Koko and GPT-3



Summary

- **Ethics in social media data analysis**
  - No harm principle
  - Cases of previous ethical issues in digital trace data research
  - Recommendations for Big Data research

- **Privacy in the digital society**
  - Digital traces predict private attributes
  - Informational self-determination
  - Online privacy is complex: the shadow profiles problem

- **A recent case**
  - Reasons to do it? benefit for participants and society?
  - Risks to participants? Harm or also privacy?
  - Other downstream risks/benefits