Sex, Lies, and Data

Shea Swauger

University of Colorado - Denver

Follow this and additional works at: https://escholarship.umassmed.edu/escience_symposium

Part of the Scholarly Communication Commons

This work is licensed under a Creative Commons Attribution 4.0 License.

Repository Citation


Creative Commons License

This work is licensed under a Creative Commons Attribution 4.0 License.

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in University of Massachusetts and New England Area Librarian e-Science Symposium by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
sex, 

lies, 

Facebook

and data
Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock

PNAS is publishing an Editorial Expression of Concern regarding the following article: "Experimental evidence of massive-scale emotional contagion through social networks," by Adam D. Kramer, Jamie E. Guillory, and Jeffrey T. Hancock, which appeared in issue 26, June 24, 2014, of Proc Natl Acad Sci U S A (111:8787-8790; first published June 2, 2014; 10.1073/pnas.1320081111). This paper represents an important and emerging area of social science research that can be approached with sensitivity and with vigilance regarding potential ethical issues. Questions have been raised about the principles of informed consent and opportunity to opt out in connection to the research described in this paper. The authors noted in their paper, "(The work) was conducted with Facebook’s Data Use Policy, to which all users agree prior to creating an account on Facebook, containing informed consent for this research." When the authors prepared their paper for publication in PNAS, they stated that "Because the experiment was conducted by Facebook, Inc. for internal purposes, the Cornell University IRB [Institutional Review Board] determined that the project did not fall under Cornell’s Human Research Protections Program. This statement has been corrected.

Obtaining informed consent and allowing participants to opt out are best practices in most instances under the US Department of Health and Human Services Policy for the Protection of Human Research Subjects (the “Common Rule”). Adherence to the Common Rule in PNAS papers, but as a private company Facebook was not under any obligation to conform to the provisions of the Common Rule when they collected the data used by the authors, and the Common Rule does not preclude their use of the data. Based on the information available to the authors, PNAS editors deemed it appropriate to publish this paper. It is nevertheless a matter of concern that the collection of the data by Facebook may involve an imbalance of interest in an agreement that is not consistent with the principles of obtaining informed consent and allowing participants to opt out.

SOBIA J. VIERTER
Editor-in-Chief
OkCupid's data
The OKCupid dataset: A very large public dataset of dating site users

Emil O. W. Kirkegaard 1
Julius D. Bjerrekrer 1

Abstract

A very large dataset (N=41,371,263 variables) from the dating site OKCupid is presented and made publicly available for use by others. As an example of the analyses one can do with the dataset, a cognitive ability test is constructed from 14 suitable items. To validate the dataset and the test, the relationship of cognitive ability to religious beliefs and political interest/position is examined. Cognitive ability is found to be negatively related to all measures of religious beliefs (latent correlations -0.26 to -0.30), and found to be positively related to all measures of political interest and position (latent correlations 0.19 to 0.22). To further validate the dataset, we examined the relationship between Zodiac sign and every other variable. We found very scant evidence of any influence (the distribution of p-values from chi square tests was flat).

Limitations of the data are discussed.

Keywords: open data, big data, OKCupid, dating site, cognitive ability, IQ, intelligence, scale construction, religiosity, politics, astrology, Zodiac sign

1 Introduction

Despite many years of advocacy of preprints, it is still uncommon for social scientists to publicly share their datasets and even sharing data on request is rare (Kranzusch & Stenholm, 2012; Savage & Vickers, 2009; Tenopir et al., 2011). Wicherts et al. (2011)5. Woeser, there is no obvious reason (which indicates that those who refuse to share data upon request make more statistical errors than those who share data (Wicherts et al., 2013). This is doubly problematic because the mistakes cannot be corrected by other researchers. Furthermore, a given dataset may have many uses not all of which are known to those who collected it. Even if they are known, the collectors may not be interested in them (or even interested in hiding the results (Duarte et al., 2015)), or they may simply not have enough time. Whatever it is, if the analyses are not done by the collectors and the data are unsuitable to anyone else, the data is not used to its full extent. Big data is usually funded by the public; this wastes an incredible amount of public funds.

The lack of data sharing probably slows down the progress of science immensely because other researchers would use the data if they could. Supporting this, it can be observed that known public datasets enjoy widespread use. For instance, Project TALENT is a very large (N=440,000) dataset of American high school students collected in 1960. The dataset is still used by researchers examining questions that were never considered of when it was collected. According to Google Scholar (https://scholar.google.com), there were 296, 148 and 116 papers published in 2013, 2014 and 2015 that had “project talent” in their title, indicating that it can still be used despite being about 56 years old. For instance, (Dunkel, 2014) used the data to examine the cognitive ability of children by the language spoken by their parents, while Mayer et al. (2014) examined the linearity of relationships between cognitive ability and personality traits, and Damon & Roberts (2017) examined whether birth order showed relationships to cognitive ability and personality. The NLST, CNLSY and NSLS95 are nationally representative (using probability sampling) longitudinal

1 The wording of the surveys over time is perhaps an effect of the way Google Scholar finds articles. Many newer articles have probably not been fully indexed yet, hence lowering the numbers.
2 NSLS stands for National Longitudinal Survey of Youth, the

1

You Are What You Eat... Listen to, Watch, and Read

Mason Breeten
Georgia Institute of Technology
masonbreeten@gmail.com

Abstract

This article describes a data driven method for deriving the relationship between personality and media preferences. A quantifiable representation of such a relationship can be leveraged for use in recommendation systems and ameliorate the "cold start" problem. Here, the data is comprised of an original collection of 1,166 Okcupid dating profiles. Of these profiles, 800 are labeled with one of 16 possible Myers-Briggs Type Indicators (MBTI). A personality specific topic model describing a person's favorite books, movies, shows, music, and food was generated using latent Dirichlet allocation (LDA). There were several significant findings, for example, intuitive thinking types preferred sci-fi/fantasy entertainment, extraversion correlated positively with upbeat dance music, and jazz, folk, and international cuisine correlated positively with those characteristics by openness to experience. Many other correlations confirmed previous findings describing the relationship among personality, writing style, and personal preferences. (For complete word/personality type associations see the Appendix.)

1 Introduction

Recommendation systems have become increasingly popular over the last several years. These systems are catered towards a multitude of mediums including music (Pandora, LastFM), movies (Netflix), literature (Amazon), food and restaurants (Yelp, UrbanSpoon), and people (Okcupid). Many of these systems rely on user history within the specific domain combined with collaborative filtering techniques to recommend new options. For example, Netflix uses a user's previous views to create suggestions. Pandora uses a thumbs up and down rating system from users to inform their playlists. Other systems such as Amazon suggest content based off of purchases from similar buyers.

These methods have proven to be relatively successful, however, we ponder if additional information about users not directly related to the content can be useful for predicting his or her preferences. Additionally, it is difficult for current systems to make recommendations to first time users when zero previous data is available to inform the system of what a person might buy, watch, or listen to. Netflix may recommend movies and television shows using what is currently popular, which seems like a logical approach, but if a user is outside of the norm she may be turned off to the suggestions and first impressions can be tantamount to an application's success. Ideally the system should leverage some prior knowledge about a user before making suggestions to ameliorate this “cold start” problem. In this work we examine traits that are descriptive of a person and correlate with a person's preferences. Specifically, we examine the relationship between an individual's personality and her preferences in music, movies, TV shows, food, and literature.

2 Objectives

There is a plethora of work examining the relationship between personality and media preferences. Personality traits such as psychism and neuroticism have been linked to preferences for horror/violent films

1

1

2
Do Personal Ethics Influence Corporate Ethics?*

John M. Griffin  Samuel Kruger  Gonzalo Maturana

First draft: November 18, 2015
This draft: August 18, 2016

ABSTRACT

We introduce a new measure of personal ethics in the form of marital cheating to examine the relationship between personal ethics and corporate misconduct. Firms with CEOs and CFOs who use marital infidelity websites are more likely to engage in two forms of corporate misconduct. The relationship is not explained by a wide range of regional, firm, and executive characteristics or by the infidelity website usage of other executives. Additionally, white-collar SEC defendants also have elevated levels of infidelity website usage. Our findings suggest that personal and professional ethics are not as distinct as some believe.

*We are grateful to Keith Brown, Jonathan Cohn, Andres Donangelo, Cesare Fracassi, Paul Gendreau, Stefan Lewallen, Issac Liskovich, Robert Parchio, Alessandro Previtero, Laura Stacks, and Luigi Zingales, as well as seminar participants at the 2016 Wake Forest CSR Conference, the 2016 Center for Accounting Research and Education Conference, the Securities and Exchange Commission, and the University of Texas at Austin for helpful comments. We thank Melissa Hall, Ziqian Ju, Jangwoo Lee, Marc LeBetecke, Felix Guzman, and Kashi Patel for research assistance. We are grateful for research support from the McCombs Research Excellence Fund and Integra FEG LLC for providing LexisNexis access. Supplementary results can be found in an Internet Appendix at the authors’ websites.

Griffin and Kruger are at the McCombs School of Business at the University of Texas at Austin. Maturana is at the Galvani School of Business at Emory University. Griffin: John.Griffin@utexas.edu; Kruger: Sam.Kruger@mccombs.utexas.edu; Maturana: Gonzalo.Maturana@emory.edu

Infidelity and the Internet: The Geography of Ashley Madison Usership in the United States

Michael L. Chohaney, Kimberly A. Panozzo

First published: 9 November 2016  Full publication history
DOI: 10.1111/greve.12235  View entire citation
Cited by: 0 articles  Citation tools

The authors would like to thank Dr. David Nemer of the University of Toledo for his input and support, as well as the constructive improvements to the study recommended by the editor, Dr. David Kaplan, and the three anonymous reviewers.

[Correction added on February 6, 2017, after initial online publication: The email address of Kimberly A. Panozzo was changed from “Kpanozzo@ricketts.utoledo.edu” to “Kimberly.Panozzo@ricketts.utoledo.edu.”]

Abstract

Ashley Madison (com) has earned several million dollars facilitating extramarital affairs online; however, the market determinants of online infidelity matchmaking have not been researched. The now-infamous customer data breach in 2015 provided a unique opportunity to analyze a large population of individuals (N=702,309) who paid to engage in extramarital affairs using Ashley Madison. Aggregating this sensitive data into spatial units, we measured the relationship between several theorized market determinants and Ashley Madison subscription and spending rates in major United States markets. We found income is the leading market determinant for Internet-facilitated infidelity, indicating the service behaves as a luxury good; further, several characteristics related to infidelity at the individual-level were also significant, including the negative relationship between religiosity and infidelity. Strong regression model performance suggests these results are robust insights into the market for online infidelity matchmaking.

Introduction

In July 2015, an anonymous hacker group self-described as “The Impact Team” accessed the user account database of the online dating website Ashley Madison.com and stole internal documents and private customer records. The hacker group demanded the website to cease operation on ethical grounds and threatened to publicly release the entire database of over 30 million user account records if it failed to do so. Ashley Madison did not meet the demand and the data was publicly released via file-sharing networks. The data breach affected 37 million user accounts, 15 million in the United States. The data included names, email addresses, passwords, phone numbers, and personal details. Of the 37 million accounts, over 20 million were in the United States. The data breach affected millions of individuals, and the consequences of the breach have been extensive. The breach led to a loss of trust and a decrease in user confidence in Ashley Madison. The website was forced to shut down and rebrand its service under a different name. The breach also raised questions about the security of personal data and the ethics of online matchmaking services.
STRUCTURES for PROTECTION
INSTITUTIONAL REVIEW BOARD
TUSKEGEE SYPHILIS EXPERIMENTS
Syphilis Victims in U.S. Study Went Untreated for 40 Years

By JEAN HELLER
The Associated Press

WASHINGTON, July 25—For 40 years the United States Public Health Service has conducted a study in which human beings with syphilis, who were induced to serve as guinea pigs, have gone without medical treatment for the disease and a few have died of its late effects, even though an effective therapy was eventually discovered.

The study was conducted to determine from autopsies what the disease does to the human body.

Officials of the health service who initiated the experiment have long since retired. Current officials, who say they have serious doubts about the morality of the study, also say that it is too late to treat the syphilis in any surviving participants.

Doctors in the service say they are now rendering whatever other medical services they can give to the survivors while the study of the disease's effects continues.

Dr. Merlin K. DuVal, Assistant Secretary of Health, Education and Welfare for Health and Scientific Affairs, expressed shock on learning of the study. He said that he was making an immediate investigation.

The experiment, called the Tuskegee Study, began in 1932 with about 600 black men.
The Belmont Report
Ethical Principles and Guidelines for the Protection of Human Subjects of Research

The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research

DHEW Publication No. (OS) 78-0012
“A chilling study of the way power corrupts.”
NEW ROADS AHEAD

FURIOUS 16

APRIL 14

IN THEATERS AND IMAX
ALGORITHMIC DISCRIMINATION
SOMETIMES,
THE WRONG THING TO DO
MAY BE ALLOWED.
SOMETIMES,

THE RIGHT THING TO DO MAY BE AGAINST THE RULES.
POLICY VIGILANTISM
SOME GROUND RULES
1. SILENCE DOES NOT EQUAL CONSENT.
2. ONLINE AVAILABILITY DOES NOT MAKE SOMETHING PUBLIC.
3.
TOOLS ARE NOT NEUTRAL.
4. DATA ARE NOT NEUTRAL.
ETHICS ARE COMPLEX & INTERSECTIONAL.
6.

OUR ONTOLOGIES & EPISTEMOLOGIES ARE SUSPECT.
ONTOMETRY
ONTOLOGY
values are built into all research.

(this includes yours)
so build something into all research that interrogates those values.
sex,
lies,
and data
articles and documents mentioned


image credits


image credits continued


further reading

• S Mannheimer, SWH Young, D Rossmann On the Ethics of Social Network Research in Libraries. Journal of Information, Communication, and Ethics in Society 14 (2)

• S Mannheimer, EA Hull Sharing selves: Developing an ethical framework for curating social media data. International Digital Curation Conference 2017

• Data and Society: https://datasociety.net/

• Council for Big Data, Ethics, and Society: http://bdes.datasociety.net/

• Literally anything by these people (seriously):
  Anna Lauren Hoffmann http://annaeveryday.com/
  danah boyd https://datasociety.net/people/boyd-danah/
  Safiya Umoja Noble https://gseis.ucla.edu/directory/safiya-umoja-noble/