INTERNATIONAL WORKSHOP ON

OB{FUS}CATION:

SCIENCE, TECHNOLOGY, AND THEORY

WORKSHOP REPORT

NYU | APRIL 7-8, 2017

OBFUSCATIONWORKSHOP.IO

INTERNATIONAL WORKSHOP ON OBFUSCATION: SCIENCE, TECHNOLOGY, AND THEORY

APRIL 7-8, 2017 | NYU

ORGANIZERS

Finn Brunton, New York University
Helen Nissenbaum, Cornell Tech & New York University

INTERNATIONAL PROGRAM AND ORGANIZING COMMITTEE

Paul Ashley, Anonyome Labs
Benoît Baudry, INRIA, France
Finn Brunton, New York University
Saumya Debray, University of Arizona
Cynthia Dwork, Harvard University
Rachel Greenstadt, Drexel University
Seda Gürses, Princeton University
Anna Lysyanskaya, Brown University
Helen Nissenbaum, Cornell Tech & New York University
Alexander Pretschner, Technische Universität München
Reza Shokri, Cornell Tech

RESEARCH ASSISTANT

Harris Kornstein, New York University

SPONSORS

NYU STEINHARDT

INFORMATION LAW INSTITUTE

NSF

FULL SCHEDULE & DETAILS

obfuscationworkshop.io

Workshop sponsored by NSF Award SES-1642553.

CONTENTS

4 Obfuscation Going Forward: A Research Agenda
Finn Brunton, New York University and Helen Nissenbaum, Cornell Tech and New York University

7 PrivacyVisor: Privacy Protection for Preventing Face Detection from Camera Images
Isao Echizen, National Institute of Informatics, Tokyo

10 Circumvention Through Obfuscation
Amir Houmansadr, University of Massachusetts Amherst

12 Political Rhetoric as Obfuscation and Finding Solutions with Neural Networks
Nicole Cote and Rob Hammond, New York University

14 Obfuscating Data to Prevent Discrimination
Sorelle Friedler, Haverford College

18 Using Ethically-Constrained Game Theory to Protect Our Privacy
Jeffrey Pawlick and Quanyan Zhu, New York University Tandon School of Engineering

22 Identity Obfuscation Through Fully Functional Avatars
Paul Ashley, Anonyome Labs

25 Go Rando: Resisting Emotional Surveillance with Noisy Feelings
Ben Grosser, University of Illinois at Urbana-Champaign

28 Hiding Data Flows with Covert Channels
Saumya Debray and Jon Stephens, University of Arizona

31 Software Diversification as an Obfuscation Technique
Nicolas Harrand and Benoit Baudry, Inria, France

35 Software (De-)Obfuscation: How Good Is It?
Alexander Pretschner, Technische Universität München, Germany

38 On Missing Datasets
Mimi Onuoha

41 Obfuscating 15M US Criminal Records and Mugshots for the Right to Remove Them
Paolo Cirio

45 HyperFace: Emerging Strategies for Obfuscating Computer Vision Algorithms
Adam Harvey

47 Place vs. Space: On the Future of Location Obfuscation
Seda Gürses

51 Obfuscation in Bitcoin: Techniques and Politics
Arvind Narayanan & Malte Möser, Princeton University

56 Obfuscation and the Threat of Centralized Distribution
Daniel C. Howe, School of Creative Media, City University of Hong Kong

OBFUSCATION GOING FORWARD: A RESEARCH AGENDA

FINN BRUNTON, New York University and
HELEN NISSENBAUM, Cornell Tech and New York University

Since we began planning the "International Workshop on Obfuscation: Science, Technology, and Theory" a year ago, there have been numerous shifts in the world's technological, political, and economic landscape, from a US election influenced by email leaks and algorithmically-promoted fake news stories, to the merger of some of the world's largest telecom and media companies into data-driven advertising behemoths, to data breaches of major entertainment companies, healthcare providers, and voting systems (to name but a few). We define *obfuscation* as the production of noise modeled on an existing signal in order to make data or information more ambiguous, uncertain, and difficult to exploit—an idea that is particularly salient in the era of big data technologies. In concert with other practices and tools, obfuscation offers a novel and unique means of evading data surveillance, building privacy-respecting platforms without sacrificing utility, and improving security (including through obfuscating code or hardware itself). However, while obfuscation has long been a methodology engaged by researchers and developers in certain subfields of computer science, engineering, and applied technologies, it has only recently been taken up and studied as a broader strategy or set of tactics by humanists, social scientists, policymakers, and artists.

Building off the 2014 Symposium on Obfuscation, as well as the myriad case studies we researched for our 2015 book *Obfuscation: A User's Guide for Privacy and Protest*, our intention for this workshop was to bring together a group of interdisciplinary scholars, industry researchers and practitioners, independent software producers, and privacy artists and activists to help shape this nascent field and seed the beginnings of a more holistic research community. Of course, obfuscation is not a singular solution, but instead operates across diverse scenarios, fields, and sociotechnical contexts, and can be wielded by and against many different actors—including both those with and without power. In most of the applications we consider, it serves as a means for individuals to evade scrutiny and create spheres of freedom and privacy, including freedom from being locked into an increasingly consolidated set of technologies and technology owners. Still, it has become clear that governments, corporations, or other institutional actors may also engage techniques of obfuscation, often for more nefarious ends.

Our goal, then, was not to attempt to nail down obfuscation, but rather to open up its myriad forms and applications to critical consideration. Following the shifting and provisional structure of obfuscation as a strategy—and in true

workshop format—we intended not merely to present typical academic papers, but to spark conversations across disciplines, methodologies, and applications, through a variety of formats that included prototypes of products, artistic interventions, and speculative proposals, as well as theoretical and empirical research from a range of fields. Moreover, as an interdisciplinary and multi-sector endeavor, our hope was to simultaneously engage technical issues, ethical and political concerns, and evaluations of the strengths and weaknesses of various use cases, and to consider the potential benefits and limitations of obfuscation overall.

To that end, we identified four key themes that not only helped structure the presentations themselves, but also emerged out of discussions throughout the weekend:

1. Threat Models: Put simply, understanding threat models is key to determining in which cases obfuscation is the solution and in which cases it is not. Throughout the workshop, it became clear that effective threat modeling includes understanding adversarial capacities and dependencies, levels of coordination, and questions of time and scale. For example, obfuscation may be an effective defense against lower-level adversaries (such as a jealous spouse or boss), while still being breakable by those with greater (or networked) computational, financial or legal, resources; similarly, what seems hidden today may be revealed tomorrow. Several speakers raised concerns about the dangers of under- or over-estimating the threat level, as well as perpetuating an "adversarial arms race."

2. Benchmarks and Metrics: Anyone who has created obfuscation tools has been asked, "But does it work?" Throughout the course of the workshop, several different models for assessing success were offered for different contexts, including measurements based on financial costs, computing power, time to de-obfuscate, political efficacy, and so on. There are many cases where identifying benchmarks can be socially or technically challenging, such as in determining training data for "real" or "fake" news. Moreover, it became clear that often those creating obfuscating systems have only a partial view of their adversaries (or of the effects of their own actions), and often are forced to make quick assessments based on limited information—which can be quite dangerous. In addition to identifying quantitative metrics for measuring particular tactics' efficacy, some also suggested considering equally rigorous qualitative standards in evaluating its performative or aesthetic impacts.

3. Ethical Justifications: Many have questioned the ethical justifications for obfuscation techniques, highlighting cases in which tactics may seem to enable instances of free-loading, wasting resources, or failing to challenge larger structures of power. These must be carefully considered in context—as well as with regards to asymmetries of information or other forms of power—while also offering legitimacy to obfuscation as a concept and offering confidence to those creating such systems. Several participants also raised ethical questions about who participates in obfuscation tactics and in what ways, noting that people with diverse and intersecting identities may be unevenly impacted by different forms of surveillance and resistance.

4. Safeguarding Obfuscation: Widely-available technologies and platforms may not support or allow for obfuscation, as a matter of function or policy. In anticipation of those who would not like obfuscation to take place, many asked how to best create space for the development of a toolkit

of obfuscation. Several speakers also raised questions about who is best engaged to support obfuscating tactics: technologists, researchers, activists, or everyday users?

Needless to say, the workshop provoked more questions than answers, and we anticipate these themes and queries will continue to shape obfuscation research moving forward.

In keeping with the workshop format, rather than publishing a formal set of proceedings, we have asked our panelists to each provide a brief essay summarizing their project, concept, application—but with an emphasis on the questions, challenges, and discussions raised during the weekend, as well as those they anticipate will guide future research in this area. In such a way, the pieces in this collection constitute a small taste of the wide range of research in the field of obfuscation, some of which may be found by consulting their recent publications, but much of which continues to be a work in progress. As with the workshop itself, this report is a starting point rather than an end point.

To conclude, we would like to express our gratitude to all of our presenters for taking on the difficult task of defining this emerging field, particularly as it engages diverse theoretical backgrounds, methodologies, and applications. We would also like to thank our planning committee (see page 2), as well as our sponsors: the NYU department of Media, Culture, and Communication and NYU Law School's Information Law Institute, and the National Science Foundation. We look forward to continuing to build this research community and to improving our means of putting obfuscation in practice.

FINN BRUNTON is a scholar of the relationships between society, culture and information technology — how we make technological decisions, and deal with their consequences. He focuses on the adoption, adaptation, modification and misuse of digital media and hardware; privacy, information security, and encryption; network subcultures; hardware literacy; and obsolete and experimental media platforms. He is the author of Spam: A Shadow History of the Internet (MIT, 2013), along with numerous articles and talks. Brunton received an MA from the European Graduate School (Saas-Fee, Switzerland) and a PhD from the University of Aberdeen's Centre for Modern Thought. Prior to his NYU appointment, he was an Assistant Professor of Information at the University of Michigan's School of Information.

HELEN NISSENBAUM is Professor of Information Science at Cornell Tech and currently on leave from New York University, Media, Culture, and Communication and Computer Science. Prof. Nissenbaum's work spans societal, ethical, and political dimensions of information technologies and digital media. Her books include Obfuscation: A User's Guide for Privacy and Protest, with Finn Brunton (MIT Press, 2015), Values at Play in Digital Games, with Mary Flanagan (MIT Press, 2014), and Privacy in Context: Technology, Policy, and the Integrity of Social Life (Stanford, 2010). Grants from the National Science Foundation, Air Force Office of Scientific Research, Ford Foundation, the U.S. Department of Health and Human Services Office of the National Coordinator, and the Defense Advanced Research Projects Agency have supported her research on privacy, trust online, cyber security, and values in design. Recipient of the 2014 Barwise Prize of the American Philosophical Association, Nissenbaum has contributed to privacy-enhancing software, including TrackMeNot (for protecting against profiling based on Web search) and AdNauseam (protecting against profiling based on ad clicks). Nissenbaum holds a Ph.D. in philosophy from Stanford University and a B.A. (Hons) from the University of the Witwatersrand.

PRIVACYVISOR: PRIVACY PROTECTION FOR PREVENTING FACE DETECTION FROM CAMERA IMAGES

ISAO ECHIZEN, National Institute of Informatics, Tokyo

Due to the popularization of portable devices with built-in cameras and advances in social networking services and image search technologies, information such as when and where a photographed person was at the time of the photograph can be revealed by the posting of photos online without the person's permission. This has resulted in a greater need to protect the privacy of photographed individuals. A particularly serious problem is unauthorized information revelation through the posting of images of people captured unintentionally and shared over the Internet. If, for example, your face or figure is unintentionally captured in an image taken by someone, and then that image is shared by posting it on a social networking site, information about where you were and when can be revealed through the face recognition process of an image retrieval service (e.g., Google Images) that can access the geographic location and shooting date and time information contained in the image's geotag without your permission.

An experiment conducted at Carnegie Mellon University showed that the names of almost one-third of the people who participated could be determined by comparing the information in photographs taken of them with the information in photographs posted on a social networking site. Furthermore, other information about some of the participants, including their interests and even their social security number, was found [1]. A commercial facial recognition application for smartphones called FindFace that was released in 2016 in Russia can identify people in public from their profile image on a Russian social media site. After the release of FindFace, a Russian photographer initiated a project called "Your face is big data." It showed that 70 of 100 people photographed in the subway without permission could be identified using FindFace [2].

Unlike the capturing of images with surveillance cameras, in which the images are managed by an administrator and are generally not posted online, the situation addressed here is that in which a person's image is captured unintentionally in a photograph with the person possibly being unaware that a photograph is being taken, such as in a photograph taken at a tourist attraction. The photograph may then be posted online without that person being able to control the posting, resulting in possible unauthorized disclosure of personal information.

Methods proposed for preventing unauthorized face image revelation include hiding the face

with an unfolded shell [3] and painting particular patterns on one's face [4]. The first method physically protects the user's privacy by using material in the shape of a shell that can be folded and unfolded. When folded, it functions as a fashion accessory; when unfolded, it functions as a face shield, preventing unintentional capture of the wearer's facial image. The second method prevents identification of the person by using particular coloring of the hair and special paint patterns on the face that cause facial recognition methods to fail. However, such methods interfere with face-to-face communication because they hide a large portion of the face and/or distract the attention of the person to whom the wearer is communicating.

We previously proposed using invisible noise signals to prevent privacy invasion [5]. This method uses infrared LEDs as a light source to add noise to a captured image. Although the infrared rays do not affect the human eye, there are two problems: a power supply is needed for the LEDs, and some digital cameras are unaffected by the rays. Consumer camcorders, for example, use infrared wavelengths to enable them to adjust the settings for dark situations. The sensitivity to infrared varies among cameras, and some cameras do not react to infrared rays. A method using infrared rays is thus ineffective against them [6].

We have developed a method for overcoming these two problems [7]. It prevents face image detection without the need for a power supply by using materials that naturally absorb and reflect incident radiation. It is effective against all digital cameras because it uses visible rather than infrared light, and it negligibly interferes with face-to-face communication in physical space. The small amounts of light- reflecting and absorbing materials attached to a goggle-like visor (a "PrivacyVisor") effectively obscure the Haar-like features used for face detection and thus cause face detection to fail [8]. Moreover, no new functions need to be added to existing cameras and/or networking services.

Increased performance of sensors enables biometric identities to be obtained in more ways than could have ever been anticipated. Future work will thus focus on protecting against biometric identity theft via image sensors. We are developing a method (a "BiometricJammer") that will prevent the surreptitious photographing of fingerprints and subsequent acquisition of fingerprint information while still enabling the use of the fingerprint authentication methods that are normally used by smartphones and the like [9]. Other forms of biological information besides fingerprints that could be used for personal identification or authentication include the patterns of the iris and of the veins in the fingers or palms. We plan to continue researching and developing methods aimed at preventing the illegal acquisition of each type of information.

REFERENCES:

[1] A. Acquisti, R. Gross, and F. Stutzman, "Face Recognition Study—FAQ," August 2011, http://www.heinz.cmu.edu/~acquisti/face-recognition-study-FAQ/ [Accessed May 26, 2017]

[2] E. Wilson, "Your face is big data," BBC News, [Online] April 13, 2016, http://www.bbc.com/news/av/magazine-36019275/your-face-is-big-data [Accessed May 26, 2017]

[3] R. Hernandez, "Veasyble by GAIA," March 8, 2010, https://www.yatzer.com/Veasyble-by-GAIA [Accessed May 26, 2017]

[4] A. Harvey, "CV Dazzle," https://cvdazzle.com/ [Accessed May 26, 2017]

[5] T. Yamada, S. Gohshi, and I. Echizen, "Privacy Visor: Method for Preventing Face Image Detection by Using Differences in Human and Device Sensitivity," Proc. of the 14th Joint IFIP TC6 and TC11 Conference on Communications and Multimedia Security (CMS 2013), LNCS 8099, pp. 152-161, Springer (September 2013)

[6] See: "Privacy visor glasses jam facial recognition systems to protect your privacy #DigInfo," Ikinamo, June 19, 2013, https://www.youtube.com/watch?v=LRj8whKmN1M&t=23s [Accessed May 26, 2017]

[7] T. Yamada, S. Gohshi, and I. Echizen, "Privacy Visor: Method based on Light Absorbing and Reflecting Properties for Preventing Face Image Detection," Proc. of the 2013 IEEE International Conference on Systems, Man, and Cybernetics (IEEE SMC 2013), 6 pages (October 2013)

[8] See: "Privacy visor fools facial recognition," IDG.tv, August 20, 2015, https://www.youtube.com/watch?v=HbXvZ1XKdWk&t=1s [Accessed May 26, 2017]

[9] See: "'Peace' signs risk fingerprint theft, says Japanese study," Reuters, January 16, 2017, https://www.youtube.com/watch?v=vJn9cx-CyPE [Accessed May 26, 2017]

ISAO ECHIZEN is a professor of the National institute of informatics (NII). He received B.S., M.S., and D.E. degrees from the Tokyo Institute of Technology in 1995, 1997, and 2003. He joined Hitachi, Ltd. in 1997 and until 2007 was a research engineer in Hitachi's Systems Development Laboratory. He was a visiting professor at the University of Freiburg in 2010 and a visiting professor at the University of Freiburg and the University of Halle-Wittenberg in 2011. He is currently conducting research in the fields of content security and privacy and of multimedia application systems. He is a member of Information Forensics and Security Technical Committee (IFS TC), IEEE Signal Processing Society.

CIRCUMVENTION THROUGH OBFUSCATION

AMIR HOUMANSADR, University of Massachusetts Amherst

THE PROBLEM OF INTERNET CENSORSHIP

The Internet plays a crucial role in today's social and political movements by facilitating the free circulation of speech, information, and ideas; democracy and human rights throughout the world critically depend on preserving and bolstering the Internet's openness. Recent events in Tunisia, Egypt, Turkey, and the rest of the world give strong indications that oppressive regimes can even be overthrown by the power of people mobilized to fight by organizing, communicating, and raising awareness through use of the Internet. Consequently, repressive regimes, totalitarian governments, and corrupt corporations regulate, monitor, and restrict access to the Internet, which is broadly known as *Internet censorship*.

CENSORSHIP TECHNIQUES

The techniques commonly used to enforce censorship include IP address blocking, DNS hijacking, and TCP content filtering to block access to certain destinations or to prevent certain forms of content from being transmitted. To ensure compliance and to detect undercover political/social activists, repressive regimes additionally utilize Deep Packet Inspection (DPI) and other techniques to disable the operation of all censorship circumvention technologies by their citizens. Consequences of non-compliance can be severe, ranging from identification and termination of employment to life-threatening prosecutions under repressive governments.

COMMON CIRCUMVENTION MECHANISMS

To help the affected users bypass censorship, various groups of researchers and practitioners have designed and deployed a toolset of systems, called circumvention systems or anti-censorship tools. Such systems use various techniques to disable the censorship mechanisms introduced above, i.e., IP address blocking, DNS interference, and DPI-based keyword filtering. We roughly classify existing censorship circumvention tools into the following groups:

Tools that Obfuscate Identity: The most common technique for censorship is to blacklist the IP addresses of the forbidden websites. Therefore, a large number of circumvention systems try to obfuscate the IP addresses (identities) of the websites or services being browsed by the censored users. Such systems include the widely used HTTP proxies, VPN services, and their variants such as the Tor network. Other recent techniques obfuscate traffic by running entangling circumvention identities with that of popular Internet services like cloud services and CDNs.

Tools that Obfuscate Content: Modern censorship technologies are able to perform

Deep-Packet Inspection, i.e., inspect the content of network traffic for forbidden keywords and content. Therefore, most circumvention tools deploy mechanisms to obfuscate the content of (forbidden) communication by the censored user. The most trivial way of obfuscating content is encrypting packet contents using keys shared between the users and the circumvention servers. To defeat omniscient censors who whitelist traffic (instead of blacklisting) a new circumvention proposals encrypts traffic such at it matches the regular expressions of normal traffic (this known as format-transforming encryption).

Tools that Obfuscate Protocol: Modern censors aim at blocking popular circumvention systems like Tor and VPNs. Such systems perform efficient mechanisms to obfuscate content and identity (IP addresses), however, the censors try to detect them based on the patterns of their network communications. For instance, Tor traffic is comprised of packets with unique sizes that easily identify Tor traffic to the censors. Therefore, modern circumvention tools aim at obfuscating their underlying protocol to evade blocking. In particular, several new mechanisms modify Tor traffic such that its traffic pattern imitate that of a non-forbidden protocol like Skype.

SUMMARY

While there are a wide range of censorship circumvention technologies, they have one thing in common: they all deploy obfuscation on way or another to defeat the censors. The implemented obfuscation trades off resistance to censorship with the quality of service provided by such systems, e.g., too much obfuscation can slow down the Internet browsing experience of the censored users. Therefore, the key challenge to designing circumvention systems is keeping the right balance between censorship resistance efficiency and usability.

AMIR HOUMANSADR is an assistant professor at the College of Information and Computer Sciences at the University of Massachusetts Amherst, where he joined in 2014. He received his PhD from the University of Illinois at Urbana-Champaign in 2012, and was a postdoctoral scholar at the University of Texas at Austin before joining UMass. Amir's area of research is network security and privacy, which includes problems such as Internet censorship resistance, statistical traffic analysis, location privacy, cover communications, and privacy in next-generation network architectures. Amir has received several awards including the Best Practical Paper award at the IEEE Symposium on Security & Privacy (Oakland) in 2013, a Google Faculty Research Award in 2015, and an NSF CAREER Award in 2016.

POLITICAL RHETORIC AS OBFUSCATION AND FINDING SOLUTIONS WITH NEURAL NETWORKS

NICOLE COTE and **ROB HAMMOND**, New York University

With the relationship between rhetoric and political language, alongside the heightened dissemination of new information by means of the Internet, it is difficult to cipher the language that purposefully eludes its audience. Rhetoric is defined by the *Oxford English Dictionary* as "the art of using language effectively so as to persuade or influence others, esp. the exploitation of figures of speech and other compositional techniques to this end" [1]. We believe that this deception is a form of obfuscation where politicians hide the ideas in their language through misleading narrative that is intended to confuse their true intentions.

We therefore suggest that in order to fully analyze a political speech with a natural language model, the existence of rhetoric needs to be considered; it is then necessary to ask if and how a model might accomplish the tasks of analyzing or filtering rhetoric. This project does not deal with obfuscation as a privacy tactic to redress a power imbalance in order to protect or hide information against a more powerful adversary. It approaches obfuscation not as a tool for users, but as something that can be exploited by those in a position of power to mislead a generally less informed audience to further the asymmetry of power. The theoretical model that we have proposed in this project would thus hope to complete two tasks in an effort to tackle rhetoric: 1) extractive summarization of the underlying ideas present in political text, and 2) categorization of the summarized political speeches into specific political ideologies; both combined are, to our knowledge, not tasks of which a singular existing language model is capable. The neural network approach we take stems from analysis of existing natural language processing and neural network approaches to political speech [2], attention [3][1], and sentence summarization and entailment [4, 5], among other methods.

Our model also theoretically connects to the CLSA model posited by Cambria, et al. [6], who suggest that a truly effective singular language model should be an ensemble of models that excel at individual tasks; they accordingly offer a hypothetical model that takes all existing individual tasks, and puts them together to create a holistic individual model. The project is accordingly posited as a means to explore rhetoric, an open issue in neural network research, by thinking about how we might create one holistic model that combines numerous individual tasks from neural network models that already exist,

and use these tasks individually for political text. With our theoretical model we would hope to be able to read a speech, extract a summary, and then classify the sentiment of the speech into a political ideology in order to inform a reader about a politician. We are interested in thinking about how to use politicians' actual words (and not those of journalists writing about them) for training a neural network language model. Specifically, we have investigated a way to train the model on transcripts from speeches made by politicians within the context of a campaign, political rally, or debate. There are some general challenges on how to incorporate concepts such as the semantic separation of message and noise in a political speech where rhetoric and content may be found to be intertwined. However, most of the individual tasks involved in the project are already well-defined tasks within language modeling where this is work being produced to actively build upon the cited works.

Because this work is an attempt to join multiple models rather than create a new task, much of the model is relatively well defined from a computational perspective, but getting adequate data for supervised learning poses the biggest problem for this work. As this is a supervised machine learning task, a nuanced model requires finding ways to define "rhetoric" for the necessary tagging of the dataset (which would need to be comprised of thousands of political speeches) on which the model would train. A non-exhaustive list of linguistic options explored include ideas such as: repetition, excessive synonyms, frequent oppositional sentences/phrases (bait and switch), frequent pronouns (we, you, they). Because of the intertwined nature of content and rhetoric, an already complex idea to define, tagging thousands of speeches in a scalable, and consistent manner will be the most challenging aspect of this project. The model remains theoretical as curating such a dataset would have substantial financial, computational, and time costs. This theoretical model asks questions of how we solve problems with neural network models and how such models address nuanced language issues, such as the obfuscation of language through rhetoric—an area we identify as needing further exploration in the obfuscation and NLP communities.

REFERENCES

[1] http://www.oed.com/

[2] Iyyer,Mohit and Enns,Peter and Boyd-Graber,Jordan and Resnik,Philip. 2014. Political Ideology Detection Using Recursive Neural Networks. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1113-1122.

[3] Olah,Chris and Carter,Shan. 2016. Attention and Augmented Recurrent Neural Networks.

[4] Bowman, Samuel R. and Angeli, Gabor and Potts, Christopher and Manning, Christopher D. 2015. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

[5] Rush, Alexander M. and Chopra,Sumit and Westo, Jasonn. 2015. A Neural Attention Model for Abstractive Sentence Summarization. *arXiv preprint arXiv:1509.00685*.

[6] Cambria, Erik and Poria, Soujanya and Bisio, Federica and Bajpai, Rajiv and Chaturvedi, Iti. 2015. The CLSA Model: A Novel Framework for Concept-Level Sentiment Analysis. *International Conference on Intelligent Text Processing and Computational Linguistics.*

Nicole Cote is an MS student in Integrated Digital Media at NYU's Tandon School of Engineering, and received an MScR in Victorian Literature from the University of Edinburgh. She is interested in the relationship between technology and social justice issues, text (particularly culturally and/or historically significant text) as data, and data visualization.

Rob Hammond is a part-time Masters student at the NYU Center for Data Science and works at Risk Management Solutions by day in catastrophe risk analysis. He is particularly driven by technology, data, privacy, and environmental issues and is keen to de-obfuscate data to find the signal in the noise where it's appropriate.

OBFUSCATING DATA TO PREVENT DISCRIMINATION

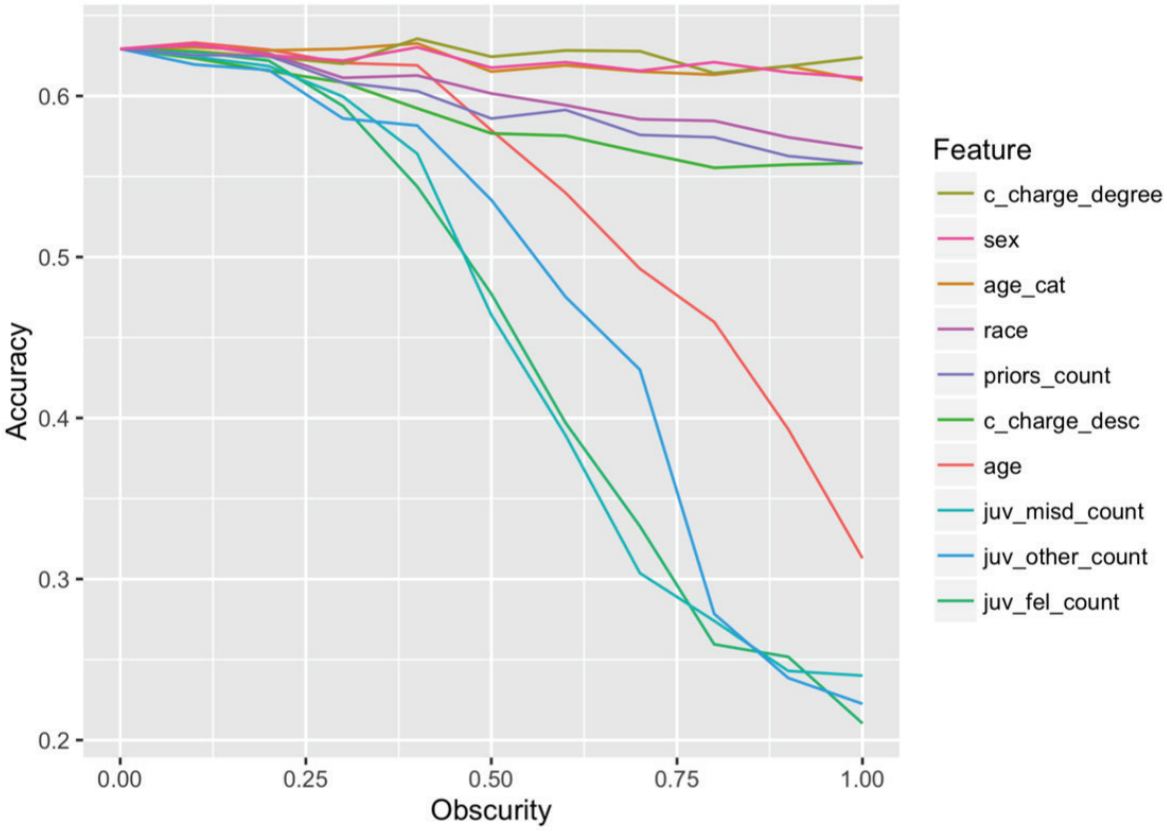
SORELLE FRIEDLER, Haverford College

Obfuscating data with respect to protected information can serve to decrease discrimination and increase transparency, even in the face of black-box proprietary models.

Algorithms are increasingly being used to make high-stakes decisions that directly impact people's lives. Such algorithms may use past data about a company's hiring practices to determine who should be hired in the future [16], determine where to send police based on historical arrest data [4], or be used to relieve overcrowding in jails by releasing those predicted most likely to reappear without bail [7]. As high-stakes decision-making about people has become more driven by machine learning processes, these algorithmic choices have also begun to come under more scrutiny. A recent Wisconsin court case challenged the right to use proprietary recidivism prediction algorithms at sentencing time [21], Philadelphia's pre-trial risk assessment has been receiving press that focuses on its potential racist impacts [20], and Chicago's predictive policing algorithms have also been viewed as automated racial profiling [18]. These worries, especially around the potential for discrimination arising from machine-learned decisions about people, have led to prominent calls for more accountability in algorithmic decisions [19, 1].

Perhaps the most obvious way that machine-learned decisions can become discriminatory is by using training data that directly encodes human biases, for example creating a hiring algorithm based on historical hiring decisions at an all-white company may lead an algorithm to discriminate against people of color. More subtly, data collection feedback loops may reinforce incorrect algorithmic notions, for example a predictive policing algorithm that keeps sending police back to the same neighborhood because the algorithm sent police there the previous day [15]. Issues may also arise due to systematic differences between groups—patterns that were machine-learned on traditionally Western names may not hold on the Native American subpopulation. This can be compounded without careful evaluation of the algorithm, since many metrics weight errors per-person instead of per-group, so an algorithm that has the incorrect outcome on all Native Americans (about 2% of the U.S. population) could be evaluated as 98% successful. Many solutions have now been put forward to try to perform fairness-aware machine learning. Broadly, the interventions can be characterized as those performed by pre-processing the training data [8, 24, 5, 11], by directly changing the machine-learning algorithm [13, 6, 23], and by post-processing the outcomes [12]. In addition, recent work has focused on fair decision making with feedback loops [10, 9, 14].

What does this have to do with obfuscation? One way to think about the removal of protected



information, such as race or sex, from a training data set (pre-processing) is as the *obfuscation* of the data set with respect to that information. In fact, it has been shown that the discriminatory impact of any classifier can be estimated from the training data by measuring the error in attempting to predict the protected information [8]; i.e., if your race can be guessed, it can be used to discriminate against you. This idea can also be used to certify a data set as safe from potential discrimination; if a protected feature can't be predicted from the remaining features, then the information from that feature can't influence the outcome of the model. In other words, if race can't be predicted (with low error) from the remaining data in the training set, a machine learning algorithm trained on this data can't discriminate based on race. Thus, obfuscating the data with respect to protected class serves to prevent discrimination.

However, referring to this procedure as obfuscation implies that the observed data being used to train the machine learning model is correct and does not itself suffer from systemic bias. Instead, if the belief is that any distributional differences between groups in the data is the result of observational bias, then this procedure can be viewed as *repairing* the data so that it more properly reflects the underlying truth. One such repair procedure works by modifying the protected class-conditioned distributions per-attribute so that they look more similar [8]. This works by effectively grouping people based on their per-group quantile and assigning all members of that quantile the same score, specifically the score that is the quantile's median over the groups. This preserves the within-group ranks for each person, serving to preserve some predictive capacity of the data. Experimentally, this procedure has been shown to result in fair

classifiers (measured using the disparate impact four-fifths ratio [22]).

Using this repair procedure can cause a drop in accuracy (or other measures of utility) in the resulting classifier. While this is often framed as a problem for the effectiveness of fairness-aware machine learning—a "tradeoff" between fairness and accuracy—it can also be viewed as a measure of the extent to which the protected class was used by the machine learning model. In fact, this procedure can be used to *remove* any attribute from the data set by *obscuring* the remaining attributes with respect to that one. The importance of the removed attribute can then be measured based on the model's drop in accuracy—removed attributes that have a larger drop in accuracy had a larger influence on the model's outcomes [2]. This procedure measures the *indirect influence* of an attribute; the effect of correlated or proxy variables is included in the overall influence ascribed to the feature, so that if zip code is used by a model as a proxy for race, race is considered to have an influence on the model's outcomes.

This tool for auditing for indirect influence allows the partial *de-obfuscation* of black-box systems [2]. For example, a groundbreaking study by ProPublica [3] recently investigated a risk assessment instrument called COMPAS [17] and found that it was biased against black defendants in the sense that the misclassification rates were skewed so that black defendants were more likely to be incorrectly labeled high risk, while white defendants were more likely to be incorrectly labeled low risk. With direct access to COMPAS, we could run the above procedure to determine the indirect influence of each variable on the outcome. Unfortunately, COMPAS is considered proprietary by Northpoint, the company that created it—even the full inputs are unknown. Without such access, we can instead attempt to model the COMPAS low / medium / high outcomes using data released by ProPublica. Modeling these outcomes using a support vector machine (SVM) model, we see that juvenile records and age are most important in predicting these outcomes (obscuring these attributes causes a large drop in model accuracy), while race is also important but to a much lesser extent. However, the somewhat low starting accuracy of the SVM model, due to the lack of access to either COMPAS or the true inputs, weakens these results. With access to COMPAS, such conclusions would more directly serve to de-obfuscate the decision-making process.

REFERENCES

This paper describes work published in [2] and [8].

[1] More accountability for big-data algorithms. *Nature*, 537(449), Sept. 21 2016.

[2] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian. Auditing black-box models for indirect influence. In *Proc. of the IEEE International Conference on Data Mining (ICDM),* 2016.

[3] J. Angwin, J. Larson, S. Mattu, and L. Kirchner Machine bias. *ProPublica*, May 23, 2016.

[4] J. Bachner and J. Lynch. Is predictive policing the law-enforcement tactic of the future? *The Wall Street Journal*, Apr. 24 2016.

[5] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *ICDM Workshop Domain Driven Data Mining*, pages 13–18, 2009.

[6] T. Calders and S. Verwer. Three naïve Bayes approaches for discrimination-free classification. *Data Min Knowl Disc*, 21:277–292, 2010.

[7] K. Colaneri. Can a computer algorithm be trusted to help relieve philly's overcrowded jails? *NewsWorks*, Sept. 1 2016.

[8] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. *Proc. 21st ACM KDD*, pages 259–268, 2015.

[9] M. Joseph, M. Kearns, J. Morgenstern, S. Neel, and A. Roth. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559, 2016.*

[10] M. Joseph, M. Kearns, J. H. Morgenstern, and A. Roth. Fairness in learning: Classic and contextual bandits. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 325–333. Curran Associates, Inc., 2016.

[11] F. Kamiran and T. Calders. Classifying without discriminating. In *Proc. of the IEEE International Conference on Computer, Control and Communication*, 2009.

[12] F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. *Proc. of the IEEE 12th International Conference on Data Mining*, 2012.

[13] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

[14] S. Kannan, M. Kearns, J. Morgenstern, M. Pai, A. Roth, R. Vohra, and Z. S. Wu. Fairness incentives for myopic agents. *arXiv preprint arXiv:1705.02321, 2017.*

[15] K. Lum and W. Isaac. To predict and serve? *Significance*, pages 14—18, October 2016.

[16] C. C. Miller. Can an algorithm hire better than a human? *The New York Times: The UpShot*, June 25 2015.

[17] Northpointe. COMPAS - the most scientifically advanced risk and needs assessments. http://www.northpointeinc.com/risk-needs-assessment.

[18] M. S. on. The minority report: Chicago's new police computer predicts crimes, but is it racist? *The Verge*, Feb. 19 2014.

[19] J. Podesta, P. Pritzker, E. J. Moniz, J. Holdren, and J. Zients. Big data: seizing opportunities, preserving values. Executive Office of the President, May 2014.

[20] J. Reyes. Philadelphia is grappling with the prospect of a racist computer algorithm. *Technically Philly*, Sept. 16 2016.

[21] M. Smith. In wisconsin, a backlash against using data to foretell defendants' futures. *The New York Times*, June 22 2016.

[22] The U.S. EEOC. Uniform guidelines on employee selection procedures, March 2, 1979.

[23] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness constraints: A mechanism for fair classification. In *ICML Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, 2015.

[24] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proc. of Intl. Conf. on Machine Learning*, pages 325–333, 2013.

SORELLE FRIEDLER is an Assistant Professor of Computer Science at Haverford College and an Affiliate at the Data & Society Research Institute. Her research interests include the design and analysis of algorithms, computational geometry, data mining and machine learning, and the application of such algorithms to interdisciplinary data. Sorelle is one of the organizers of the Workshop on Fairness, Accountability, and Transparency in Machine Learning and has received a Data & Society Fellowship and recent NSF Grant for her work on preventing discrimination in machine learning. Her work on this topic has been featured in IEEE Spectrum, Gizmodo, and NBC News and she has been interviewed about algorithmic fairness by the Guardian, Bloomberg, and NPR. Sorelle holds a Ph.D. in Computer Science from the University of Maryland, College Park.

USING ETHICALLY-CONSTRAINED GAME THEORY TO PROTECT OUR PRIVACY

Jeffrey Pawlick and **Quanyan Zhu**, New York University Tandon School of Engineering

Those who propose the use of obfuscation to protect privacy must understand the consequences of using obfuscation, and judge when those consequences are morally justifiable. Opponents accuse obfuscators of "the valorization of deceit and dishonesty… wastefulness, free riding, database pollution, and violation of terms of service" [1]. Indeed, these concerns draw upon strong cultural norms against lying and deceit—as well as particularly powerful recent movements against wastefulness and environmental pollution. What type of criteria can be used to evaluate these harms against the potential benefits of obfuscation?

As mathematical game theorists, we typically answer that a utility function can be constructed: the first term in the function can represent privacy benefits, and the second term can represent harms. If the first term outweighs the second—we might argue—then obfuscation is justified. We might imagine that this argument draws off the purity and objectivity of mathematics, and therefore is not burdened by any particular philosophical view.

Of course, that is not the case. Game theorists in this community implicitly adhere to consequentialist or utilitarian points of view, such as those advanced, for instance, by John Stuart Mill. In particular, "*act consequentialism* is the claim that an act is morally right if and only if that act maximizes the good, that is, if and only if the total amount of good for all minus the total amount of bad for all is greater than this net amount for any incompatible act available to the agent" [2,3]. Extreme versions of consequentialism might be summed up in the vernacular phrase: "the ends justify the means."

Of course, not all philosophers (or non-philosophers) are consequentialists, and therefore not everyone judges obfuscation solely by its ends. For example, those who criticize obfuscation as dishonest draw upon the idea of lying as *malum in se*. Others claim that obfuscation violates terms of service which have legal force, and is therefore *mala prohibitum*—wrong because it is illegal—regardless of its consequences.

First, we address the claims of dishonesty. While it is true that many traditions reject lying as *malum in se* (e.g., those of Kant and Aquinas), most obfuscation does not involve lying. Lying requires "making a believed-false statement" [4], and obfuscation techniques do not make statements. Obfuscation is indeed deception, but deception is probably not *malum in se*.

Second, we consider the argument that (online) obfuscation violates terms of use and is therefore *malum prohibitum*. Our argument is that terms of use do not have legal force, and even less likely moral force. It would take the average Internet user 76 days to read the privacy policies of all of the websites he or she visits each year [5]. This conveys the general idea that policies online are not effectively promulgated. It seems, then, that obfuscation is neither *malum in se* nor *malum prohibitum*.

Obfuscation, therefore, passes at least two litmus tests for moral permissibility as a means. Of course, that in itself does not make obfuscation justified. But it suggests that we can return to considering obfuscation's consequences. Arguments along these lines can perhaps be found in just war doctrine [6] or the principle of double effect [7]. According to both of these ideas, an act which produces bad effects can be tolerated (under certain criteria) if its bad effects are proportional to its intended good. Obfuscation certainly can produce harms; it may waste computational cycles, degrade personal advertising, or even distract law enforcement. These must be weighed against the good of protecting users' privacy. But how can we evaluate this proportionality?

Here, we return full circle to game theory. Utility functions are poor tools to capture complex ethical issues, but they are excellent tools to capture proportionality. Game theory is a branch of mathematics which models strategic interactions between two or more rational agents. (See, e.g., [8].) Models in game theory assume that agents choose strategies in a way that anticipates the strategies of the other agents. A game-theoretic *equilibrium* predicts the strategies at which rational agents will deadlock. We will use equilibrium predictions to assess the long-term consequences of obfuscation technologies.

Consider a game-theoretic model with *N+1* players: *N* users and one machine learning agent which computes some statistic of the users' data. The total utility function of each user is composed of an accuracy term, a privacy term, and a term that reflects the cost of perturbation:

$$U_S^i = A_S^i \exp\{\epsilon_g\} - P_S^i(1 - \exp\{\epsilon_p\}) - C_S^i \mathbf{1}_{\sigma_S^i > \mathbf{0}}.$$

Here, ϵ_g is proportional to accuracy, ϵ_p is inversely related to privacy, and $C_S^i \mathbf{1}_{\sigma_S^i > \mathbf{0}}$ implies that user *i* pays a flat cost of C_S^i for using obfuscation. If user *i* perturbs her data maximally, then she receives zero benefit for accuracy and zero loss for privacy, and she pays the cost C_S^i for using obfuscation. On the other extreme, if user *i* does not perturb at all, then she gains A_S^i for accuracy and loses P_S^i for privacy. The utility function for the learner is similar, except that it does not have a term related to privacy.

We have three tasks. The first is to determine the conditions under which user *i* will employ obfuscation. (See [9, 10] for the mathematical details.) There are three cases to consider. First, if users are much more accuracy-sensitive than privacy-sensitive, then they will never obfuscate. Second, if the opposite is true, then they will always obfuscate. The most interesting case is in the middle. In this intermediate case, the equilibrium predicts that user *i* will obfuscate if other users—on average—obfuscate above a threshold amount. This is important. It suggests a strategic reason why adoption could cascade; this reason is in addition to the type of epidemic spreading often seen in technology adoption.

Our second task is to ask how a machine learning agent can avoid this large-scale adoption of

obfuscation. For if many users perturb, then the learning agent's accuracy will be greatly decreased. We find that if the learning agent proactively provides a sufficient level of privacy protection, then the users will have no incentive to obfuscate. Their obfuscation was only a tool to protect their privacy, and if the learning agent does this himself, then the users are content to submit truthful data.

The third task is to analyze whether providing this protection is incentive-compatible for the machine learning agent. In other words, if he concedes some accuracy in order to protect privacy to some degree, can he improve his outcome in the long run by avoiding cascading adoption of obfuscation?

We find that he can—but only under certain circumstances. Protection is incentive-compatible if obfuscation is cheap for the learner and expensive for the users. Perhaps more surprisingly, protection is also incentive-compatible for the learner to the degree that he is highly accuracy-sensitive. This sensitivity will make him more cautious about cascading adoption of obfuscation. Finally, this incentive-compatibility is proportionally difficult to the degree to which users are privacy-sensitive. The more sensitive they are, the more the learning agent will have to perturb, and the more this will cost him.

While much work remains to be done, we have shown that if 1) users care about privacy enough to cause cascading adoption of obfuscation and 2) obfuscation is sufficiently cheap (and accuracy is sufficiently important) for the learning agent, then it is optimal for learning agents to avoid obfuscation by protecting privacy themselves. In this case, *the threat* of obfuscation is sufficient to accomplish its objective; and this satisfies the type of proportionality that makes obfuscation morally justified.

REFERENCES

[1] Brunton, Finn and Helen Nissenbaum. *Obfuscation: A User's Guide for Privacy and Protest.* The MIT Press, Cambridge, Massachusetts, 2015.

[2] Sinnot-Armstrong, Walter. "Consequentialism," *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition), Edward N. Zalta (ed.). (Online. Available: https://plato.stanford.edu/archives/win2015/entries/consequentialism/.) 2015.

[3] Moore, George Edward. *Ethics*. Oxford University Press, New York. 1912.

[4] Mahon, James Edwin. "The definition of lying and deception," *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.). (Online. Available: https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=lying-definition.) 2016.

[5] McDonald, Aleecia and Lorrie Faith Cranor. "The cost of reading privacy policies," *Privacy Year in Review*, 2008.

[6] *Catechism of the Catholic Church: Paragraph 2309*, 1992.

[7] McIntyre, Alison. "Doctrine of double effect," *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.). (Online. Available: https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=double-effect.)

[8] Fudenberg, Drew and Jean Tirole. *Game Theory*. The MIT Press, Cambridge, Massachusetts, 1991.

[9] Pawlick, Jeffrey and Quanyan Zhu. "A Stackelberg game perspective on the conflict between machine learning and data obfuscation," *IEEE Workshop on Information Forensics and Security*, 2016. (Available: https://arxiv.org/pdf/1608.02546.pdf.)

[10] Pawlick, Jeffrey and Quanyan Zhu. "A Mean-Field Stackelberg Game Approach for Obfuscation Adoption in Empirical Risk Minimization," Submitted to *IEEE Global SIP, Symposium on Control & Information Theoretic Approaches to Privacy and Security*, 2017. (Available: https://arxiv.org/pdf/1706.02693.pdf.)

JEFFREY M. PAWLICK studies strategic trust and deception in cyber-physical systems (CPS) and internet of controlled things (IoCT). Security in these heterogeneous and dynamic systems requires new frameworks and equilibrium concepts. CPS and IoCT security also require incentive-compatible mechanism design due to

competition between benign agents, in addition to the threat of attack from malicious actors. To capture these factors Jeffrey leverages results from game theory and dynamic systems together with psychological insights that inspire attack models as well as models for defensive deception. A Ph.D. candidate at New York University's Tandon School of Engineering, Jeffrey holds a B.S. in Electrical Engineering from Rensselaer Polytechnic Institute (Troy, NY).

QUANYAN ZHU is an assistant professor at the Department of Electrical and Computer Engineering at New York University (NYU). He received B. Eng. in Honors Electrical Engineering from McGill University in 2006, M.A.Sc. from University of Toronto in 2008, and Ph.D. from the University of Illinois at Urbana-Champaign (UIUC) in 2013. His current research interests include game theory for cyber security, cyber agility, moving target defense, cyber deception and cyber-physical system security.

IDENTITY OBFUSCATION THROUGH FULLY FUNCTIONAL AVATARS

PAUL ASHLEY, Anonyome Labs

The world we live in changes all the time. We grow, learn, experience, and evolve. We may feel strongly about something one day, and indifferent the next. We also may experience companionship, loneliness, happiness, and sorrow. We go through seasons that define our lives and while they may be intense when we're "in the moment," there is an "ebb and flow" to that intensity. Those feelings, beliefs, circumstances, and their intensity will invariably define us in our totality. We are the culmination of our life experiences where the only thing permanent *is* change.

Yet somehow, the Internet evolved differently: everything we do, say, and record is attached to us like a digital tattoo. Our emails, texts, pictures, purchases, browsing histories, and searches are literally stored *forever*. We have no control or knowledge of what is done with that information. Who has seen it? Who has purchased it? Will it come back to haunt me? Will this matter? Will it have unforeseen consequences in a future I can't predict? Will my credibility, financial credit, or livelihood be negatively impacted?

If I was once dating, and then got married, I'm still recorded somewhere as a "single person, in search of…" Why should a picture of my reckless college antics follow me to that job interview? Why is my mobile number—which I've had for years, used to identify and validate me on so many levels, and the key to my personal private information such as my home address—required to enjoy the benefits of frequenting my favorite grocery store?

With so much of our lives and personal experiences moving to the online world, we should have the same freedom to change, adapt, mature, and evolve that we enjoy in the offline world. Being online shouldn't mean, nor should it entitle anyone or any entity, free and unencumbered access to our personal history, forever and always. "Freedom from permanence" is a natural right and it's one we've enjoyed as humans since we started walking upright. Nothing should have to last forever.

The Internet is broken. It doesn't match how humans naturally want to interact and socialize. Moreover, it has unfairly shifted the balance of power to big companies and bigger governments. The erosion of this fundamental right has led us down a path to a very dystopian future; one where freedom of association, expression, and the ability to hold minority views is at significant risk. For a liberal democracy to function properly, it demands this right for its citizens. Yet somehow, through a combination of fear and

convenience, we've been convinced this erosion is for our own good.

We need a change, the same way we ourselves change. We need to take back control of our online identity, giving people the same freedoms of expression, belief, and experience that they've enjoyed since the advent of free societies.

We need Sudo identities.

Anonyome Labs was created to give people control and freedom over their personal and private information. We believe that people should be able to determine how, what, and with whom they share their personal details. Importantly, we believe the individual should decide what is permanent and what is temporary, not a company or government. We are building tools that shift this balance of power from the public and private data brokers, advertisers, and organizations demanding your personal information, back to *you*, the user.

Leveraging our experience in identity management, cybersecurity, and consumer applications, our team is devoted to giving every user the power to create a proxy, or *avatar*, that can be used in both the online and offline world. An "avatar for anything" can be used for searching, shopping, selling, or socializing. We call our avatars "Sudo" and we are clear in our promise to you:

- We do not know who you are—and we don't want to know. Each Sudo you create uses a private encryption key that resides with you and your device. Access to our systems, authorized or not, will never reveal your personal information to us or anyone else.

- We do not use targeted advertising to make money. Unlike other services, we're not providing you something in exchange for selling everything you do, and everything about you, to the highest bidder.

- Complete and absolute control for every user means securing all Sudo-to-Sudo communications with end-to-end encryption, while ensuring that when you're navigating the online world as a Sudo, you're "hidden in the crowd." While your Sudo has a digital footprint, history, and connections to others, no one will be able to connect that Sudo back to you, unless you want them to.

- Our apps are designed and built with a simple goal in mind: control, privacy, and safety *with* convenience. Using a Sudo should be *easier* and *safer* than *not* using one. We do not subscribe to the notion that being private means opting out, masking, or hiding from the world.

We believe people want a sense of control, security, and freedom over their personal information. We believe having this right brings us more safety and security than indiscriminately leaving our personal information everywhere, with everyone, forever. Rather than relying on regulation, incomprehensible terms of service, and the benevolence of faceless companies and governments to make this determination, people should be able to decide how, what, and for how long, they share with others. We should be able to do online what we do naturally offline. We should have the right and the freedom to *own* and *control* our personal information.

The outcome is the desire to build apps that support a user's ability to control and protect

their personal information:

SudoApp: A zero-knowledge communication tool for managing your calls, texts, and emails to others. As a buffer to those outside your intimate social circles, you can reveal as much or as little as you want, while having access to all the online sites and services (no more walled gardens, in-network constraints, or use-in-exchange-for-spam) you use on your mobile or desktop device. One tool, unlimited use, complete control.

SudoPay: Buy goods and services from online vendors without having to compromise your personal private information, including your physical credit card details. Stop leaving your cards with sites whose security is questionable and could be hacked, and instead use "virtual credit cards" to protect yourself and eliminate the hassles of financial fraud and identity theft.

PAUL ASHLEY is Chief Technology Officer at Anonyome Labs, a startup company focussed on identity obfuscation. The company brings technology to every day users that allow them to interact online and offline in safety, privacy and control. Paul's responsibilities at Anonyome Labs includes overall security and application architecture, product ownership, development, emerging technologies, IP protection (patents), and technical partnerships. Paul has worked extensively in software product development for more than 25 years, providing architectural leadership across a range of security products for mobile, cloud and enterprise environments. He specializes in identity and access management, privacy management, mobile security, cloud security, and advanced threat protection. He has also been the architecture leader for many large enterprise security projects across a range of clients in the US, Europe, Asia and the Middle East. Paul has degrees in Electronics Engineering, Computer Science and a PhD in information security. He is CISSP certified, and a Senior Member of the IEEE.

GO RANDO: RESISTING EMOTIONAL SURVEILLANCE WITH NOISY FEELINGS

BENJAMIN GROSSER, University of Illinois at Urbana-Champaign

Facebook's "reactions" let users express how they feel about a link, photo, or status. While such data might be helpful for one's friends, these recorded feelings also enable increased surveillance, government profiling, more targeted advertising, and emotional manipulation. Go Rando is a web browser extension that obfuscates one's feelings on Facebook. Every time a user clicks "Like", Go Rando randomly chooses one of the six "reactions" for them. Over time, the user appears to Facebook's algorithms as someone whose feelings are emotionally "balanced"—as someone who feels Angry as much as Haha or Sad as much as Love. Users can still choose a specific reaction if they want to, but even that choice will be obscured by an emotion profile increasingly filled with noise. In other words, Facebook won't know if a reaction was genuine or not.

WHAT'S WRONG WITH FACEBOOK REACTIONS?

We've known for years now that "Likes" on Facebook not only tell one's friends what they saw, but also change what the user sees on Facebook in the future. For example, Facebook uses "Like" activity to target ads, to decide which posts appear on the News Feed, and to manipulate user emotions as part of its own studies of human behavior. At the same time, Facebook shares its data with other corporations and government agencies, fueling increased surveillance and algorithmic decision making.

So if "Likes" were already shared widely, what's the harm in a user selecting "Angry", "Sad", or "Love"? When "Like" was the only option, it was a multi-purpose signifier that could mean many things, and was thus harder to algorithmically interpret. Facebook's "reactions" are still reductive of human emotion, but they suggest just enough nuance to encourage algorithmic analysis of state-of-mind. While these analyses will be of questionable accuracy at best, they'll still be used to generate an emotion profile for every Facebook user. When combined with other data available to state agencies and corporations, the potential abuses and misuses are significant.

For example, emotion profiles could affect a user's economic future. Amazon could use reactions to feed dynamic pricing. Banks might see "Sad" or "Angry" customers as a higher credit risk for a loan. Or a future employer could treat a "Sad" profile as a sign to negotiate a lower salary or to skip that candidate altogether.

Civilian police use analytics software that draws on social media data for the purposes of intelligence gathering, crowd management, and threat analysis. From "Likes" to hashtags to emojis, recent articles have revealed how

Figure 1: The "Go Rando" logo

this data gets used to track activist locations in real-time during protests, or to analyze the threat an individual poses based on how they "feel." The addition of Facebook's reactions into these systems will lead to further (questionable) analyses of state-of-mind, possibly using how one feels as partial justification for surveillance, arrest, or otherwise.

The US Government and other state actors have long been tracking everyone's digital activities in an attempt to predict future security threats. As they integrate every "Angry", "Sad", or "Wow" users post into their prediction algorithms, that data could lead to increased surveillance, placement on watch lists, and/or rejection of individuals at the border.

Finally, this should all be considered within the context of the recent US presidential election and Brexit votes. While there is still some dispute as to just how extensive their actions were, it is clear that the Trump campaign hoped to use social media data to influence citizens. For example, they engaged the predictive analytics company Cambridge Analytica to utilize such data in order to ascertain the personalities of individuals and groups. This allowed the campaign to glean people's "needs and fears, and how they are likely to behave." The analyses were then used to craft custom messages for voters based on a division of "the US population into 32 personality types." (The same company played a role in the "leave" side of the Brexit vote in Great Britain). Given the policy intentions of the Trump administration on issues like immigration, terrorism, and more, it is likely that these campaign techniques will now become government surveillance techniques.

WHY GO RANDO?

All of the varied parties interested in social media data—whether it's Facebook itself, other corporations, or governments—are enaged in a type of "emotional surveillance." They are seeking to codify each user's political and consumer identity, to figure out how users feel, how they see themselves. In reaction, Go Rando adopts the strategy of obfuscation to disrupt the increasingly fine-grained data collection practices that enable this emotional surveillance. While unlikely, if everyone started using Go Rando tomorrow, it could have broad collective effects against state and corporate emotion profiling. But regardless, for any one user it provides individual benefits by disrupting Facebook's News Feed algorithm (and

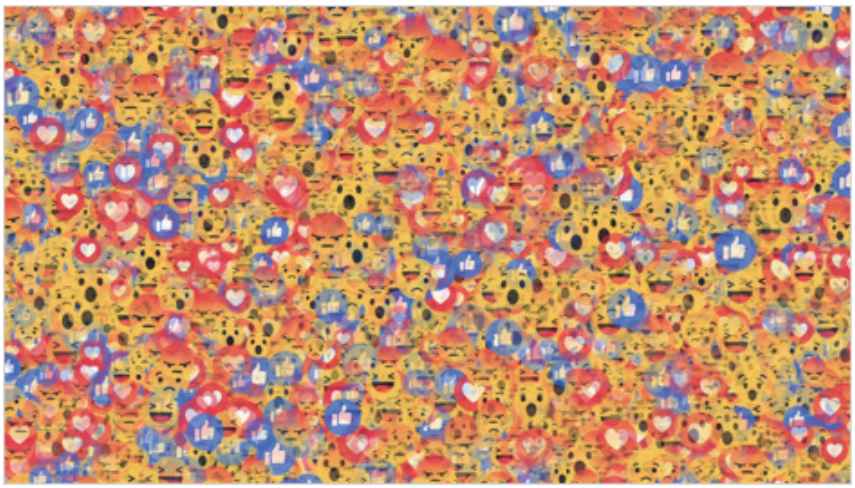


Figure 2: Go Rando fills your Facebook emotion profile with noise

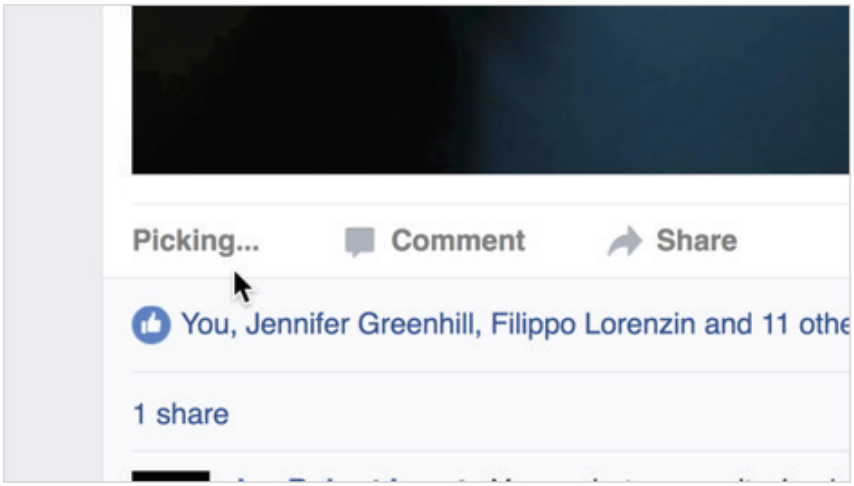


Figure 3: Go Rando randomly selects a Facebook "reaction" for you

thus, blunting the "filter bubble" effect), resisting the site's attempts at emotional manipulation, and confusing corporate and governmental surveillance.

Go Rando provokes users to question how Facebook's "reactions" are used. Who benefits when a user marks themself as "Angry" or "Sad" in response to a particular post? Which groups have the most to lose? And how might the uses of this data change the nature of privacy and democracy over the coming months or years?

Finally, when a user sees a discongruent or "inappropriate" reaction from their friends in the future, perhaps this might be a sign of a potential ally in the battle between individual freedom and the big data state-corporate machine that seeks to use our data against us.

BEN GROSSER creates interactive experiences, machines, and systems that examine the cultural, social, and political implications of software. Recent exhibition venues include Arebyte Gallery in London, Museu das Comunicações in Lisbon, Museum Ludwig in Cologne, and Galerie Charlot in Paris. His works have been featured in Wired, The Atlantic, The Guardian, The Washington Post, and Der Spiegel. The Chicago Tribune called him the "unrivaled king of ominous gibberish." Slate referred to his work as "creative civil disobedience in the digital age." Grosser's recognitions include First Prize in VIDA 16, and the Expanded Media Award for Network Culture from Stuttgarter Filmwinter. He is an assistant professor of new media and a faculty affiliate at the National Center for Supercomputing Applications, both at the University of Illinois at Urbana-Champaign.

HIDING DATA FLOWS WITH COVERT CHANNELS

SAUMYA DEBRAY and **JON STEPHENS**, University of Arizona

Secure information flow ensures the confidentiality and privacy of sensitive data from being publicly leaked or illegitimately being accessed by attackers. This is usually achieved by defining securely typed programming languages and/or run-time systems that keep track of (aka taint) sensitive data and prevent unwanted leaks [1]. Securely typed languages tag variables with different security labels and enforce some security controls at run-time, ensuring a secure flow of information. However, running a piece of code on the underlying system that facilitates the computation (including the hardware, operating system, etc.) incurs some side effects upon the system. These side effects can be exploited to infer information about the running code or the data used in the computation. These side effects are referred to as *side-channels*: they are not the primary channels of communication and are therefore easily overlooked. As a result, they are often not properly monitored to ensure the secure flow of information. As an example, the running time of a piece of code or the amount of power it consumes can be considered side-channels. Previous successful attacks using side-channels show that ignoring side-channels imposes significant security risks in protecting sensitive data.

On the other hand, current complex run-time systems, such as interpretive systems including Java/Python or even operating systems, involve many complicated and large components that actively interact with each other. For instance, an interpreter is usually equipped with a JIT compiler, which, as it will be shown later, can be affected by the input program in a way that makes it possible to exfiltrate information. Other parts of the run-time system, such as the thread library or garbage collector, can also potentially leak information. Keeping track of all the side-effects caused by input program/data is infeasible, in part due to the complexity and size of the whole run-time system. Even if the overhead of doing so is acceptable, it leads to too many false positives since normal programs also tend to cause similar side-effects. Furthermore, an attacker is not bound to use only one of the mentioned side-channels—multiple side-channels can be utilized where each side-channel contributes to leaking a bit of information. In such a case, combining all the bits from multiple channels provides the full-fledged side-channel-based obfuscation.

Here we briefly discuss some of the available challenges and opportunities in exploiting the side effects of large run-time systems in order to gain access to and propagate protected information. We will also briefly discuss the difficulties of coping with these types of information leaks due to the complexity and the wide domain where these side-channels can be utilized.

SIDE-CHANNEL-BASED OBFUSCATIONS

There are numerous ways to construct side-channel based obfuscations, but we will target interpretive systems. The attack is based on run-time JIT compilation, which is common in modern interpreters. While interpretive systems are specifically being targeted, the idea is generally applicable to run-time systems. For instance, the example can be adapted to use the CPU's data cache or the kernel's file cache rather than the JIT compiler.

In this work, we use time-based side-channels in run-time systems where the behavior of the execution is affected by the inputs. Most interpretive systems, such as Java, use a JIT compiler to optimize parts of the code that tend to take much of the execution time. With JIT compilation enabled, if a function is executed for a number of times greater than a threshold, JIT compiler compiles the function into machine code that tends to execute faster than the interpreted version. The following code in Python (Note: CPython implementation does not come with a JIT compiler but there are other versions that do) uses a time-based obfuscation technique to take advantage of the run-time JIT compilation to propagate a value.

```
#Foo is a list of k functions, each
  taking T(ms) to execute
Secret = key
Public = 0
for i in range(k):
   c = (Secret&(1 << k))*LARGE
   while c:
            Foo[k]()
            c = c-1
   s = time.time()
   Foo[k]()
   Public = Public | (time.time()-s)/T
```

The idea is to force the JIT compiler to compile a function based on the value that needs to be propagated (k bits of the variable *Secret* in the code sample). Assuming the JIT-compiled and the not-JIT-compiled versions of the function take different amounts of time to execute, it is possible to infer some information about the data. Depending on the bit that needs to be copied, we can force JIT compilation of a function. Measuring the time needed to execute the function tells us whether it has been JIT-compiled or not, and hence reveals the value of the bit.

Function *Foo[k]()* takes more than *T* milliseconds to run if not JIT-compiled. The *T* is computed such that *Foo[k]* takes more than *T* to finish if interpreted and less than *T* if JIT-compiled (*T* can be computed at run-time by choosing a value slightly less than the running time of *Foo[k]*). Line 5 assigns a large number to *c* if the *k*th bit of *Secret* is one and zero otherwise. This forces the JIT compiler to compile *Foo* only if the *k*th bit in *Secret* is one meaning that by later executing *Foo[k]()* and measuring the time taken to execute *Foo*, it is possible to infer the bit was zero or one. To leak *k* bits of information from *Secret*, it is only needed to repeat the process using *k* functions, one for each of the *k* bits.

SUMMARY

Traditionally, side-channels of information refer to the data flows derived from non-primary channels of communication (e.g., input/outputs). Carefully observing some (external) characteristics of a run-time system, such as measuring run-times and/or consumed energy, can provide a powerful channel of information that can leak data or the algorithm used in the computation.

We discuss the idea of side-channels in a

broader context. Specially, wide-spread use of complex run-time systems, such as interpretive systems, opens up different attack vectors to gain unauthorized access to confidential data. Moreover, wide-range of possible side-channel sources and the complexity of such run-time systems, makes the problem of preventing these types of attacks specially challenging.

REFERENCES

[1] Sabelfeld, Andrei, and Andrew C. Myers. "Language-based information-flow security." *IEEE Journal on selected areas in communications* 21, no. 1 (2003): 5-19.

SAUMYA DEBRAY is Professor of Computer Science at the University of Arizona, Tucson. His research interests involve various aspects of automatic program analysis, focusing in particular on code armored using various static and dynamic obfuscations and anti-analysis defenses. When he isn't playing with code, he enjoys hiking and backpacking.

JON STEPHENS is a first year master's student in the University of Arizona's Computer Science program. There, together with his good friend Jimmy'); DROP TABLE Users;-- and his advisor Dr. Saumya Debray, Jon performs research into program/malware analysis and obfuscation. Despite his status as a new graduate student, Jon uncovered the key to obfuscation, which is

SOFTWARE DIVERSIFICATION AS AN OBFUSCATION

NICOLAS HARRAND and **BENOIT BAUDRY**, Inria, France

CONTEXT: WORMS AND BOTNETS

Along the ever increasing number of devices connected to the Internet comes an everlasting plague: computer worms. From the notorious Conficker that may have infected from 7 up to 25 millions computer running Microsoft Windows [1], to the more recent Mirai and Bashlite targeting IoT devices such as DVRs or IP cameras running Linux, they share a common operating pattern: they are able to penetrate the victim system and use it to spread to new targets. Resulting botnets can cause all sorts of troubles from world-wide spamming campaigns to massive DDoS attacks (a Mirai botnet is thought to be responsible for the latest DDoS on Dyn causing an array of sites, including Twitter, Amazon, Tumblr, Reddit, Spotify and Netflix, to be partly unreachable on October 16th 2016 [2]).

Ultimately, this malware can thrive on the Internet because it presents a combination of two dangerous properties:

- Despite the large a number of nodes, there is a strong **software mono-culture**: for a given purpose, only a few different program are used by most nodes (e.g., Wordpress or Drupal for content management). This property is equally true for the environment of these programs. Indeed only a few OSs (Linux for servers, Windows for PCs, Android for mobile and embedded devices) cover most devices. The risk of this monoculture is as follows: a hacker who finds an exploit for one these programs can potentially infect all machines on which it is installed.

- It is a huge network made up of **billions of interconnected devices**. Consequently, all machines on which the same program is installed are connected to each other and an attacker who succeeds in accessing one of them can access all the others.

In the end, if a program offers an exploitable flaw (i.e. buffer overflow), it will be shared by many nodes providing entry points on each of them. Furthermore, as these nodes run in a similar environment, a unique malware may exploit their resources (i.e. shells or download tools such as wget) in order to replicate and spread over the network (see figure 1).

ON THE DANGER OF SOFTWARE MONO-CULTURE

In a network of interconnected nodes, the presence of one vulnerability on one node is not the most serious issue. The threat comes the fact that huge clusters of connected nodes are hosting the same vulnerabilities.

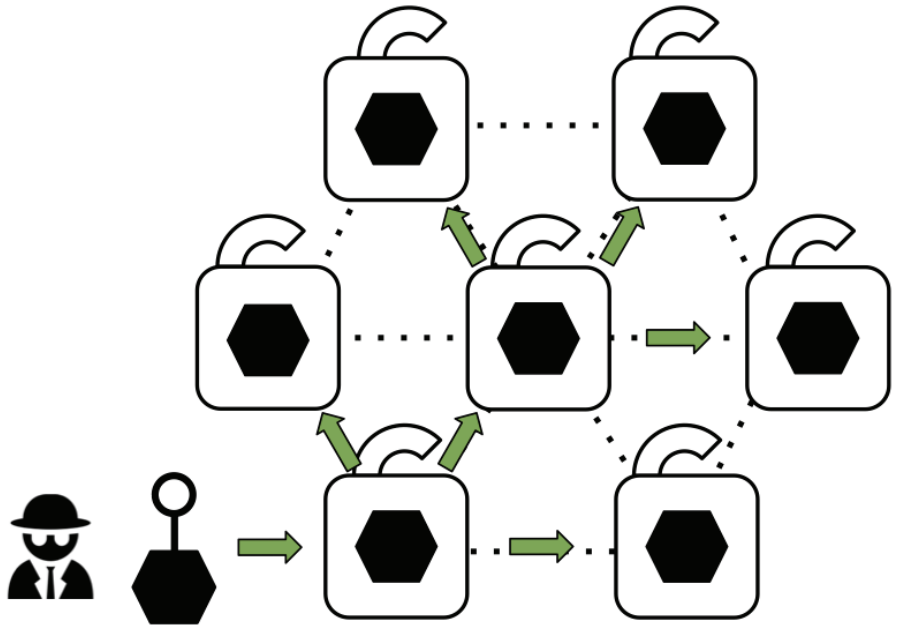


Figure 1. Software monoculture: an essential source of knowledge for malicious users

The similarity among the nodes in the network is the way to **obtain information** on the node. This lack of diversity allows an opponent to test and **predict** the behavior of its attack offline, because having the information on one node means having the information on each and every similar node. This decreases the cost of an attack.

In addition, being able to hack one node means in fact being able to compromise all the nodes that share the flaw. The benefits of such an attack increase with the popularity of the program infected by the vulnerability.

The lack of software diversity is both how and why a node is attacked.

DIVERSITY AND OBFUSCATION

Software diversity refers to the fact that multiple variants of a program can offer the same functionality through different implementations. It can appear naturally, e.g., when different persons choose to implement software for similar purposes (e.g., most languages have their maths library, some even have several), or it can be synthesized through automatic transformations [3]. Automatic diversification can occur at multiple stages of the chain: at the source code level [4], at compile time [5, 6], at load time, or even at run time [7].

Diversification represents a form of obfuscation not because it makes it hard to predict the behavior of one variant, but rather because observing one variant does not leak information about other variants. Therefore it mitigates the capacity to gain information on the node making it a moving target [7] (see figure 2). Software diversification could greatly increase the cost of an attack, as it would have to be either tailored to each variant or adaptable enough to ignore their differences. Furthermore it may cripple the propagation of worms unable to penetrate the system that present different flaw, and prevent from using at will available resources.

Yet, automatic software diversification raises many challenges. In particular, there is a lack of appropriate metrics to evaluate diversity and its efficiency in achieving the desired properties. In

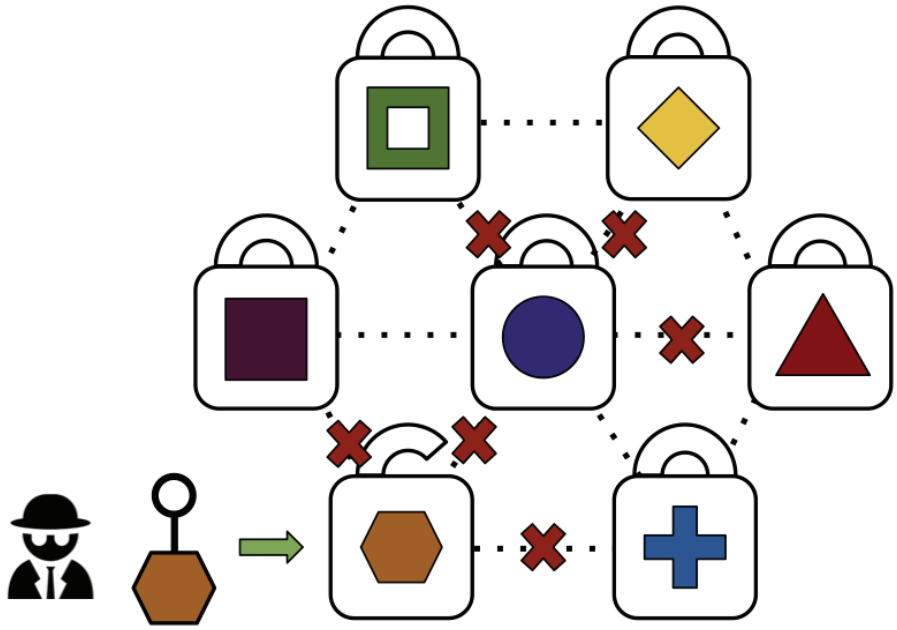


Figure 2. Software diversity: an approach for global obfuscation and system-wide protection

this perspective, our work addresses the following aspects: (i) measure actual diversity through number of diverging code blocks, instructions, or execution traces, (ii) and directly test known attacks on a vulnerable program and compare the results with its variants. Furthermore depending on the goals of diversification, all software regions are not equally interesting to diversify, and independently of the goals, all regions do not offer the same opportunities for diversification. The characterization of such software regions is essential for the success of massive diversification.

HOW TO EVALUATE THE EFFICIENCY OF DIVERSITY?

Determining the effectiveness of diversity at avoiding widespread vulnerabilities is a major challenge in the current state of the art. Given a set of N variants of a program, a set of attacks, we currently investigate the following metric. If we cluster variants based on their vulnerability or non vulnerability against each attack: $\mathbb{V} = \bigcup V_i$, where V_i is the set of variants vulnerable to the attack i. Its cardinality is noted $|\mathbb{V}|$.

The distance between two clusters is defined as follows:

$$d : \mathbb{V}^2 \mapsto [0,1], d(V_i, V_j) = 1 - \frac{|V_i \cap V_j|}{|V_i \cup V_j|}$$

This distance can be seen as the ratio of variants that differ from one group to the other.

The quadratic entropy $H(\mathbb{V})$ is defined as follows:

$$H(\mathbb{V}) = \sum_{i=0}^{|\mathbb{V}|} \sum_{j=0}^{|\mathbb{V}|} P_i P_j d(V_i, V_j)$$

where $P_i = \frac{|V_i|}{N}$ is the probability that a variant belongs to V_i. This quadratic entropy based on Rao's quadratic entropy [8], can be seen as the measure of the average spread of the set of attacks on our set of variants.

Worst case scenario: each variant is vulnerable to each attack. (That is the case if we have only one variant). The distance between every two given groups is null because all of our groups overlap

exactly, therefore $H(\mathbb{V}) = 0$.

Best case scenario: $\mathbb{V}$ partitions the set of variants, meaning every group is disjoined, and each group has the same size, and then $H(\mathbb{V}) = 1$. Note that some variants may not have any vulnerability and do not belong to any group.

From there, we can see that our goal is to maximize entropy on a set of variants grouped by vulnerabilities or, said in a different way, to minimize the information on one variant leaked by another. The more entropy a diversification scheme can generate the more efficient it is to achieve our fixed goal of preventing widespread vulnerabilities.

FEEDBACK FROM THE WORKSHOP

We presented this work at the International Workshop on Obfuscation in April 2017. The audience acknowledged the relevance of diversity for obfuscation and also provided extremely useful feedback about the challenges to actually deploy diverse versions of a program.

One challenge is about the transparency of the code and the build pipeline. Several open source communities aim at building secure programs by having a completely transparent build process that can be understood and analyzed by everyone. This transparency is necessary to ensure true open source. The integration of automatic diversification in such processes can be challenging to keep the transparency while preserving the benefits of diversity.

Software maintenance in the presence of diversity was mentioned as another challenge of this approach for obfuscation.

REFERENCES

[1] Seungwon Shin and Guofei Gu. Conficker and beyond: A large-scale empirical study. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 151–160, New York, NY, USA, 2010. ACM.

[2] See: Hacked Cameras, DVRs Powered Today's Massive Internet Outage. Krebs on Security. Oct 16, 2016. https://krebsonsecurity.com/2016/10/hacked-cameras-dvrs-powered-todays-massive-Internet-outage/.

[3] Benoit Baudry and Martin Monperrus. The multiple facets of software diversity: Recent developments in year 2000 and beyond. *ACM Computing Survey*, 48(1):16:1–16:26, 2015.

[4] Benoit Baudry, Simon Allier, and Martin Monperrus. Tailored source code transformations to synthesize computationally diverse program variants. *CoRR*, abs/1401.7635, 2014.

[5] S. Forrest, A. Somayaji, and D. Ackley. Building diverse computer systems. In *Proceedings of the 6th Workshop on Hot Topics in Operating Systems (HotOS-VI)*, HOTOS '97, pages 67–, Washington, DC, USA, 1997. IEEE Computer Society.

[6] Frederick B. Cohen. Operating system protection through program evolution. *Computers & Security*, 12(6):565—584, 1993.

[7] Hamed Okhravi, Thomas Hobson, David Bigelow, and William Streilein. Finding focus in the blur of moving-target techniques. *Security Privacy, IEEE*, 12(2):16–26, Mar 2014.

[8] C Radhakrishna Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1):24–43, 1982.

BENOIT BAUDRY is a research scientist at INRIA, France. His research is in the field of software engineering, focusing on program analysis, testing and diversification. In his recent work, he has been investigating ways to exploit software diversity to obfuscate browser fingerprints, and to synthesize large quantities of program variants. These works are part of interdisciplinary collaborations with ecologists, lawyers and artists.

SOFTWARE (DE-)OBFUSCATION: HOW GOOD IS IT?

ALEXANDER PRETSCHNER, Technische Universität München, Germany
joint work with Sebastian Banescu

Software obfuscation aims at hiding data, code, or logic. Examples for obfuscating data include license or cryptographic keys. Code is obfuscated in order to avoid the easy detection and subsequent disabling of license checks, or runtime integrity checkers that verify if the code has been tampered with. Finally, algorithms or logic often constitute IP that its owner does not want to publicly divulge. There is a plethora of software obfuscation techniques that are described in detail in a forthcoming article [BP17].

At first glance, one may be tempted to argue that the problem can simply be solved by encrypting relevant pieces of (machine) code and data. This, however, does not solve the problem because in order to be executed or used, code or data need to be decrypted. This requires their existence as plaintext at some moment in time—at which the code or the data can be read by an attacker. Moreover, this leads to the problem of protecting the key itself, which requires a root of trust. These roots of trust can be implemented in software (e.g., white box crypto) or hardware (e.g., Intel SGX or TPM) which both suffer from various disadvantages that we cannot discuss here.

The goal of software obfuscation is to defend against so-called Man-At-The-End (MATE) attackers. These are attackers that have access to code in binary or source code format and, in principle, can make use of unlimited resources in order to de-obfuscate a piece of software. In practice, of course, resources are not unlimited, and the effort that attackers invest will depend on the anticipated gains of doing so.

The "effort" to de-obfuscate a piece of software is hard to characterize. Intuitively, one would expect it to strongly correlate with the "quality" of the obfuscation strategy that was applied which means that the quality of obfuscation and de-obfuscation are dual concepts. Collberg, Thomborson and Low [CTL97] suggest considering two aspects of the quality of obfuscation: potency and resilience. Potency is the property of an obfuscated piece of software to resist a human attacker to de-obfuscate it. Resilience, in contrast, characterizes the resistance with respect to automated attackers. Collberg et al. understood that resistance is relative to the resources (that is, cost) that an attacker is willing to spend on a de-obfuscation attack.

Unfortunately, it is not clear how to characterize the power of an obfuscation transformation in a practical way. There are several essentially probabilistic, rarely complexity-theoretic,

characterizations that give rise to the beginning of a theory of obfuscation. However, these rather theoretical characterizations tend to be of limited value to a practitioner because they do not state, for a given obfuscator, precisely how potent or resilient it is.

Interestingly, the situation is similar for the strength of cryptographic encryption. There are a few studies that estimate how long it will take to break a key of a specified length, but we are, again from a practitioner's rather than a theoretician's perspective, not aware of hard lower bounds on the effort needed by a clever attacker to break a specific cipher. (One may indeed see cryptography as a special case of obfuscation: Given (1) an original artifact—code or data in the case of obfuscation, and plaintext data in the case of cryptography—and (2) a set of parameter—a set of transformations, their ordering, and transformation-specific parameters in the case of obfuscation; and the cryptographic key in the case of crypto—obfuscation and encryption transformations yield obfuscated and encrypted artifacts. De-obfuscation and decryption are the respective inverse transformations).

In terms of potency, it seems naturally hard to make solid statements about the quality of the respective obfuscation transformation—because human ingenuity is hard to predict. When considering fixed automated attacker models, the situation slightly improves. It is then possible to apply a specific automated attacker on a sample set of de-obfuscation problems and use machine learning technology to build respective prediction models. In recent work [BCP17], we have shown that it is possible, in the context of the study presented in the paper, to predict the time for attacks that are based on symbolic execution technology [BCG+16] with an accuracy of >90% for >80% of the considered programs. An obvious observation is that this study suffers from several threats to external validity, most notably the dependence on the specific attack technology (symbolic execution) that we use. However, this seems to be the nature of the beast, and we do see hope that there may only a limited number of realistic automated attacks that could, and should, be studied in a way similar to our work.

In the future, we consider it of utmost importance to technically characterize the de-obfuscation problem, probably as a search problem, and use this characterization as a basis for a systematic study, and understanding, of the "quality" of obfuscation. We believe that one of the most pressing questions in software obfuscation is the question for which there are only very partial answers today: from a practitioner's perspective, how good is obfuscation? Can we provide estimates of the cost (the resources needed by the attacker) to de-obfuscate in a way that is similar to physical safes that are assessed by the time that an attacker with a standard set of tools needs to break it [UL11]?

Finally, the title of this brief statement deliberately plays with the meaning of "good." So far, we have considered obfuscation technology to be "good" if it raises the bar for the attacker, and have equated quality with effort for an automated attack to de-obfuscate. A second relevant meaning of "good" pertains to the moral, or ethical dimension. Hiding code or data may not be considered "good" by the proponents of open source software or by opponents of digital rights management technology, and it certainly is not good if the technology is used by malware to evade detection.

REFERENCES

[BCG+16] Banescu, Collberg, Ganesh, Newsham, Pretschner: Code Obfuscation Against Symbolic Execution Attacks. Proc. ACSAC, pp. 189-200, 2016

[BCP17] Banescu, Collberg, Pretschner: Predicting the Resilience of Obfuscated Code Against Symbolic Execution Attacks via Machine Learning. Proc. USENIX, to appear, 2017

[BP17] Banescu, Pretschner: A Tutorial on Software Obfuscation. Advances in Computing, to appear, 2017

[CTL97] Collberg, Thomborson, Low: A Taxonomy of Obfuscation Transformations. Technical Report #148, The University of Auckland, 1997

[UL11] UL 687 Standard for Burglary-Resistant Safes. Underwriters Laboratories, 2011. https://standardscatalog.ul.com/standards/en/standard_687

ALEXANDER PRETSCHNER is a full professor for software engineering at the Technical University of Munich, scientific director at the fortiss institute for research and technology transfer for software-intensive systems and services, and director of the Munich Center for Internet Research. Research interests include all aspects of software systems engineering, specifically testing and information security. Prior appointments include a full professorship at Karlsruhe Institute of Technology, an adjunct associate professorship at the Techincal Unviersity of Kaiserslautern, a group management position at the Fraunhofer Institute for Experimental Software Engineering in Kaiserslautern, and a senior research associate position at ETH Zurich. PhD in Computer Science from the Technical University of Munich. Alexander is recipient of a Google Focused Research Award, two IBM Faculty Awards, a Fulbright scholarship and various best paper awards.

ON MISSING DATASETS

MIMI ONUOHA

To talk about obfuscation is to talk about negotiations of access to data. But what about cases where the presence of data itself is a luxury? These spaces offer new viewpoints for considering many of the same themes—of surveillance, privacy, power, and access—that obfuscation is concerned with. Because it is in these arenas that much of my work is situated, I have my own term for these spaces where omitted data live: missing datasets.

More specifically, "missing datasets" refer to the blank spots that exist in spaces that are otherwise data-saturated. My interest in them stems from the observation that within many spaces where large amounts of data are collected, there are often correlating empty spaces where data are not being collected.

The word "missing" is inherently normative, it implies both a lack and an ought: something does not exist, but it should. That which should be somewhere is not in its expected place; an established system is disrupted by distinct absence. These absences are significant, for that which we ignore reveals just as much (if not more) than what we give our attention to. It's in these things that we find cultural and colloquial clues to what is deemed important. Spots that we've left blank reveal our hidden biases and indifferences.

In addition, by paying attention to missing datasets we are able examine the wider culture of data gathering; in a world in which data collection is routine, explicit, and the de facto business model for an increasing number of industries, missing datasets force us to consider the spaces that remain removed from this emergent value system.

WHY ARE THEY MISSING?

Below I present four reasons, accompanied by real-world examples, for why a data set that seems like it should exist might not. Though these are not exhaustive, each reveals the quiet complications inherent within data collection.

1. Those who have the resources to collect data lack the incentive.

Police brutality towards civilians in the United States provides a powerful example of this maxim. Though incarceration and crime are among the most data-driven areas of public policy, traditionally there has been little history of standardized and rigorous data collection around brutality in policing.

Recently, public interest campaigns like Fatal Encounters and the Guardian's The Counted have filled this void, and today there is data about the issue [1, 2]. But the fact remains that a task that is arduous and time-consuming for these individuals and organizations would be relatively easy for the law enforcement agents who are most closely tied to the creation of the dataset in the first place—they merely lack any significant incentive to gather it.

2. The data to be collected resist simple

quantification (corollary: we prioritize collecting things that fit our modes of collection).

The defining tension of data collection is the challenge of defining a messy, organic world in formats that are structured. This complication is magnified for information that is difficult to collect by nature of its very form. For instance, since there's no reason for other countries to monitor US currency within their countries, and the very nature of cash and the anonymity it affords renders it difficult to track, we don't know how much US currency is outside of the country's borders [3].

Other subjects resist quantification entirely. Things like emotions are hard to quantify (at this time, at least). Institutional racism is similarly subtle; it often reveals itself more in effect and results than in deliberate acts of malevolence. Not all things are quantifiable, and at times the very desire to render the world more abstract, trackable, and machine-readable is an idea that itself should invite examination.

3. The act of collection involves more work than the benefit the data is perceived to give.

Sexual assault and harassment are woefully underreported [4]. While there are many reasons informing this reality, one major one could be that in many cases the very act of reporting sexual assault is an intensive and painful process. For some, the benefit of reporting isn't perceived to be equal or greater than the cost of the process.

4. There are advantages to nonexistence.

To collect, record, and archive aspects of the world is an intentional act. As the concept of obfuscation illustrates, there are situations in which it can be advantageous for a group to remain outside of the oft-narrow bounds of collection. In short, sometimes a missing dataset can function as a form of protection, such as in the case of sanctuary cities deleting identifying data related to undocumented immigrants [5].

FINAL THOUGHTS

It is important not to interpret the highlighting of missing datasets as a direct call or invocation to fill these gaps. Rather, the topic lends itself to specific and general considerations of our wider system of data collection.

If we begin from the understanding that there will always be data missing from any collection system, we allow ourselves the space to address resulting patterns of inclusion and exclusion.

For more examples of missing datasets, please visit my GitHub repository [6], and essays on Quartz [7] and Data & Society's Points blog [8].

REFERENCES

[1] "Fatal Encounters," *Fatal Encounters*, accessed May 22, 2017, http://www.fatalencounters.org/.

[2] "The Counted: people killed by police in the US," *The Guardian* (New York City), https://www.theguardian.com/us-news/ng-interactive/2015/jun/01/the-counted-police-killings-us-database.

[3] United States, Federal Reserve Board, The Federal Reserve Board, by Ruth Judson, November 2012, accessed May 22, 2017, https://www.federalreserve.gov/pubs/ifdp/2012/1058/default.htm.

[4] United States, Federal Reserve Board, The Federal Reserve Board, by Ruth Judson, November 2012, accessed May 22, 2017, https://www.federalreserve.gov/pubs/ifdp/2012/1058/default.htm.

[5] Colin Lecher, "NYC will stop retaining data that could identify immigrants under Trump administration," *The Verge*, December 2016, accessed May 22, 2017, https://www.theverge.com/2016/12/8/13882676/new-york-idnyc-trump-de-blasio-data-new-policy.

[6] Mimi Onuoha, "On Missing Datasets," GitHub, accessed September 29, 2017 https://github.com/MimiOnuoha/missing-datasets.

[7] Mimi Onuoha, "Broadway won't document its dramatic race problem, so a group of actors spent five years quietly gathering this data themselves," *Quartz*, December 4, 2016, accessed September 29, 2017, https://qz.com/842610/broadways-race-problem-is-unmasked-by-data-but-the-theater-industry-is-still-stuck-in-neutral/.

[8] Mimi Onuoha, "The Point of Collection," *Points (Data & Society)*, February 10, 2016, accessed September 29, 2017, https://points.datasociety.net/the-point-of-collection-8ee44ad7c2fa.

MIMI ONUOHA is a Brooklyn-based artist and researcher using code and writing to explore the process, results, and implications of data collection. Recently she has taught at NYU's Interactive Telecommunications Program and been in residence at the Royal College of Art and Data & Society Research Institute. Currently she writes for Quartz's Things Team and is a Research Resident at Eyebeam, where she is investigating data collection, missing datasets, and strategies for intervention and response.

OBFUSCATING 15M US CRIMINAL RECORDS AND MUGSHOTS FOR THE RIGHT TO REMOVE THEM

PAOLO CIRIO

ABOUT THE ARTWORK

Control and access to information, the right to privacy, mass surveillance and profiling, and the system of participation within social dynamics are explored in this socio-critical Internet artwork. Ultimately, the project questions the legal frameworks surrounding public policies on privacy and profiling of citizens and engages the public in a debate about them.

By engaging with law, millions of individuals, bad business practices, and general public opinion, Obsurity seeks to embody a practical discourse about the aesthetics, functions, and ethics of information systems affecting social structures that resonates within and outside the contemporary art dialogue.

This artwork is made with millions of mugshots, obtained through a screen-scraping software, from websites such as Mugshots.com, Usinq.com, Justmugshots.com, MugshotsOnline.com, etc. This project has cloned these mugshot websites using similar domain names and then shuffled the data associated with the individuals listed to obfuscate their identities.

The algorithm created for obfuscating the data makes sure that an individual's name and picture are never associated with the actual person arrested. It scans for individuals with a common gender, age, race, and location and shuffles their first and last names along with their respective mugshots, while maintaining accurate all the other details about the individuals, including charges and the location of the arrest.

Then the algorithm republishes this data on the open web using search engine optimization (SEO) techniques to boost the search rankings of the cloned websites and promote the version with the scrambled criminal records. The republished obfuscated data maintains the layout and watermarks of the original mugshots, and by using similar domain names the project would effectively interfere with the activity, reputation, and business of mug-shot websites.

ABOUT THE PUBLICATION OF MUGSHOTS IN THE U.S.

Mugmshot websites have been exposing tragic photos of people who have been arrested regardless of the amount of time spent in jail, often just for minor offenses, or even if they were later found to be innocent or the charges against them had been dropped.

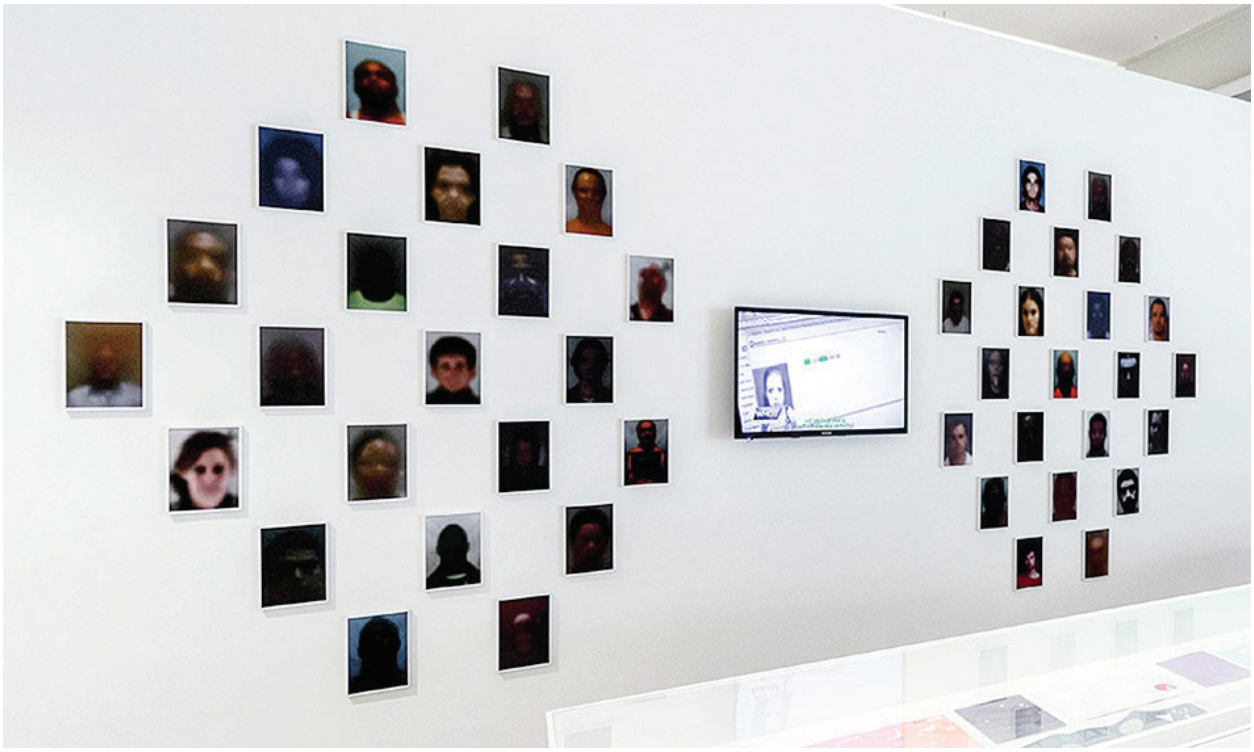


Figure 1: Obscurity installation view

These websites are designed to embarrass and shame since searchable booking photos can effectively ruin someone's reputation with social stigmas and attendant prejudgments in their communities, families, and workplaces, especially when they are seeking employment, obtaining insurance or credit. These mugshots are often of the most vulnerable members of society: victims of mass incarceration, economic inequality, and racial discrimination, along with those who lack treatment for mental illness, in a country with poor welfare policies, coupled with a severe criminal justice system and unforgiving law enforcement agencies.

The online mugshots problem came to public attention in 2013. Since then, the number of these websites has multiplied, and existing websites continuously change their brand name in order to keep collecting and monetizing on mugshots. The United States has the highest rate of imprisonment in the world. Every year in the U.S., city and county jails across the country admit between 11 and 13 million people. Approximately 2.2 million people are currently locked up in the U.S. Of those in jail, 60 percent haven't been convicted of anything.[1] Of the non-convicted defendants behind bars, 75% of them are held on non-violent offenses.[2]

The publication of booking photos online is legal under the freedom of information and transparency laws in most U.S. states. Furthermore, many freedom of press organizations and legislators[3] have been opposing bills that would regulate the publication of mugshots. Already many federal bills related to mugshots have failed or have been pending for years.[4]

Mug-shot websites monetize by placing advertising for reputation management services alongside booking data, and they often charge a picture removal fee. The controversial request

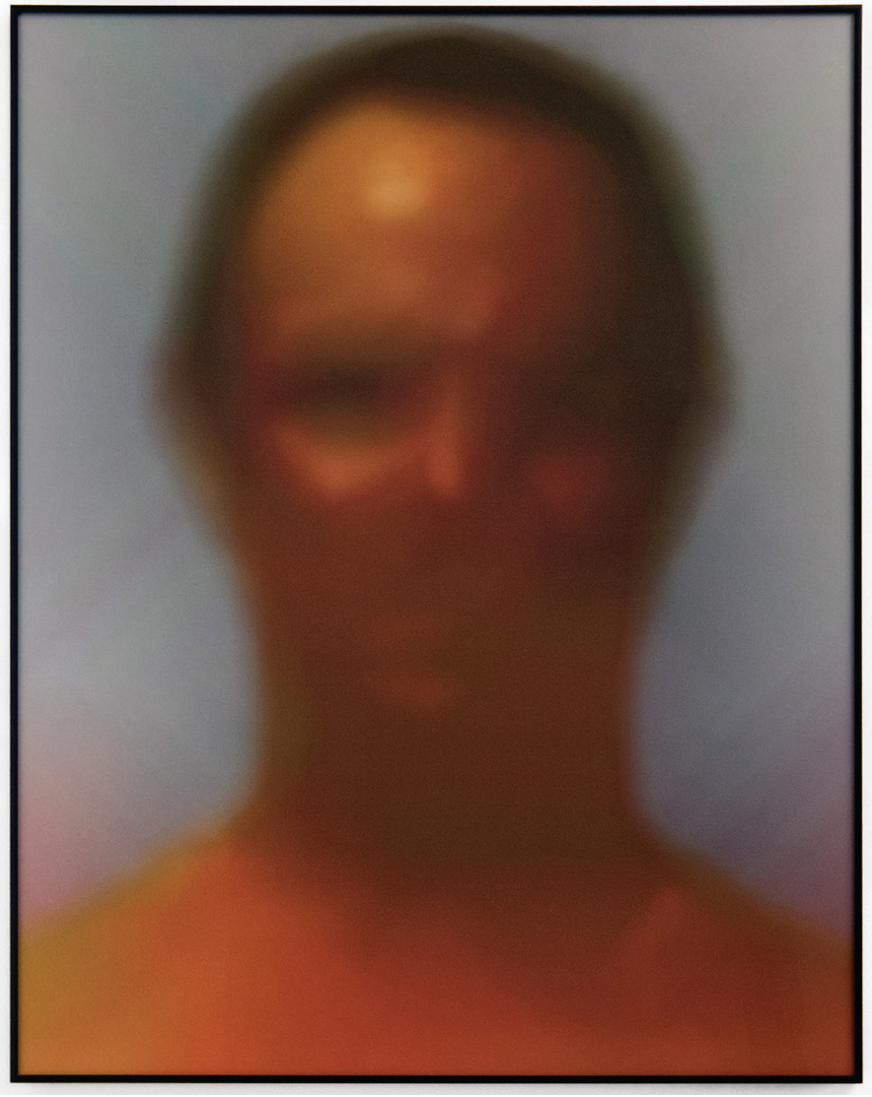


Figure 2: Obscurity detail

for payment to remove mugshots has led some state legislatures to propose bills to regulate the industry to comply with the right-of-publicity statute, which gives individuals some control over the commercial use of their name and likeness.[5] They also violate extortion laws since removal and online reputation services can charge thousands of dollars[6] to remove mugshots. However, some mug-shot websites operate in offshore jurisdictions and their owners are in hiding [7], which makes difficult any kind of legal and effective action to remove the mugshots and crack down on extortionary practices such as demanding a fee to stop their dissemination online. And yet, the simple publication of mugshots across the Internet still hasn't been regulated nationwide in the U.S.

Search engines such as Google are complicit as they could do what no legislator could—demote mugshot sites and thus reduce, if not eliminate, their power to stigmatize. Commercial search engines collect and exploit as much information as possible in order to monetize on data analysis and brokerage and sell it to advertisers. Any information taken out from search results might affect a search engine's profit, and as such they will always oppose and slow down any removal

of personal information from their datasets. The core business model of most search engines that profile people is often unscrupulous by design.

ABOUT THE PARTICIPATION OF THE PUBLIC

A participatory element of the project allows anyone to both judge typical criminal case scenarios sampled from the database and send a complaint to search engines and mug-shot websites. The visitors of the cloned mugshot websites, as participants of the online artwork, are able to decide whether to report individual profiles or instead keep them public by opting between two buttons: "Keep it" or "Remove it," which automates the sending of removal requests of individual mugshots to search engines as complaints for the publishing of unethical content and activity.

The main goal of the project Obscurity is thus to report all the URLS of the mugshot websites and advocate for statewide regulation on the publication of court information. More specifically, Obscurity proposes to keep all the information on civil cases filed in courtrooms and in law enforcement offices on web platforms that require registration to ensure that only qualified professionals are able to access certain data.

REFERENCES

[1] Nick Pinto, "The Bail Trap," *New York Times*, August 13, 2015. http://www.nytimes.com/2015/08/16/magazine/the-bail-trap.html

[2] Chandra Bozelko, "The cash bail system should be eliminated rather than reformed," *Guardian*, February 5, 2016. http://www.theguardian.com/commentisfree/2016/feb/05/the-cash-bail-system-should-be-eliminated-rather-than-reformed

[3] "Mug shot publishing industry," *Wikipedia*, last modified January 14, 2016. https://en.wikipedia.org/wiki/Mug_shot_publishing_industry

[4] "Mug Shots and Booking Photo Websites," *National Conference of State Legislatures*, December 11, 2016. http://www.ncsl.org/research/telecommunications-and-information-technology/mugshots-and-booking-photo-websites.aspx

[5] "Lawsuit goes after 'extortion' mugshot websites," *WCPO Channel-9 Cincinnati*, February 13th, 2013 http://www.wcpo.com/news/local-news/lawsuit-goes-after-extortion-mugshot-websites

[6] David Kravets, "Mug-Shot Industry Will Dig Up Your Past, Charge You to Bury It Again," *Wired*, August 2, 20111. http://www.wired.com/2011/08/mugshots/

[7] Natasha Del Toro, Dan Lieberman, and Rachel Schallom, "The Digilantes Try to Find Out Who Is Behind Mugshots.com," *Fusion*, February 9, 2016. http://fusion.net/interactive/252451/digilantes-mugshots-dotcom-investigation/

PAOLO CIRIO engages with legal, economic and semiotic systems of the information society, such as privacy, copyright, democracy and economy. Because of his artistic provocations, Cirio has often been subject to investigations, legal and personal threats by governmental and military authorities, powerful multinationals and financial institutions, as well as crowds of ordinary people. His controversial artworks have unsettled institutions as such as Facebook, Amazon, Google, VISA, Pearson, Cayman Islands and NATO, and are often covered by global media outlets, such as CNN, Fox News, Washington Post, Der Spiegel, and El Pais, among others. He has had solo shows at International Kunstverein Luxemburg, NOME gallery (Berlin), Bellegarde Centre Culturel (Toulouse), Kasa Gallery (Istanbul), Aksioma Institute for Contemporary Art (Ljubljana), and he has won a number of awards, including Golden Nica first prize at Ars Electronica, Transmediale second prize and the Eyebeam Fellowship, among others.

HYPERFACE: EMERGING STRATEGIES FOR OBFUSCATING COMPUTER VISION ALGORITHMS

ADAM HARVEY

HyperFace is a new type of camouflage to obfuscate computer vision algorithms. It is being designed to decrease the efficiency and accuracy of automated facial recognition. As an obfuscation strategy HyperFace's primary goal is to introduce face-like noise into the visual wavelength signal domain by displaying maximally activated false-face regions that are not perceivable as faces to a human observer. As a countersurveillance item, HyperFace can be used as a decoy in combination with CV Dazzle to allow the wearer's true face to become hidden in the background of higher-confidence face scores.

In any automated facial recognition system, the first step is to isolate the facial region using a face detection algorithm. The three most widely used approaches for 2D face detection in the

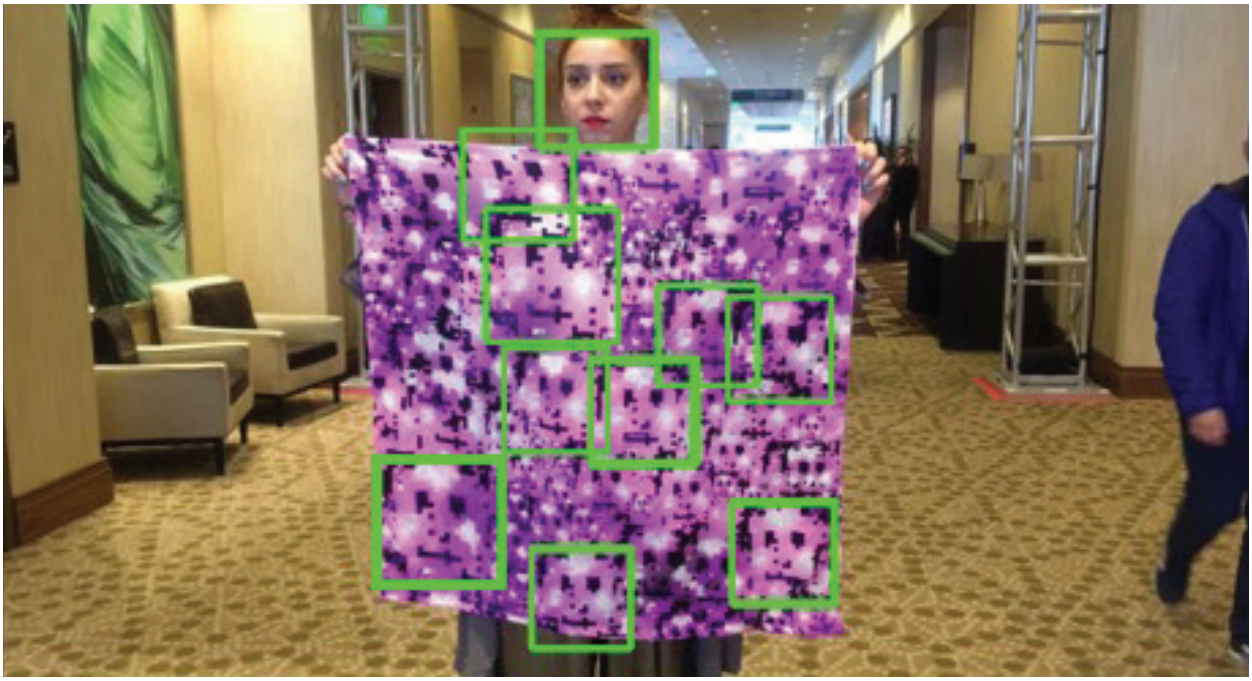


Figure 1: HyperFace demonstration by Hypen-Labs. Face regions detected using haarcascade frontal face default detector. March 14, 2017.

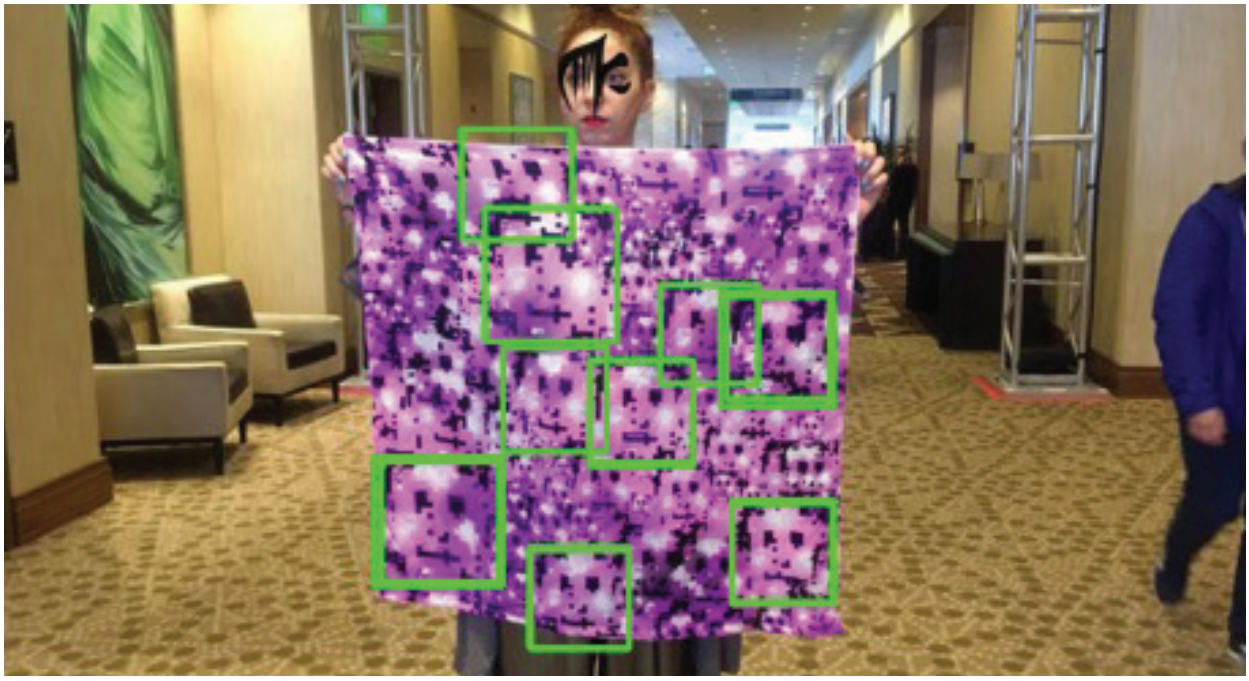


Figure 2: HyperFace demonstration by Hypen-Labs with superimposed CV Dazzle pattern. Face regions detected using same haarcascade frontal face default detector settings as in Figure 1.

visible spectrum are haarcascades, histogram of oriented gradients (HoG), and convolutional neural networks (CNN). The first iteration of the HyperFace project targets haarcascade detectors, which were chosen as starting point because of their ubiquity in consumer applications and previous research from CV Dazzle (2010). Future versions will target HoG and CNN detectors, which are much more effective at detecting faces in varying poses, illumination, expression, and skin tone. The first prototype from February 2017, shown below on a printed textile displayed by Hyphen Labs, has activated 9 false-positives.

In this example (Fig. 1) the true face is still detected and HyperFace only adds a small amount of computational excess for processing the image. But when combined with a strategy such as CV Dazzle (Fig. 2) the true face is now hidden and the face detection system is fooled.

In addition to the technical goal of the HyperFace project, it also aims to engage the public in dialogue about the potential risks of computer vision by using a non-technical and analog approach. HyperFace, as a fashion accessory, becomes accessible to groups outside of academic research communities. Already, the project has seen a large amount of interest online and offline from large media and fashion outlets.

In summary, the HyperFace project aims to increase the difficulty of automated face recognition systems by introducing noise, then using this noise source to provide a camouflaging background for the true face to escape detection.

ADAM HARVEY is an artist and independent researcher based in Berlin. His work includes developing camouflage from face detection (CV Dazzle, 2010), thermally reflective anti-drone garments (Stealth Wear, 2013), and a WiFi geolocation emulator (SKYLIFT, 2016). Harvey's multidisciplinary approach to exploiting surveillance technologies has appeared in a wide range of media from fashion magazines to a tweet from the Pentagon.

PLACE VS. SPACE: ON THE FUTURE OF LOCATION OBFUSCATION

SEDA GÜRSES, University of Leuven

Obfuscation has proven to be a successful strategy in preserving location privacy, and yet, we may have only scratched the surface of its potential in location services. Location privacy is a challenging topic that came to prominence with the introduction of location based services (LBS) (e.g., Beresford and Stajano 2003, Gruteser and Grunwald 2003, Wernke 2014). The objective of the research is to develop mechanisms that would protect individuals from other parties that may be able to "algorithmically discover a subject's whereabout and other information" throughout time (Krumm, 2009). Throughout the years, researchers have demonstrated that achieving location privacy is extremely hard and improvements to proposed mechanisms have come hand in hand with the development of mathematical foundations of obfuscation (e.g., Theodorakopoulos et al. 2014, Olteanu 2017). Building on these advances, and given the state of the art in Location Based Services, I would like to explore whether we can expand the scope of the research questions related to privacy and autonomy of individuals in space and time. To do so, I propose to revisit the common premises of location privacy research: its assumptions about space and time, its objectives and relationship to obfuscation.

Location privacy work builds on a specific understanding of space and time. Specifically, most of the models used in location privacy treat space as a kind of container that shapes people's movements and from which unwanted inferences can be made. When speaking about privacy risks, the authors often refer to inferences that can be made based on "proximity to an abortion clinic, crack house, AIDS clinic, business competitor, or political headquarters." Movement trajectories and behavioral patterns are captured too, typically represented as the sequence of locations associated with that individual. These trajectories are so unique, that researchers go as far as arguing "location is identity." Given this container model of space, technical strategies to ensure location privacy have followed two main paths. The first set of techniques aspire to conceal the identity of the person associated with the given location or movement (anonymity). The second set, of interest to us, protect location privacy by providing inaccurate, imprecise, or vague information concerning the location of that person (obfuscation) (Duckham and Kulik 2006).

The container model of space implicit to most location privacy research contrasts with more relational and dynamic understandings of space that conceive it as something that is constituted. In this latter framing, places are co-constructed in relation to the patterns of activity of people who use, observe or experience them. For

example, a park may be designated as such by the municipality, and marked as such on a map, but when it becomes the site of a political demonstration, what can be inferred from this location changes. In this more dynamic model, the relationship between space and people is bidirectional: a specific place may determine what are meaningful activities in that location, but people may also transform the meaning attached to that location with their actions. Once we accept the premise that space is not static and that people play a role in the making of a place, they can be treated as active geographic agents in the constitution of space in time.

In their work tracing the history of geodemographic systems, Phillips and Curry show that with the advances made in GPS and cellular technologies, Location Based Services have started working with this relational and dynamic conception of constituted space (Phillips and Curry, 2003). LBS systems are no longer only concerned with tracking individuals and building profiles, but also with leveraging this information to manipulate the behavior of their users to create "ideal" geographies. Most strikingly, based on the state of the art in 2003, they predict that "new [LBS] systems will potentially allow the instantaneous reconfiguring of spatial elements toward any emergent strategic end" (Phillips and Curry, 2003). The authors worry that harnessing such capabilities will accelerate the way in which the meaning of space gets negotiated and such negotiations will become increasingly invisible to their inhabitants.

Today, services like Waze, for sharing real time traffic and road info, Uber, for cab hailing, and Pokemon Go, an augmented reality game, are second nature to the billions of users of mobile devices. These services exemplify exactly the kind of LBS systems that Philips and Curry predicted over a decade ago. They gather location information based on which they make behavioral and spatial inferences. But, more importantly, they treat their users like active geographical agents that not only sense environments but can also be brought to co-create them based on notifications from LBS. The instantaneous feedback on how these users react to the notifications provided by these services are leveraged to devise experiments on how well their services create optimized geographies in line with their business interests while upholding a valuable user experience.

Exemplary of current day LBS, Waze, Uber and Pokemon Go also provide all actors involved in these systems new opportunities to apply obfuscation. Some of these techniques make news headlines. For example, when Waze rerouted cars avoiding freeway traffic jams to residential neighborhoods, residents turned to reporting road blocks in order to get the algorithm to re-route the cars elsewhere. They aspired to preserve the state of their neighborhood using obfuscation to change the availability that Waze attributed to their roads. Similarly, researchers have reported that they created hundreds of ghost Waze accounts to simulate a traffic jam, only to get Waze to reroute all other cars on the highway, sweeping the road empty for them to whiz through. Pokemon Go users have been reported to spoof their GPS to scoop Pokemon's where they are more densely available, not very different from app developers who want to test their new apps in different locations. The short history of Uber is cluttered with stories about the numerous ways in which the company developed systems to deceive its drivers, customers and authorities, obfuscating their actual prices, capacity and practices.

What is common to these stories of obfuscation

is that they all function on the premise of constituted space and its byproducts. In order to infer, manage and shape events in constituted space, LBS call other novel kinds of space into being. For example, to optimally manage space, LBS services providers are constantly in the process of modeling and converging upon an 'optimized space'. This optimized space is a reference object towards which all geographic agents are managed. In order to shape events, they may make use of indicators about a possible 'future space' to mobilize different geographic agents towards it. The interplay between optimized and future space is exemplified in the notifications that Uber sends to its drivers about an upcoming surge. That they initially included information about the price increase expected with the surge, and later removed this indicator, can be interpreted as the introduction of vagueness to gamify drivers into action in a future space given outcomes for an optimized space. When developers spoof GPS or use simulations to test their applications, one could argue that they are acting in "simulated space." Analysis of movements in physical space and simulations can be used to generate predictions about future space, develop optimized space, mobilize or gamify geographic agents, and to sort desirable and undesirable behaviors across these spaces.

Users now knowingly, or not, participate in these different spaces, and using obfuscation, can come to create yet other spaces. When Pokemon Go players spoof their GPS signals, one could argue that they enter a sort of "ghost space." Assuming the users spoof their location perfectly, to the game servers they may be located in Manhattan, to other players in Manhattan they are representatives of the increasing number of ghosts registered on their devices. Similarly, when Uber drivers synchronize to turn off their apps in order to stimulate a surge, one could argue that they generate a "resistance space." In constituted space, it seems that obfuscation acts on another level of abstraction. These users deliberately generate new spaces that are intended to escape or intentionally confuse the LBS capture mechanisms to identify, manage and shape events. The objective is not to obfuscate individual identities in a static space, but to obfuscate the many models of space captured in LBS systems.

So, what other questions can we pose in location privacy research if we shift from the container to the constituted model of space as a premise? LBS have moved from being systems intended to enrich a static understanding of space to dynamic systems of capture. It is true that if location privacy is in place, the inferences necessary for such LBS to work would be removed. However, we can still ask whether there are ways to use obfuscation to not conceal the static but dynamic aspects of geographic behavior. For example, would it be able to obfuscate that people are converging upon a political demonstration. In such a case, the obfuscation techniques may remain the same, but what is evaluated may include more than the effectiveness of the strategy in concealing an individual's movements.

We further see that the impact of location information leakage includes inferring information about people and their behaviors, but also extends to being able to leverage that information to optimize locations and behaviors. Researchers in location privacy have studied the impact of past, present and future locations on possible inferences. Moreover, those inferences and real-time observations of movements in space can co-exist and impact each other. Predictions are not only that, but can be used to nudge people to change their location, movement

and behavior in order to create ideal geographies. What does location privacy have to say about this practice of co-constructing space using past and predicted data: should we also study ways to do privacy preserving simulation? Would that be sufficient to deal with the different societal and individual autonomy concerns that may arise from such practices?

All of this may seem very complex to fit into the problem frame called location privacy. To shift assumptions about space and time that underlie an already challenging research question can easily be seen as daunting. However, some of these challenges may be seen as an opportunity to bring together communities that work on location privacy, malicious and deceptive behavior, and geography around a table. Once there, we may also discover many potential uses of obfuscation, good and bad, which may contribute greatly to the endeavors of researchers using this strategy in their work. If we do so, we may also make some necessary contributions to building systems that respect individuals and communities in space and time.

REFERENCES

Beresford, Alastair R., and Frank Stajano. "Location privacy in pervasive computing." IEEE Pervasive computing 2.1 (2003): 46-55.

Duckham, Matt, and Lars Kulik. "Location privacy and location-aware computing." Dynamic & mobile GIS: investigating change in space and time 3 (2006): 35-51.

Gruteser, Marco, and Dirk Grunwald. "Anonymous usage of location-based services through spatial and temporal cloaking." Proceedings of the 1st international conference on Mobile systems, applications and services. ACM, 2003.

Krumm, John. "A survey of computational location privacy." Personal and Ubiquitous Computing 13.6 (2009): 391-399.

Olteanu, Alexandra-Mihaela, et al. "Quantifying interdependent privacy risks with location data." IEEE Transactions on Mobile Computing 16.3 (2017): 829-842.

Phillips, David. and Michael Curry (2003) 'Privacy and the Phenetic Urge: Geodemographics and the Changing Spatiality of Local Practice', in David Lyon (ed.) Surveillance as Social Sorting: Privacy, Risk and Automated Discrimination, pp 137–152. London: Routledge.

Theodorakopoulos, George, et al. "Prolonging the hide-and-seek game: Optimal trajectory privacy for location-based services." Proceedings of the 13th Workshop on Privacy in the Electronic Society. ACM, 2014.

Wernke, Marius, et al. "A classification of location privacy attacks and approaches." Personal and ubiquitous computing 18.1 (2014): 163-175.

This research was completed with the generous support of the Research Foundation of Flanders (FWO), IMEC-COSIC KU Leuven and Center for Information and Technology at Princeton University.

SEDA GÜRSES is a Postdoctoral Research Associate at CITP, Princeton University and an FWO fellow at COSIC, University of Leuven in Belgium. She works on privacy and requirements engineering, privacy enhancing technologies and surveillance. Previously she was a post-doctoral fellow at the Media, Culture and Communications Department at NYU Steinhardt and at the Information Law Institute at NYU Law School, where she was also part of the Intel Science and Technology Center on Social Computing.

OBFUSCATION IN BITCOIN: TECHNIQUES AND POLITICS

ARVIND NARAYANAN AND **MALTE MÖSER,** PRINCETON UNIVERSITY

OBFUSCATION TECHNIQUES

Bitcoin's design is centered around a widely distributed, global database which stores all transactions that have ever taken place in the system. Thus, there is no avenue for redress if a user wishes to retrospectively hide a transaction. Further, nothing in the ledger is encrypted, and digital signatures are mandatory, ensuring cryptographic attribution of activities to users. On the other hand, account identifiers in Bitcoin take the form of cryptographic public keys, which are pseudonymous. Anyone can use Bitcoin "wallet" software to trivially generate a new public key and use it as a pseudonym to send or receive payments without registering or providing personal information. However, pseudonymity alone provides little privacy, and there are many ways in which identities could be linked to these pseudonyms (Narayanan et al., 2016).

To counter this, Bitcoin and its users employ a variety of obfuscation techniques to increase their financial privacy. We visualize a representative selection of these techniques in Figure 1 based on their time of invention/creation and our assessment of their similarity to obfuscation vs cryptography. We make several observations. First, techniques used in Bitcoin predominantly fall into obfuscation, with stronger techniques being used exclusively in alternative cryptocurrencies (altcoins). Second, there is a trend towards stronger techniques over time, perhaps due to a growing interest in privacy and to the greater difficulty of developing cryptographic techniques. Third, obfuscation techniques proposed at later points in time are seeing less adoption, arguably a result of their increased complexity and need for coordination among participants (Möser & Böhme 2017).

Among the techniques used in Bitcoin, the most prevalent can be characterized as "ambiguating obfuscation" (Brunton & Nissenbaum 2015): effectively reducing the information an adversary is able to extract from a particular transaction. Examples include using a new pseudonym for every new transaction and randomizing the structure of transactions to make the spend to the "true" recipient indistinguishable from "change" going back to the sender.

A second type of obfuscation, namely "cooperative obfuscation", has risen in popularity over the last years. For example, users can send their money to a service that will "mix" their funds with those of other users, thereby obfuscating the flow of payments (cf. Möser, Böhme & Breuker 2013). A similar technique called CoinJoin works in a peer-to-peer fashion and doesn't require a trusted intermediary is CoinJoin. Due to the need for

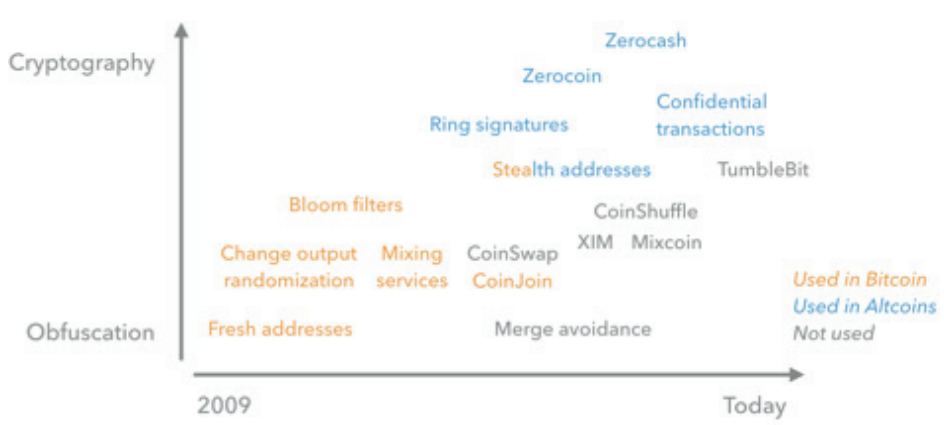


Figure 1: Privacy-Enhancing Technologies for Bitcoin. The X-axis is the date of invention and the Y-axis is an informal measure that combines the sophistication of the technique and the strength of the privacy guarantee. See Appendix 1 for references.

these users to find and transact with each other, markets for anonymity have arisen that bring together providers and receivers of anonymity (Möser & Böhme 2016).

THE CASE FOR OBFUSCATION

Critically, none of the techniques discussed provide provable privacy guarantees through cryptography. While these do exist and have been deployed (e.g., Zcash), they are far from being adopted by the Bitcoin community, for both technical and political reasons. On the technical side, Bitcoin's decentralization already incurs a severe performance penalty compared to centralized payment systems such as Paypal. Achieving cryptographic privacy would further degrade performance. Obfuscation also has a lighter impact on the usefulness of the blockchain for non-currency applications. The current design allows selectively employing obfuscation, leaving room for other uses that prioritize different goals, such as Colored Coins (Rosenfeld 2012), a protocol for representing assets on top of the Bitcoin blockchain.

On the political side, providing stronger privacy through cryptography might make Bitcoin even more attractive for activities such as money laundering, ransomware, or terrorism financing, and thereby tempt a government crackdown. Much of the Bitcoin community is invested in its mainstream adoption, and therefore keen to avoid such an outcome. When Bitcoin began to be noticed by the press, members of the community went to work explaining it to policy makers. They framed the technology as neutral and unthreatening, and the Bitcoin ecosystem as subject to existing regulations and amenable to new ones (cf. Brito 2013, Brito & Castillo 2013, Lee 2013, Murck 2013, Hattem 2014).

The use of obfuscation in Bitcoin may have achieved a balancing act between the financial privacy of its users and the investigatory needs of law enforcement and regulators. Law enforcement agencies have two important advantages over everyday adversaries: the budget for specialized Bitcoin tracking tools and services (Cox 2017), and subpoena power. The latter allows deanonymizing selected actors by obtaining user records from exchanges and cross-referencing them with the results of blockchain analysis (Meiklejohn et al. 2013). Since only a few governmental actors possess these powers, users still enjoy a measure of financial privacy. Thus, the imperfect privacy protection in Bitcoin may be one of the keys to its success.

REFERENCES

Bissias, G., Ozisik, A. P., Levine, B. N., & Liberatore, M. (2014). Sybil-Resistant Mixing for Bitcoin. In Proceedings of the 13th Workshop on Privacy in the Electronic Society (pp. 149-158). ACM.

Bonneau, J., Narayanan, A., Miller, A., Clark, J., Kroll, J. A., & Felten, E. W. (2014). Mixcoin: Anonymity for Bitcoin with Accountable Mixes. In International Conference on Financial Cryptography and Data Security (pp. 486-504). Springer Berlin Heidelberg.

Brito, J., & Castillo, A. (2013). Bitcoin: A Primer for Policymakers. Mercatus Center at George Mason University.

Brito, J. (2013). Beyond Silk Road: Potential Risks, Threats, and Promises of Virtual Currencies. Testimony to the Senate Committee on Homeland Security and Governmental Affairs. Available online at https://www.mercatus.org/system/files/Brito_BeyondSilkRoadBitcoin_testimony_111313.pdf (retrieved on 2017-06-02).

Brunton, F., & Nissenbaum, H. (2015). Obfuscation: A User's Guide for Privacy and Protest. MIT Press.

Cox, J. (2017). US Law Enforcement Have Spent Hundreds of Thousands on Bitcoin Tracking Tools. Motherboard. Available online at https://motherboard.vice.com/en_us/article/us-law-enforcement-have-spent-hundreds-of-thousands-on-bitcoin-tracking-tools (retrieved on 2017-06-02).

Hattem, J. (2014). Bitcoin Gets a Lobbyist. The Hill. Available online at http://thehill.com/policy/technology/207085-bitcoin-investors-register-lobbyist (retrieved on 2017-06-02).

Hearn, M. (2013). Merge Avoidance. Available online at https://medium.com/@octskyward/merge-avoidance-7f95a386692f (retrieved on 2017-06-02).

Heilman, E., Baldimtsi, F., Alshenibr, L., Scafuro, A., Goldberg, S. (2017). TumbleBit: An Untrusted Tumbler for Bitcoin-Compatible Anonymous Payments. In Network and Distributed System Security Symposium (NDSS).

Lee, T. B. (2013). Here's How Bitcoin Charmed Washington. The Washington Post. Available online at https://www.washingtonpost.com/news/the-switch/wp/2013/11/21/heres-how-bitcoin-charmed-washington (retrieved on 2017-06-02).

Maxwell, G. (2013a). CoinJoin: Bitcoin Privacy for the Real World. Available online at https://bitcointalk.org/index.php?topic=279249.0 (retrieved on 2017-06-02).

Maxwell, G. (2013b). CoinSwap: Transaction Graph Disjoint Trustless Trading. Available online at https://bitcointalk.org/index.php?topic=321228 (retrieved on 2017-06-02).

Maxwell, G. (2015). Confidential Transactions, the Initial Investigation. Available online at https://www.elementsproject.org/elements/confidential-transactions/investigation.html (retrieved on 2017-06-02).

Meiklejohn, S., Pomarole, M., Jordan, G., Levchenko, K., McCoy, D., Voelker, G. M., & Savage, S. (2013). A Fistful of Bitcoins: Characterizing Payments Among Men with No Names. In Proceedings of the 2013 Conference on Internet Measurement (pp. 127-140). ACM.

Miers, I., Garman, C., Green, M., & Rubin, A. D. (2013). Zerocoin: Anonymous Distributed E-Cash from Bitcoin. In 2013 IEEE Symposium on Security and Privacy (S&P) (pp. 397-411). IEEE.

Möser, M., Böhme, R., & Breuker, D. (2013). An Inquiry Into Money Laundering Tools in the Bitcoin Ecosystem. In eCrime Researchers Summit, 2013 (pp. 1-14). IEEE.

Möser, M., & Böhme, R. (2016). Join Me on a Market for Anonymity. In Workshop on the Economics of Information Security (WEIS).

Möser, M., & Böhme, R. (2017). Anonymous Alone? Measuring Bitcoin's Second-Generation Anonymization Techniques. In IEEE Security & Privacy on the Blockchain (IEEE S&B). IEEE.

Murck, P. (2013). Testimony of Patrick Murck General Counsel, the Bitcoin Foundation to the Senate Committee on Homeland Security and Governmental Affairs "Beyond Silk Road: Potential Risks, Threats, and Promises of Virtual Currencies". Available online at https://www.hsgac.senate.gov/download/?id=4CD1FF12-312D-429F-AA41-1D77034EC5A8 (retrieved on 2017-06-02).

Narayanan, A., Bonneau, J., Felten, E., Miller, A., & Goldfeder, S. (2016). Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction. Princeton University Press.

Rosenfeld, M. (2012). Overview of Colored Coins. Available online at https://bitcoil.co.il/BitcoinX.pdf (retrieved on 2017-06-02).

Ruffing, T., Moreno-Sanchez, P., & Kate, A. (2014). CoinShuffle: Practical Decentralized Coin Mixing for Bitcoin. In European Symposium on Research in Computer Security (pp. 345-364). Springer International Publishing.

Sasson, E. B., Chiesa, A., Garman, C., Green, M., Miers, I., Tromer, E., & Virza, M. (2014). Zerocash: Decentralized anonymous payments from bitcoin. In 2014 IEEE Symposium on Security and Privacy (S&P) (pp. 459-474). IEEE.

Todd, P. (2014). Stealth Addresses. Available online at https://lists.linuxfoundation.org/pipermail/bitcoin-dev/2014-January/004020.html (retrieved on 2017-06-02).

Van Saberhagen, N. (2013). CryptoNote v2.0. Available online at https://cryptonote.org/whitepaper.pdf (retrieved on 2017-06-02).

APPENDIX 1: REFERENCES FOR PRIVACY-ENHANCING TECHNIQUES FOR CRYPTOCURRENCIES

NAME	REFERENCE
Bitcoin mixers	cf. Möser, Böhme & Breuker 2013
CoinJoin	Maxwell 2013a
CoinShuffle	Ruffing 2014
CoinSwap	Maxwell 2013b
Confidential transactions	Maxwell 2015
Merge avoidance	Hearn 2013
Mixcoin	Bonneau et al. 2014
Ring signatures	Van Saberhagen 2013
Stealth addresses	Todd 2014
TumbleBit	Heilman et al. 2017
XIM	Bissias et al. 2014
Zerocash	Sasson et al. 2014
Zerocoin	Miers et al. 2013

APPENDIX 2: THE LIFECYCLE OF OBFUSCATION

The success of obfuscation in Bitcoin motivates studying the adoption of obfuscation in sociotechnical systems more generally. To this end, we present a simplified model of the adoption of obfuscation, visualized in Figure 2.

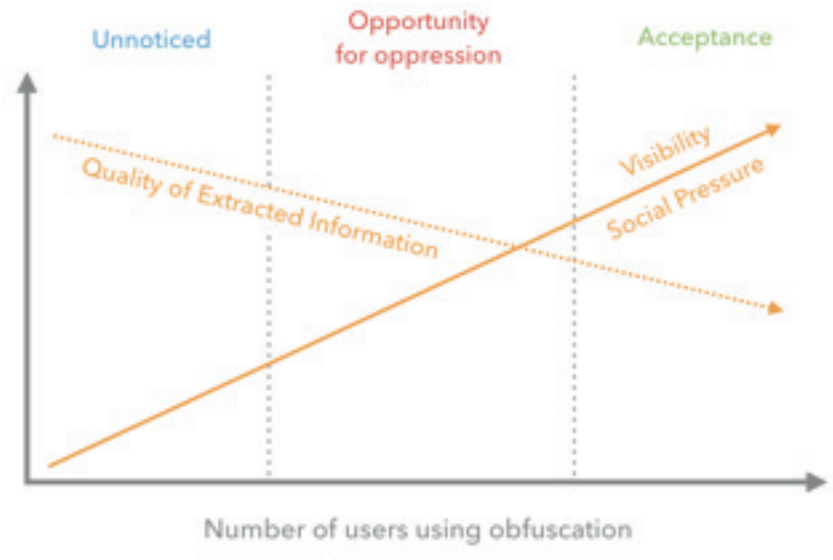


Figure 2: Lifecycle of Obfuscation

We conjecture that as the number of users of obfuscation grows, the visibility of the use of obfuscation increases as well. It also reduces the quality of the information that can be extracted from the system. We argue that initially, the use of obfuscation is mostly unnoticed as the user base and its impact is small. On

the other hand, once obfuscation has reached a critical mass, social pressure helps against the platform owner's (or government's) wish to oppress obfuscation, leading to general acceptance. A good example is "Nymwars", i.e. Google's (and other companies) attempt to forbid the use of pseudonyms on social networks. Due to a large, negative public reaction Google had to reverse its decision to ban the use of pseudonyms. This suggests a critical phase in between these two, where there is opportunity for oppression by the platform owner or government. For those who aim to establish obfuscation as a means of defense against a system, this suggests two related strategies to minimize the window of oppression. The first is to hide the use of obfuscation for as long as possible through both social and technical means. The second is to maximize the visibility of obfuscation and campaign for its acceptance once it can no longer remain unnoticed.

ARVIND NARAYANAN is an Assistant Professor of Computer Science at Princeton. He leads the Princeton Web Transparency and Accountability Project to uncover how companies collect and use our personal information. Narayanan also leads a research team investigating the security, anonymity, and stability of cryptocurrencies as well as novel applications of blockchains. He co-created a Massive Open Online Course as well as a textbook on Bitcoin and cryptocurrency technologies. His doctoral research showed the fundamental limits of de-identification, for which he received the Privacy Enhancing Technologies Award. Narayanan is an affiliated faculty member at the Center for Information Technology Policy at Princeton and an affiliate scholar at Stanford Law School's Center for Internet and Society. You can follow him on Twitter at @random_walker.

MALTE MÖSER is a first year PhD student in the Department of Computer Science at Princeton University and a Graduate Student Fellow at the Center for Information Technology Policy, advised by Professor Arvind Narayanan. He is broadly interested in the security and anonymity of cryptographic currencies. You can follow him on Twitter at @maltemoeser.

OBFUSCATION AND THE THREAT OF CENTRALIZED DISTRIBUTION

DANIEL C. HOWE, School of Creative Media, City University of Hong

Earlier this year, Helen Nissenbaum, Mushon Zer-Aviv, and I released an updated version of AdNauseam with a number of new features. AdNauseam is the adblocker that clicks every ad in an effort to obfuscate tracking profiles and inject doubt into the economics driving advertising-based surveillance. Soon after the release, we learned that Google had banned AdNauseam from its store, where it had been available for the past year. We've since learned that Google was disallowing users from manually installing or updating AdNauseam on their Chrome browser.

The fact that the distribution of open-source extensions is now largely in the hands of a few multinational corporations, operating with little oversight, highlights the threat of recent moves toward centralized distribution. Whether or not you agree with AdNauseam's approach, it is chilling to realize that Google can quietly make one's extensions and data disappear at their whim. Today it is a privacy tool that is disabled, tomorrow it could be your photo app, chat program, or even password manager. And you don't simply lose the app, you lose your stored data as well: photos, chat transcripts, passwords. For developers, who incidentally must pay a fee to post items in the Chrome store, this should cause one to think twice. Not only can your software be banned and removed without warning, but all ratings, reviews and statistics are deleted as well.

When we wrote Google to ask the reason for the removal, they initially responded that AdNauseam had breached the Web Store's Terms of Service, stating that "An extension should have a single purpose that is clear to users." Only months later did Google admit the actual reason for the block: that AdNauseam was interfering with their ad networks. In Google's final official response, Senior Policy Specialist Dr. Michael Falgoust confirmed that: "[AdNauseam] appears to simulate user behavior by sending automated clicks in such a way that may result in financial harm to third party systems such as advertising networks." As one could also claim economic harms from adblockers (which are, as yet, not blocked in the Chrome store), we are left to speculate whether they might be other reasons behind the takedown. Our guess is that part of Google's antipathy toward AdNauseam can be traced to a new feature: specifically our built-in support for the EFF's Do Not Track mechanism [1].

For anyone unfamiliar, this is not the ill-fated DNT of yore, but a new, machine-verifiable (and potentially legally-binding) assertion on the part of websites that commit to not violating the privacy of users who send the Do-Not-Track header. A new generation of blockers, including

the EFF's Privacy Badger and AdNauseam, have support for this mechanism enabled by default; which means that they don't block ads and other resources from DNT sites, and, in the case of AdNauseam, don't simulate clicks on these ads.

So why is this so threatening to Google? Perhaps because it could represent a real means for users, advertisers, and content-providers to move away from surveillance-based advertising. If enough sites commit to Do Not Track, there will be significant financial incentive for advertisers to place ads on those sites, and these too will be bound by DNT, as the mechanism also applies to a site's third-party partners. And this could possibly set off a chain reaction of adoption that would leave Google, which has committed to surveillance as its core business model, out in the cold.

But wait, you may be thinking, why did the EFF develop this new DNT mechanism when there is AdBlock Plus' "Acceptable Ads" programs, which Google and other major ad networks already participate in? That's because there are crucial differences between the two. For one, "Acceptable Ads" is pay-to-play; large ad networks pay Eyeo, the company behind Adblock Plus, to whitelist their sites. But the more important reason is that the program is all about aesthetics—so-called "annoying" or "intrusive" ads—which the ad industry would like us to believe is the only problem with the current system. An entity like Google is fine with "Acceptable Ads" because they have more than enough resources to pay for whitelisting [2]. Further, they are quite willing to make their ads more aesthetically acceptable to users (after all, an annoyed user is unlikely to click) [3]. What they refuse to change—though we hope we're wrong about this—is their commitment to surreptitious tracking on a scale never before seen. And this, of course, is what we, the EFF, and a growing number of users find truly "unacceptable" about the current advertising landscape.

NOTES

Note: a version of this argument was published on the "Freedom to Tinker" blog as "AdNauseam, Google, and the Myth of the 'Acceptable Ad'".

[1] This is indeed speculation. However, as mentioned in [1], the stated reason for Google's ban of AdNauseam does not hold up to scrutiny.

[2] In September of this year, Eyeo announced (https://adblockplus.org/blog/new-acceptable-ads-platform-launches-bringing-feedback-to-rtb-and-help-to-small-websites) that it would partner with a UK-based ad tech startup called ComboTag to launch the "Acceptable Ads Platform" with which they would act also as an ad exchange, selling placements for "Acceptable Ad" slots. Google, as might be expected, reacted negatively (http://arstechnica.com/tech-policy/2016/09/adblock-plus-starts-selling-ads-but-only-acceptable-ones/), stating that it would no longer do business with ComboTag. Some assumed that this might also signal an end to their participation in "Acceptable Ads" as well. However, this does not appear to be the case. Google still comprises a significant portion of the exception list (https://easylist-downloads.adblockplus.org/exceptionrules.txt) on which "Acceptable Ads" is based and, as one ad industry observer put it (http://adage.com/article/digital/ad-blocking-battle-side-google/305984/), "Google is likely Adblock Plus' largest, most lucrative customer."

[3] Google is also a member of the "Coalition for Better Ads" (https://www.betterads.org/), an industry-wide effort which, like "Acceptable Ads," focuses exclusively on issues of aesthetics and user experience, as opposed to surveillance and data profiling.

DANIEL C. HOWE is an artist and critical technologist whose work focuses on the social, cultural and political implications of algorithmic technologies. His projects includes TrackMeNot, RiTa, AdNauseam, AdLiPo, ChinaEye, and Advertising Positions. He currently divides his time between New York and Hong Kong, where he teaches at the School of Creative Media.