Auditing Algorithms From the Outside: Methods and Implications
a proposal for a half-day workshop

Mike Ananny, Communication and Journalism, University of Southern California
Karrie Karahalios, Computer Science, University of Illinois (***primary contact person***)
Christian Sandvig, Communication and Information, University of Michigan
Christo Wilson, Computer Science, Northeastern University

Short Abstract
An emerging area of research we call "algorithm auditing" allows researchers, designers, and users new ways to understand the algorithms that increasingly shape our online life. This research investigates algorithms "from the outside," testing them for problems and harms without the cooperation of online platform providers. So far, researchers have investigated the systems that handle recommendations, prices, news, commenting, and search, examining them for individually and societally undesirable consequences such as racism or fraud. In this workshop we treat this new area as a development of the traditional social science technique—the audit study—that creates new ways of investigating the function and relevance of algorithms. Through a cumulative set of activities designed by a group of multidisciplinary organizers, participants will create, document, trick, interrogate, and propose revisions to social media and other algorithms, leaving the attendees with an appreciation of both the ethical and methodological challenges of studying networked information algorithms, as well as a chance for the development of new research designs and agendas to guide future work in this area.

Workshop Description
The most prevalent social scientific method for the detection of employment and housing discrimination is the audit study (Mincy, 1993). Audit studies are field experiments in which (in one common design) a fictitious correspondence is created purporting to be from actual job applicants seeking employment, targeted at a real employer, but some variable of interest such as race, geography, or gender is manipulated to detect undesirable discrimination (Riach & Rich, 2002; Pager, 2007). In computing research about personalization, security, social media, and Internet platforms, researchers faced with understanding "black box" Internet algorithms (Pasquale, 2014) for news feeds, search results, recommendations, and prices are now turning to research designs that are variations on the audit (Sandvig et al., 2014a). In this workshop, we propose to explore the methods and implications of this emerging research area.

Importantly, auditing an algorithm is not "reverse engineering" by an attacker or competitor to recreate a proprietary mechanism. Rather, auditing is research undertaken to investigate governmental or commercial Internet-based intermediaries with large data repositories (e.g., YouTube, Google, Facebook, Netflix, and so on) to ascertain whether they are operating in some normatively harmful way--such as unlawful discrimination by class, race, gender or the undesirable promotion of antisocial content (for an overview, see Barocas, Hood, and Ziewitz, 2013; Gillespie, 2014). Like traditional auditing, it is research undertaken "from the outside:" without the cooperation of the platforms themselves, and this raises significant design, legal and ethical challenges (Sandvig et al., 2014).

While algorithm audits evoke the earlier history of housing and employment audits, harmful online algorithms also differ from earlier processes of harmful discrimination (such as redlining in real estate) in a number of crucial ways. One difference is that these algorithms are often so complex that they cannot be understood or interpreted even if the source code is available. That is, an expert in the area (or even the algorithm's author) may not be able to predict what results an algorithm would produce in a particular case without plugging in example data. Indeed, many algorithms are created or managed by teams of individuals, making

it almost impossible to identify a single author or expert who *might* understand the algorithm's functions or effects. Furthermore, algorithms increasingly depend on huge amounts of (personal) data as inputs. Even given the algorithm's code, its outputs cannot be understood in the absence of the underlying input data. Finally, machine learning algorithms also routinely produce results that can not be reasonably foreseen or intended by the algorithm's author.

Algorithm auditors and topics are diverse. Influential examples of audits have come from computer scientists, social scientists, and journalists working in venues that include university research labs--but also government agencies and even ad hoc groups of disgruntled customers. For example, auditors of online platforms have discovered the following: searches for African American-identifying names are more likely to produce advertisements that suggest the person holding that name has an arrest record (Sweeney, 2013); a New York State welfare case management system was secretly designed to reduce payments rather than increase efficiency (Eubanks, 2014); Google search inadvertently gave Obama a large coverage advantage in search results during the last presidential election (Angwin 2012); sending e-mails containing words associated with depression produces predatory advertising for spirit healing (Lecuyer et al., 2014); non-African American hosts on AirBnB charge 12% more than African American hosts for an equivalent rental (Edelman & Luca, 2014); users of Android phones are charged 6% more on Home Depot's Web site (Hannack et al. 2014); and Facebook implemented "likes" in a way that can produce an "X likes Y" message that is the opposite of the user's intent (Meisler, 2012). This multidisciplinary context is an ideal fit for ICWSM's mission to contribute to both social and computational research.

Research methods in this area of study are refinements of the traditional social science audit, but are also still emerging. A "classic" audit design investigating an online platform would entail the insertion of fictitious data, requests, or user accounts ("sock puppets"), just as housing economists gave landlords fictitious rental applications in the 1970s to diagnose racial discrimination (Mincy, 1993). However, online audits also demand new techniques. A second, modified audit design might involve correlating existing user or public data without inserting new data or an experimental manipulation. A third strategy might involve securing access to the source code, input data, or design specifications for a system. These designs vary in scale and complexity. While some designs involve a small army of Mechanical Turkers elaborately generating test data that are analyzed with sophisticated statistics, successful audits have been the result of a single user experimenting on their own.

In this workshop, we propose to explore this new area of research by interrogating and advancing its methods and research designs, considering its implications, and assessing the current challenges researchers in the area face. We also anticipate a key problem of this kind of research to be properly sampling, quantifying, analyzing, visualizing and interpreting the results in a powerful and meaningful way such that they can be used as evidence in public debates. We propose to address these issues by convening a group of four co-organizers who span computer science and social science. We have prior experience in research design, public policy, online audit research, and research ethics. Our prior work is both conceptual (e.g., Sandvig et al., 2013) and empirical (Hannak et al., 2013 and 2014), and includes studies that required scraping data from major Internet platforms (e.g., Gilbert et al., 2010). We will facilitate interactive activities that are designed to allow attendees to collaborate in small groups on specific research and design problems. This will be done in a context of peer discussion and will not employ invited papers or formal presentations. A concrete deliverable will be a workshop report to guide new work in the area and potentially organizing a tradition of this research as a future feature of ICWSM conferences.

Format and Timetable

We anticipate that we will adjust the format and timetable as we receive registrants. If possible, we would like to query registrants in advance for the conference about their prior knowledge and interests. (We will also promote this workshop to the emerging research community.) We will also plan to be flexible on the day of the workshop itself in order to accommodate the progress that we make through our interactive group activities. This timetable is thus very provisional.

Major Themes of Discussion Throughout the Workshop

At the introduction, conclusion, and between activities the co-convenors will be dedicated to facilitating the open discussion of the following major themes. These themes will be provided to participants in advance, if possible:

- The Co-Construction of Algorithmic Systems that Involve Machine Learning, User Data, and Online Platforms
- The Requirements for Successful Auditing vs. Other Research Interventions (e.g., vs. Improving Algorithms, Finding Security Vulnerabilities in Algorithms, or Reverse Engineering Algorithms)
- The Societal Problems of Harmful Algorithms, Including Unlawful Discrimination and Problems of Equity and Justice
- The Public Awareness of Algorithms -- Including “Processual Thinking” and “Computational Thinking” as Described in the STEM Education Literature
- Revisiting Other ICWSM Research Areas But With the Perspective of Detecting Algorithm and Platform Misbehavior, e.g., Recommendation Systems, Information Retrieval, Social Media Friend Recommendation, Feed Curation, etc.
- Research Ethics and Research Design
- Research Methods from the Practical to the Statistical that Allow Researchers to Obtain and Analyze Audit Data
- Legal Problems and Solutions from the Perspective of Researchers (e.g., Terms of Service), Users (e.g., Privacy), and Platform Providers (e.g., Trade Secrets)

Provisional Timeline

While this half-day workshop can be morning or afternoon, we propose a morning timeline here, ending with a group lunch for the workshop participants.

8:00am	Introduction of organizers and participants
8:30am	Brief description of audits, the main challenges and explanation of Exercise 1
9:00am	Exercise 1: Facebook groups exercise
9:45am	Short break
10:00am	Explanation of Exercise 2
10:15am	Exercise 2: Group audit using five different strategies
11:15am	Group discussion
11:45am	Closing
12:00pm	Group lunch for those interested

Exercise 1: The first exercise will begin with the creation of a Facebook group consisting of the workshop participants and organizers. We will spend the first fifteen minutes of this session populating this group feed with articles and comments about algorithmic curation, awareness, application, ethics, and policy. The purpose for this is two-fold: (1) We plan to use this group throughout the workshop and perhaps after as an archive of our discussion, (2) Group feeds differ from the basic Facebook News Feed in that there is no option to view the feed in chronological order. From the onset, prioritization algorithms dictate the order of

presentation for the audience. We will use this data for the remainder of Exercise 1 to create a systematic face-to-face real-time collective group approach to discovering the prioritization elements that curate this group feed.

Exercise 2: We briefly present five audit strategies in the Workshop Description above. In Exercise 2, the workshop participants will divide into five groups, one for each of the strategies. The groups will be presented with the same data set, tentatively, the Facebook group data. The goal of each group is to use their assigned audit strategy to uncover feed personalization and prioritization features. We will then compare the resulting features compiled in each of the five groups.

Historical Workshops and Panels

1. "Governing Algorithms," a 2013 symposium at NYU, was a broad introduction to the politics of algorithms, how they govern us, and what governance is necessary for them. The current proposal is an in-depth exploration of implications from this 2013 symposium. All four co-organizers of this proposal participated. This symposium was full.
2. "Data and Discrimination," was a full-day workshop at ICA 2014 organized by the Open Technology Institute. This proposal's co-organizers, Karahalios and Sandvig, participated and developed the idea of an algorithm auditing framework during this workshop. This workshop was full.
3. "We Guess Your Algorithm," a two-hour experimental session at AoIR 2014 was proposed and run by two co-organizers of the current proposal. The session involved a software demo and produced the beginning of thinking about algorithm audits and research methods. This session was full and produced very favorable feedback to AoIR organizers, as well as positive feedback on Twitter.
4. "Algorithms and the Future of Accountability Reporting," a panel at the 2014 Computation+Journalism symposium at Columbia University. http://computation-and-journalism.brown.columbia.edu/ Two co-organizers participated in a roundtable with journalists and ethicists. This session was nearly full, the discussion was extremely lively, and the Q&A line could not be satisfied in the time allowed. Several attendees suggested that a workshop or symposium would be a more appropriate venue for the topic.

Related Workshops and Panels

There are currently a large number of related workshops, typically focusing on social media, "big data," and/or privacy. However, these often include some discussion of algorithms and even auditing them, a topic that has usually emerged just in the last two years. Here is a small selection:

1. "Ethics of Data in Civil Society," A 2014 symposium at Stanford University. The proposed workshop develops the algorithms portion of this event in more detail. One co-organizer participated. http://pacscenter.stanford.edu/sites/all/files/SYNTHESIS%20Ethics%20of%20Data%20in%20Civil%20SocietyFINAL%20%281%29%20%281%29.pdf
2. "Fairness, Accountability, and Transparency in Machine Learning" (FAT ML), hosted by the Center for Information Technology Policy (CITP) at Princeton. http://fatml.org/index.html One co-organizer participated.
3. "The Social, Cultural, and Ethical Dimensions of Big Data," a conference organized by the White House Office of Science and Technology Policy. http://www.datasociety.net/initiatives/2014-0317/

Bios

Mike Ananny is an Assistant Professor at the University of Southern California's Annenberg School for Communication & Journalism, an Affiliated Faculty with USC's Science, Technology and Society research cluster, and co-leader of the Civic Paths research group. He studies the public significance and sociotechnical dynamics of networked news infrastructures, focusing on how they encode journalistic values and normative theories of the press.

He has held fellowships and scholarships with Harvard's Berkman Center on Internet and Society, Stanford's Center on Philanthropy and Civil Society, the Pierre Elliott Trudeau Foundation, the LEGO Corporation, and Interval Research. He was a founding member of Media Lab Europe's research staff, a postdoc with Microsoft Research's Social Media Collective, and has worked or consulted for LEGO, Mattel, and Nortel Networks. His PhD is from Stanford University (Communication), SM from the MIT Media Lab (Media Arts & Sciences), and BSc from the University of Toronto (Human Biology & Computer Science). He has published in a variety of venues including *Critical Studies in Media Communication*, *International Journal of Communication*, *Journal of Computer Mediated Communication*, *First Monday*, *American Behavioral Scientist*, *Television & New Media*, *Digital Journalism*, and the proceedings of the ACM's conferences on *Computer-Human Interaction* and *Computer Supported Collaborative Learning*. He is writing a book on press freedom and a public right to hear in the age of networked journalism (under contract with MIT Press).

Karrie Karahalios (*primary contact person*) is an Associate Professor at the University of Illinois in Urbana-Champaign where she heads the Social Spaces Group. Her work focuses on the interaction between people and the social cues they emit and perceive in face-to-face and mediated electronic spaces. Her work is informed by communication studies, sociology, art&design, computer science, linguistics, and psychology. Of particular interest are interfaces for public online and physical gathering spaces such as Twitter, chatrooms, cafes, parks, etc. Research projects range from studying tie strength between people to encouraging vocalization through visualization. A major theme in the work is to create interfaces that enable users to perceive conversational patterns that are present, but not obvious, in traditional communication interfaces. Her PhD (Media Arts and Sciences), MS (Media Arts and Sciences), MEng (Electrical Engineering and Computer Science) and SB (Electrical Engineering and Computer Science) are from the Massachusetts Institute of Technology. She received the Alfred P. Sloan Research Fellowship, the A. Richard Newton Breakthrough Research Award and the Faculty Early-Career Development Award from the US National Science Foundation (NSF CAREER) in the area of human-centered computing to better understand and visualize relationship and conversation dynamics.

Christian Sandvig is Steelcase Research Professor and Associate Professor in both the Department of Communication Studies and the School of Information at the University of Michigan. He is also a faculty associate of the Berkman Center for Internet & Society at Harvard University. Sandvig is a social scientist studying the implications of the algorithmic curation of culture. He is also a computer programmer with industry experience in the Fortune 500, government, and a start-up. He holds the MA and PhD in Communication Research from Stanford University (2002) and received the US National Science Foundation's Faculty Early-Career Development Award (NSF CAREER) in the area of human-centered computing. Sandvig was previously named a "next-generation leader" in technology policy by the American Association for the Advancement of Science. He has published and presented in a variety of venues, including ICWSM (previous honorable mention), ACM CHI (best paper), ICA (top paper), TPRC (first prize), and AEJMC (top paper). His research has appeared in *The New York Times*, *The Economist*, *New Scientist*, *National Public Radio*, and *CBS News*. He previously organized a related experimental session at AoIR in 2014.

Christo Wilson is an Assistant Professor in the College of Computer and Information Science at Northeastern University. His current research focus is on understanding the types of data that online entities track about individuals, and how this data is used to personalize content on the Web. His current work in this space has examined personalization on major search engines, as well as price discrimination and price steering on major e-commerce and travel websites. He completed his PhD in Computer Science at the University of California, Santa Barbara under the direction of Ben Y. Zhao in 2012. Professor Wilson received a Best Paper: Honorable Mention award at SIGCOMM 2011, and his work has been featured on CBS Nightly News, Scientific American, NPR Marketplace, MIT Technology Review, the Wall Street Journal, and the Boston Globe.

Bibliography of Related Work and Sources Cited

Ananny, Mike. (2011). The Curious Connection Between Apps for Gay Men and Sex Offenders. *The Atlantic*, April 14, 2011.

Angwin, Julia. (2012, November 4). On Google, a Political Mystery That's All Numbers. *The Wall Street Journal.* http://www.wsj.com/articles/SB10001424052970203347104578099122530080836

Barocas, Solon, Hood, Sophie, and Ziewitz Malte. (2013). Governing Algorithms: A Provocation Piece. A paper presented to *Governing Algorithms*, NYU School of Law, New York, NY. http://governingalgorithms.org/resources/provocationpiece/

Edelman, Ben & Luca Michael. (2014). Digital Discrimination: The Case of Airbnb.com. Discussion Paper. Online: http://www.benedelman.org/publications/airbnb011014.pdf

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Voung, A., Karahalios, K., Hamilton, K., Sandvig, C. (2015). "I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in the news feed. Manuscript submitted for publication.

Eubanks, V. (2014). Big Data and Human Rights. In: S. P. Gangadharan (ed.), *Data and Discrimination: Collected Essays*, pp. 48-52. Washington, DC: New America Foundation. http://newamerica.org/downloads/OTI-Data-an-Discrimination-FINAL-small.pdf

Gilbert, Eric; Karahalios, Karrie; and Sandvig, Christian. 2010. The Network in the Garden: Designing Social Media for Rural Life. *American Behavioral Scientist 53*(9), 13671388.

Gilbert, Eric and Karahalios, Karrie. (2009). Predicting Tie Strength from Social Media. Proceedings of the Conference on *Human Factors in Computing Systems (CHI).*

Gillespie, Tarleton. (2014). The Relevance of Algorithms. in: Tarleton Gillespie, Pablo J. Boczkowski and Kirsten A. Foot (eds.), *Media Technologies: Essays on Communication, Materiality, and Society.* Cambridge: MIT Press.

Hamilton, K., Karahalios, K., Sandvig, C., & Eslami, M. (2014, April). A path to understanding the effects of algorithm awareness. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (pp. 631-642). New York: ACM.

Hannack, A., Sapiezynski, P., Kakhki, A. M., Krishnamurthy, B., Lazer, D., Mislove, A., & Wilson, C. (2013). Measuring Personalization of Web Search. Paper presented to the *ACM International World Wide Web Conference (WWW)*, Rio de Janeiro, Brazil. http://personalization.ccs.neu.edu/papers/web_search.pdf

Hannack, A., Soeller, G., Lazer, D., Mislove, A., & Wilson, C. (2014). Measuring Price Discrimination and Steering on E-commerce Web Sites. Paper presented to the *ACM Internet Measurement Conference (IMC)*, Vancouver, BC, Canada. http://personalization.ccs.neu.edu/papers/price_discrimination.pdf

Lecuyer, M, Ducoffe, D, Lan, F., Papancea, A., Petsios, T., Spahn, R., Chaintreau, A., & Geambasu, R. (2014). XRay: Enhancing the Web's Transparency with Differential Correlation. *23rd USENIX Security Symposium.*
https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-lecuyer.pdf

Light, J. (2011). Discriminating Appraisals: Cartography, Computation, and Access to Federal Mortgage Insurance in the 1930s. *Technology and Culture* 52, 3, 485-52.

Meisler, B. (2012, December 11). Why Are Dead People Liking Stuff on Facebook? *ReadWrite.*
http://readwrite.com/2012/12/11/why-are-dead-people-liking-stuff-on-facebook

Mincy, Ronald. (1993). The Urban Institute Audit Studies: Their Research and Policy Context. In Fix and Struyk, eds., *Clear and Convincing Evidence: Measurement of Discrimination in America*. Washington, DC: The Urban Institute Press, pp. 16586.

Pager, Devah. (2007). The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future. *The Annals of the American Academy of Political and Social Science*, Vol. 609, No. 1, pp. 10433.

Pasquale, F. (2010). "Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries," 104. *Northwestern University Law Review* 105.

Pasquale, F. A. (2011). Restoring Transparency to Automated Authority. *Journal on Telecommunications & High Technology Law*, 9 (235).

Pasquale, F. A. (2014). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge: Harvard University Press.

Riach, Peter A., and Judith Rich. (2002). Field Experiments of Discrimination in the Market Place. *The Economic Journal*, Vol. 112, No. 483, November, pp. F480518.

Rickman, A. & Sandvig, C. (2014) Broke and buying rides: Adolescent girls and social media Brokering. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM '14)*.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014a, May). Auditing algorithms: Research methods for detecting discrimination on internet platforms. Paper presented to Data and Discrimination Pre-conference at the 64th annual meeting of the International Communication Association, Seattle, WA, USA.

Sandvig, C., Karahalios, K., and Langbort, C. (2014b, July 22). Uncovering Algorithms: Looking Inside the Facebook News Feed. Berkman Center for Internet and Society Series Talk, Harvard University. Retrieved from http://cyber.law.harvard.edu/events/luncheon/2014/07/sandvigkarahalios

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2013). Re-Centering the Algorithm.
Paper presented at Governing Algorithms: A Conference on Computation, Automation, and Control, New York University. Retrieved from http://governingalgorithms.org/wp-content/uploads/2013/05/4
-response-karahalios-et-al.pdf

Sweeney L. (2013). Discrimination in Online Ad Delivery. *Communications of the ACM, 56* (5): 44-54. http://cacm.acm.org/magazines/2013/5/163753-discrimination-in-online-ad-delivery/fulltext