How to Throw the Race to the Bottom: Revisiting Signals for Ethical and Legal Research Using Online Data

Working Paper
Forthcoming, ACM SIGCAS Computers and Society
Special Issue on Security, Privacy and Human Behavior
(February, 2015)

Erin Kenneally[*] University of California San Diego
Center for Applied Internet Data Analysis
Center for Evidence-based Security Research
9500 Gilman Dr
San Diego, CA
ekenneally@ucsd.edu

ABSTRACT

With research using data available online, researcher conduct is not fully prescribed or proscribed by formal ethical codes of conduct or law because of ill-fitting "expectations signals" – indicators of legal and ethical risk. This article describes where these ordering forces breakdown in the context of online research and suggests how to identify and respond to these grey areas by applying common legal and ethical tenets that run across evolving models. It is intended to advance the collective dialogue work-in-progress toward a path that revisits and harmonizes more appropriate ethical and legal signals for research using online data between and among researchers, oversight entities, policymakers and society.

1. INTRODUCTION

The stability of trust on the Internet has implications for political diplomacy, innovation, economic stability, social and civil relations, and individual self-determinism. The degree of online trust is a reflection of the gap between individual and collective Netizens' expectations formed by laws and ethics, and their capabilities enabled by technology. Law and ethics, just as with familiar offline society, act as ordering forces that inform the acceptability of our behaviors and relationships with other person and organizations. The migration of analog activities online has exposed a rather sweeping gap between expectations and capabilities, where legal and ethical ordering forces are challenged to re-examine, -interpret, and -apply the tenets and principles upon which they moor. As this gap widens, so too does ambiguity between asserted rights, interests, and threats to same.

This gap between expectations and capabilities is manifest most prominently in the current industrial and geo-political struggle to define rules of engagement for cyber conflict and national security, as well as with online advertising and data brokering. A related context where ordering forces are challenged, lower on the public notoriety index but no less considerable, is network and computer security research. Specifically, the ability to easily collect and combine massive amounts of existing, "publicly-available" information of a sensitive nature (personal or confidential) online exposes deficiencies in the legal and ethical structures that directly and indirectly inform and reflect our expectations about the acceptability of online research.

Researchers increasingly encounter data online that may be beneficial, if not indispensable, to their research, such as: personal health, financial or behavioral records; usernames and passwords lists; corporate manuals and technical documents; email and voice communications databases; and, network traffic traces, device vulnerabilities and machine-to-machine communications. This data is located in various online locations ranging from normal websites and social networks to underground criminal forums, Internet relay chat rooms, and publicly-obscured (e.g., deep or dark web) sites.

[*] Erin Kenneally is a licensed Attorney specializing in information technology law, including legal risk, technology policy, information forensics and cybercrime, privacy technology, applied ethics, and trusted information exchange. She is a Technology Law Specialist at the University of California San Diego, and Founder and CEO at Elchemy, Inc.. Ms. Kenneally holds Juris Doctorate and Master of Forensic Sciences degrees.

And, its availability is often a product of malicious, negligent, or ignorant collection or disclosure by a third party.

What are researchers' responsibilities when they obtain sensitive information online that is a product of malicious (criminal theft or illicit fraud), negligent, or ignorant (data that were poorly secured or protected from public browsing) acquisition and/or disclosure? Does the calculation change when researchers use of collection (scanners, crawlers) and analysis (data mining, probabilistic reasoning) tools either magnify the quantity or quality of the sensitivities?

With research using data available online, researcher conduct is not fully prescribed or proscribed by formal ethical codes of conduct or law because of ill-fitting "expectations signals" – indicators of legal and ethical risk. This article describes where these ordering forces breakdown in the context of online research[1] and suggests how to identify and respond to these grey areas by applying common legal and ethical tenets that run across evolving models. It is intended to advance the collective dialogue work-in-progress toward a path that revisits and harmonizes more appropriate ethical and legal signals for research using online data between and among researchers, oversight entities, policymakers and society.

2. ISSUE SPACE: THE GAP BETWEEN ORDERING FORCES AND OUR EXPECTATIONS

"We don't have the norms, the rules of engagement, the rules of the road for how we and other countries should operate in this space." [13]

What can we infer from the fact that the ability to anonymously observe, collect and use new and existing online data without directly interacting with the subject of the data can equally characterize cyber espionage and surveillance by corporations and nation-states, targeted online advertising and data brokering by industry, and network and security research by public and private knowledge workers?

While it seems intuitive if not obvious that the three scenarios are distinguishable in terms of degrees of acceptability, they share a common thread– opaque (if not hidden) acts of potentially harmful data acquisition and usage, without normative or prescriptive procedures and revelations to the entities whose rights or interests may be negatively impacted. Law and ethics are the social ordering forces that direct our collective attention to harms and differentiates acts. When they are silent or unclear the risk of harms may be unattended or conflated, prompting us to revisit the legal and ethical calculus with good cause.

What is changing and challenging our legal and ethical ordering forces when it comes to online security research? The objectives of network and security research remain the same, such as generating theoretic and applied knowledge of networks, malicious threats and vulnerabilities; and developing new and improved cyber security products and strategies [8, 37, 7]. What *has* changed is the data substrate that researchers are engaging to reach those objectives: what *type of data* is involved (quantitatively and qualitatively), *how* and *where* the data is collected, *who* has interests in that online data, and what is the *impact* of secondary research use and disclosure.

The character of the data openly available online is often sensitive (private or confidential) and its original acquisition or disclosure online is the product of illicit or unclean hands. While such encounters are not unprecedented in research, the greater volume, dynamism and variety of data online stresses legal and ethical ordering forces that were designed for offline research. Moreover, the nature of the tools at researchers' disposal can magnify and even generate sensitivities, such as the case with automated collection and analysis tools like scanners, crawlers, data mining, and probabilistic reasoning techniques. The new data and tools invariably drive an opportunistic, not theory-driven research paradigm, accompanied by impacts of first impression.

To prohibit the collection and use of some sensitive information publicly available online for research is to threaten the empirical foundation of important public policy from infrastructure protection and law enforcement, to health and social welfare and commercial innovation. Yet equally undesirable is a regime of unbridled engagement with information that invokes persons' protected interests or rights under the cover of intermediary amnesty. One result is diminished trust and legitimizing the socially unacceptable acts that exposed the data originally.

2.1 Brave New Context

The difference between some incumbent legal and ethics ordering forces and current and forthcoming ones reflect the changed underlying contexts within which these forces originate. As a result, the expectation signals associated with incumbent standards can be incongruent with the reality of what researchers face online today. For one, the threat model to end user privacy is less bounded [29, 19, 35, 31] and implicates fewer individuals (i.e., sensitive data is more intermingled and subject to compound interests such as with social network sites [25]) than in offline data research contexts. The trigger for data protection obligations used to be active collection, but now because of the relative ubiquity and accessibility of data online, focus has shifted to acceptable *uses* [40] of data since it is often automatically and passively collected viz. machine-to-machine transactions. Also, the risk can be more emergent when seemingly disparate, non-sensitive public data is analytically combined to generate information that may expose sensitivities in unforeseen ways [15].

The meaning of "personally identifiable" data has shifted from pre-determined and static, to one where identity attributes are dynamic and encompass many more variables than biographical data typical of analog contexts [8, 39]. With regard to research purpose, traditional approaches were more static and specific whereas research online has fostered a more emergent approach to the purpose of collecting data

[1] This term "online research" is specifically not intended to address research study procedures where data is collected from individuals online in social networks or otherwise, such as is involved in behavioral manipulation, surveys, clinical interviews, structured tests, self-reports, deception, or ethnographic interviews.

since economical value and innovation is derived from creating derivative data and the anticipated value of subsequent, new purposes.

Online research also challenges traditional research tenets insofar as it enthuses opportunistic research that is data driven where there is increased incentive to engage short-term, innovative studies than long-term, directed research (not unlike what is occurring in industry with *big data* commercial opportunism). This fosters a different paradigm from offline research that leads with a hypothesis and is only then followed by data collection and theory validation. Finally, data quality risks demand a different focus in the new context where researchers are more challenged to assess the reliability and provenance of information available in open environments.

2.2 Common Scenarios That Expose Expectation-Capability Gaps

Wrapping our collective heads around a path(s) forward starts with realizing that what may seem like an issue of first impression is really just a recurrence of technology exposing fractures in our control structures, the colloquial "law lagging behind technology." This is manifest as a disconnect between the expectations of researchers, the public, and oversight entities about the ethical and legal propriety of the collection, use and disclosure of data available on the Internet. The disconnect is born out as tensions between and among individuals' rights and interests and the public good, where individual and corporate privacy face-off against forces of innovation, public safety (counter fraud and crime), and infrastructure security.

Increasingly, ordering forces are challenged to treat researchers differently than investigative journalists, private investigators and intelligence analysts, the government, and even the underground cyber vigilantes whom they study. The common research scenarios that expose these gaps include such things as:

1. Network layer information (e.g., maps, traffic, machine-to-machine (M2M) communications) about Internet-wide consumer and industrial vulnerabilities (e.g., open embedded devices in the energy, telco and transportation networks) is readily searchable and downloadable from a website as a result of port scanning from a distributed botnet of poorly protected embedded devices [10, 21].

2. Location information about individuals and corporate networks (e.g., subnets, hosts, open ports and banners) from different public sources (e.g., search engines, databases, public archiving sites like the Wayback Machine) are accessible using free open source tools [41, 26].

3. Personal private data (e.g., email addresses, names, device identifiers, financial account credentials, user name/password combinations, disease information) and business confidential manuals/technical docs leaked by negligent employees, fraudulent insiders or malicious hackers onto a publicly-accessible website and then collected by an automated script and posted openly on an open chatroom [34, 42, 28].

4. Links to a readily downloadable dumps of stolen credentials, corporate financial ledgers and billing data, goods and services price lists (e.g., stolen credit cards, accounts, botnets, cash out services) are posted openly on underground forums [11, 41, 37].

3. FRACTURED ETHICS SIGNALS

The ethical parameters for research using online data are ambiguous and contested. [27] Institutional Review Boards (IRB)[2] are often uncertain when it comes to their own policies about online research and lack uniformity in their collective advisement about what protections should apply to online research. The signals they have traditionally used to gauge ethical propriety are no longer stable indicators in the online context. IRBs anchor around whether the research involves acquiring private information about or otherwise interacting with a living individual. These signals– "identifiable", "private information", and "interaction"/"intervention"– become noisy when applied to online research. Is information collected by researchers on a social media site or available via a web crawl considered publicly available, for example? What if access to that site requires a user account with no special eligibility [43]? Does it change if payment or some other broad qualification is a condition to entry? The murkiness does not improve when we look to the traditional signals for expediting or exempting research from ethical oversight: if the data involves observing "public behavior" or collecting existing data; and the data collected is 'publicly-available" or "non-identifiable"; and its further disclosure does not cause harm.

The current application of these expectations signals are both under- and over-inclusive. For instance, interpretations of what is identifiable are strained in the context of M2M communications data that can predominate online security research data collection. Privacy-sensitive behavioral or profile information is generated by the proliferation of fingerprinting practices where device ID's, publicly observable IP addresses or URLs, and attributes are increasingly mapped to or represent individual users in databases in place of first order notions of PII [30, 5].

The scope of the ethics parameters are human-centric and haven't fully accounted for the extent to which information is an extension of our personhood and can cause harm to individuals, such as when systems and data that are distanced from users pose psychological, legal, economic, reputation and/or physical harm. Further, the application of informed consent is strained if not ill-fitting in these secondary use situations with research involving existing data online. This is because arguably the object of the research is the publication or system and not an individual person. And, other human subject protections related to the procedures for notifying users of the risks and benefits and ability to withdrawal are not applicable with secondary information that is not collected by the researcher from the subject-user. Also, oftentimes online data involves shared information that is not

[2] These are more broadly referred to as, "Ethical Review Boards" (ERB).

disclosed by every person implicated, so it is unclear from whom consent would be obtained even if it were required.

These uncertainties are recognized in the proposed changes to the Common Rule, such as the recommendation to exclude from oversight public information and data that is publicly-available[3], and that future use of pre-existing data should not require re-consent [32]. Even IRBs attempts to manage this gap by creating public-use datasets fall short because of the inconsistent treatment and definitional disagreement about secondary analysis of public-use data across IRBs [1, 2] ... not to mention the enormous quantitative and qualitative disadvantage to researchers from such a restriction.

4. FRACTURED LEGAL SIGNALS

Not unlike ethical parameters for research using online data, there are gaps between signals of legal risk and researchers' capabilities. Liability for use or disclosure of confidential information does not offer guidance because common law duty applies only to regulated contexts like doctor-patient and attorney-client. And, there is no contractual risk viz. nondisclosure agreement for researchers because there is lack of privity between them and the persons or organizations whose confidential data was accidentally or wrongfully published. Even if the researcher was contractually bound, information that is publicly known and made generally available in the public domain prior to disclosure is usually excepted under confidentiality provisions.

Invasion of privacy liability for use of online data for research is not triggered since the researcher is not the actor who originally exposed the user's private information online. Also, many privacy and identity theft-related claims require proof of harm. Subjects of misappropriated and leaked data have found nary little success overcoming this legal hurdle against the negligent or illicit actor, let alone against a researcher who may secondarily use or disclose what was published online. Increased risk of harm and fear of future injury has proven inadequate to meet legal damages requirements. Notwithstanding that challenge, proving that a researcher's secondary posting of the data actually caused the harm, when the data could have been collected and used by anyone online, renders that legal contour highly improbable.

The sector-specific laws offer poor signals as well. Researchers are not covered entities under various industry data protection laws that apply to organizations in the healthcare, financial services or commercial sectors, for example, so those flagship privacy protections offer no contours. Relevant communications privacy and computer trespass laws present definition application challenges, as key elements such as "unauthorized," "access," "interception" and "consent" are inapt [16, 3, 12] when researchers acquire data that is widely available on the web. Open questions abound, such as to what extent a researcher is authorized to collect or use online content in the absence of a website Terms of Service or from a URL on the deep web that is assumed obscured, or whether a researcher violates wiretap laws when s/he sets up a honeypot server that records traffic for research purposes without the consent of all the communicating parties.

Legal signals are strained furthermore in light of the myriad interests that may be at play when researchers avail themselves of data online. Namely at issue are the First Amendment rights of researchers to engage in research using this data free from censorship, as well as for the protection of institutional accreditation if academic freedom is restricted by oversight authorities. This match has already played out analogously in the private sphere when two commercial license plate recorder firms sued the Governor and Attorney General of Utah for impeding their claimed First Amendment right to collect data on license plates displayed in the public on open roads.

Other ancillary legal risk signals that may seem intuitively relevant offer little pre-/proscriptive guidance. For instance, novel application of the possession of stolen goods liability fails because mere possession of sensitive information such as passwords, usernames, or other credentials is not illegal. While mere possession of account numbers is illegal[4], researcher lack of intent to defraud eliminates this claim as a behavior-shaper. Trafficking-type charges miss the mark if there is no intent to transfer the data, which is likely the case with researchers who may secondarily collect dumps of passwords from public sites (e.g., chatrooms or Pastebin-like websites). Conspiracy and aiding--abetting breakdown because it is unlikely that researchers have actual knowledge whether the source acquisition of the sensitive data was illegal.

5. DERIVING NEW EXPECTATION SIGNALS TO MIND THE GAP

It is important to be mindful that the grey canvas that is ethics and legal risk with online research can be both a sword and shield. While aforementioned signals may fail to provide guidance to researchers, that is not dispositive of the ethics and legal risk they face. Researchers don't get a free pass because they purportedly act in the interests of enhancing generalizable knowledge for society. Instances where prosecutors, overseers, disputants, and the court of public opinion interpret these signals to the detriment of researchers are neither remote nor conjectural [6, 11, 42, 43, 4]. Having pointed out where current ordering forces breakdown in the context of research using online data, we briefly overview a strategy to kickstart pragmatic guidelines to deal with these grey areas. It involves identifying and synchronizing common legal and ethical tenets that run across current efforts to manage the capabilities-expectations gap in other non-research contexts.

The model involves first galvanizing the common parameters across ethics and law:

- *What* is the nature of the data engaged by researchers
- *Where* the data is acquired online and *how* it is collected

[3] Note that this recommendation is conditioned on whether the data subject has no "reasonable expectation of privacy" in that public data. This begs the question of what signals should define that condition.

[4] As an "unauthorized access device" under 18 U.S.C. 1029 Access Device Fraud

- *Who* has an interest in the data
- What is the *impact* of the secondary use and disclosure (i.e., the risk and type of harm involved, any mitigation measures, the purpose of research use)

Next, we can distill expectations signals from current and emerging legal and ethics ordering forces in issue spaces that may not necessarily be specific to online research, but whose purpose is to protect individuals whose interests are implicated by data in a manner proportional to the risk and in consideration of balancing interests. A sampling of some exemplar signals that could be applied to research involving online data include:

State and Federal Data Breach Laws and Regulations [14, 22]- These signals anchor on the nature of the "personally identifiable data" (PII) involved in data leakage and draws contours for what would trigger protection and notification obligations by the data holder, such as if sensitive financial account-related data is involved. They also address impact with risk of harm triggers related to the degree to which a breach is likely to compromise the security, confidentiality or integrity of sensitive personal information.

Similarly, the U.S. national healthcare law has replaced the former "risk of harm" standard for gauging impact with a four factor test that bases the data steward's obligations on the nature and extent of the "personal health information"(PHI), whether an unauthorized person was involved, whether the relevant data was actually acquired or viewed, and the extent to which any risk of exposure has been mitigated. Further, it entirely exempts de-identified data from oversight.

Secondary Use Health Data for Research (SHIP) [38]- These signals include assessing the privacy risks in the data and the likelihood of sensitive data being breached, the impact of a privacy breach, the reputational impact on the researcher, the researcher's motive, the public expectation of research use of the data, the public interest served by use of the data, whether the data is handled consistent with any relevant legal or ethical requirements, and whether the policies and procedures for collection and use are transparent.

Copyright Fair Use [18]- The Fair Use criteria may be helpful expectations signals for online research since it is specifically engineered to help reduce tensions between rights. The analog to copyright versus freedom of expression would be user privacy and confidentiality rights versus academic freedom of speech. Specifically, it looks at: the nature of the data (the analog to less protection afforded to factual works than creative works would be sensitive private or confidential information inuring more protection from research use than other data available online); the purpose and character of the data use (e.g., whether it is commercial or educational); how much of the (sensitive) data is used; and what the effect of using the data is on the value of the data.

FTC Act Section 5, 'Unfairness' standard [20]- If we substitute *data subjects* for *consumers* and *research* for *practices* , the relevant signals here attempt to gauge unfairness if a practice causes or is likely to cause harm to consumers, is not reasonably avoidable by consumers, and is not outweighed by countervailing benefits to consumers. This approach reflects a growing philosophy that shifts focus to responsible use of data and away from reliance on up front notice and consent. In the online research context, this argues for allowing the collection of identifiable data but restricting risky uses and disclosures.

EU Data Protection Act [17]- Some relevant signals in this broad-reaching law include allowing secondary information to be processed if it does not support actions or decisions with respect to particular individuals, if the processing does not put substantial distress on the data subject, and if research results are disclosed so that no individual is identifiable. As applied to the online research space, for instance, data found in a hacked database posted online might be permissible to use if it informs a study about the ecosystem of insurance benefit fraud but not not to investigate and report about specific individuals dodging the law.

Menlo and Companion Reports [24, 23]- This report and its supplement provide more directed guidance for online research since its primary motivation was to translate the first-order ethical principles and applications set forth in the authoritative guidance for general human subject research oversight by Institutional Review Boards (i.e., the Belmont Report and the Common Rule) to modern day information communication technology research. The signals it offers online research anchor around the broad principles of respect for persons, maximizing benefits and minimizing harms, justice and fairness, and respect for law and public interest.

Privacy Marketplace [33, 25, 9]- Although not formal ordering forces, nonetheless there is much value in inferring expectation signals from the deployment and adoption of privacy enhancing controls (products, services, policies) by website providers, consumer-users and industry.

6. IMPLICATIONS FOR SOCIETY

Just as the Sony-North Korea hacking scandal [36] illuminated for the national cyber conflict issue space, there is no consensus – community wide, nationally or globally– on what is acceptable research using data available online. The predominantly hypothetical, abstract scenarios that have prompted efforts to provide guidance will continue to give way to palpable instances where researchers are acting in a gray zone and will be subject to more exacting public scrutiny, liability and oversight.

This article approaches such research challenges and opportunities in the era of increased information availability with a summary initial conceptual model to understand, evaluate and address ethical and legal issues surrounding the use of online public data for research. It does so by abstracting common first order parameters and signals across ethical and legal ordering forces. This model is intended to foster novel cyber security research prospects while discouraging a race to the bottom – opportunistically exploiting or engineering logical vulnerabilities in our ordering forces for research. This approach can be used to anticipate public concerns and scientific needs in concert, promote transparency and accountability, and engender fair allocation of responsibilities and balancing of expectations among the research community, the public, funding agencies, and oversight entities (regulators, ERBs, law enforcement). These are the collective ingredients for effective stewardship of public funds and enhanced public trust in online research.

By outlining and developing a common understanding and conduct for ethical and legal research using online data we can avoid unattended harm and researchers enduring blowback by association that may result from undifferentiated comparisons to public or private surveillance or other adverse institutional exploitation of cyber capabilities.

7. ACKNOWLEDGMENTS

Support for this paper and its associated work is provided by the Center for Applied Internet Data Analysis (CAIDA) and the Center for Evidence-based Security Research (CESR) at the University of California San Diego under grants from the Department of Homeland Security and the National Science Foundation.

8. REFERENCES

[1] University of Michigan Human Research Protection Program. Research Using Publicly Available Data Sets: UM Policy. http://www.hrpp.umich.edu/initiative/datasets.html (revised May 2008).

[2] University of Washington. Human Subjects Division. Public Data Sets. https://www.washington.edu/research/hsd/topics/Public+Data+Sets.

[3] Cal. Penal Code Section 632 (California makes it a crime to record or eavesdrop on any confidential communication, including a private conversation, without the consent of all parties to the conversation).

[4] Updated Administration Proposal. http://www.whitehouse.gov/sites/default/files/omb/legislative/letters/updated-law-enforcement-tools.pdf (Executive Office of the President, proposed new laws against hacking that would arguably make it a felony to intentionally access unauthorized informationăeven if it has been posted to a public website, as well as to traffic in information like passwords, including posting a link) (visited January 15, 2014).

[5] Opinion 9/2014 on the Application of Directive 2002/58/EC to Device Fingerprinting. 14/EN, WP 224 (adopted on 25 November 2014). =http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp224_en.pdf.

[6] United States v. Aaron Swartz. 1:11-cr-10260, 106 (D. Mass. filed Jan. 14, 2013) (accessed at: http://s3.documentcloud.org/documents/217115/20110719schwartz.pdf.

[7] S. Afroz, V. Garg, D. Mccoy, and R. Greenstadt. Honor among thieves: A common's analysis of cybercrime economies. In *eCrime Researchers Summit (eCRS), 2013*, pages 1–11, Sept 2013.

[8] S. Afroz, A. Islam, A. Stolerman, R. Greenstadt, and D. Mccoy. Doppelganger finder: Taking stylometry to the underground. In *Security and Privacy (SP), 2014 IEEE Symposium on*, pages 212–226, May 2014.

[9] J. Bonneau and S. Preibusch. The privacy jungle: On the market for data protection in social networks. In *Economics of information security and privacy*, pages 121–167. Springer, 2010.

[10] Internet Census 2012, Port scanning /0 using insecure embedded devices. http://internetcensus2012.bitbucket.org/paper.html.

[11] D. Carr. A journalist-agitator facing prison over a link, September 2013. http://www.nytimes.com/2013/09/09/business/media/a-journalist-agitator-facing-prison-over-a-link.html?pagewanted=all_r=0.

[12] Computer Fraud and Abuse Act, 18 U.S. Code Section 1030 Fraud and related activity in connection with computers. Available at: http://www.law.cornell.edu/uscode/text/18/1030.

[13] M. Crowley. No rules of cyberwar. December 2014. Quote from Keith Alexander, former NSA Director and head of U.S. Cyber Command in response to the November 2014 Sony hacking incident.

[14] Security breach notification laws. Available at: http://www.ncsl.org/research/telecommunications-andinformationtechnology/securitybreachnotification-laws.aspx.

[15] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep.*, 3, 03 2013/03/25/online.

[16] The Electronic Communications Privacy Act of 1986 (ECPA), Pub. L. 99-508, oct. 21, 1986, 100 Stat. 1848 (1986). Available at: http://www.law.cornell.edu/topn/electronic _communications% $_{p}rivacy_act_of_$1986.

[17] Data Protection Act 1998. http://www.legislation.gov.uk/ukpga/1998/29/part/IV.

[18] 17 U.S. Code Section 107 Limitations on exclusive rights: Fair use. http://www.law.cornell.edu/uscode/text/17/107.

[19] J. Ferro, L. Singh, and M. Sherr. Identifying individual vulnerability based on public data. In *2013 Eleventh Annual International Conference on Privacy, Security and Trust (PST)*, pages 119–126, July 2013.

[20] Federal Trade Commission Act, Section 5, 15 U.S. Code 45 - unfair methods of competition unlawful. Accessed at: http://www.law.cornell.edu/uscode/text/15/45.

[21] S. Gorman. *Networks, Security and Complexity: The Role of Public Policy in Critical Infrastructure Protection*. Edward Alger Publishing, London, UK, 2005. Sean Gorman Ph.D thesis publication that documented SCADA vulnerabilities).

[22] Health Insurance Portability and Accountability Act of 1996 (HIPAA) Breach Notification Rule, 45 CFR 164.400-414. Available at: http://www.gpo.gov/fdsys/pkg/FR201301-

25/pdf/201301073.pdf.

[23] E. Kenneally and D. Dittrich. Applying Ethical Principles to Information and Communication Technology Research-A Companion to the Menlo Report. https://predict.org/Portals/Documents/Menlo-Report-Companion.pdf.

[24] E. Kenneally and D. Dittrich. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. https://predict.org/Portals/Documents/Menlo-Report.pdf.

[25] K. J. Lee and I.-Y. Song. Modeling and analyzing user behavior of privacy management on online social network: Research in progress. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 1344–1351, Oct 2011.

[26] K. Levchenko, A. Pitsillidis, N. Chachra, B. Enright, M. Felegyhazi, C. Grier, T. Halvorson, C. Kanich, C. Kreibich, H. Liu, D. McCoy, N. Weaver, V. Paxson, G. Voelker, and S. Savage. Click trajectories: End-to-end analysis of the spam value chain. In *Security and Privacy (SP), 2011 IEEE Symposium on*, pages 431–446, May 2011.

[27] A. Markham and E. Buchanan. Ethical decision-making and internet research recommendations from the aoir ethics working committee (version 2.0). December 2012. http://aoir.org/reports/ethics2.pdf".

[28] D. McCoy, H. Dharmdasani, C. Kreibich, G. M. Voelker, and S. Savage. Priceless: The role of payments in abuse-advertised goods. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 845–856. ACM, 2012.

[29] M. A. McGrath. Prescriber information and privacy: The costs of innovation in the healthcare industry. 2013.

[30] N. Nikiforakis, G. Acar, and D. Saelinger. Browse at your own risk. *Spectrum, IEEE*, 51(8):30–35, August 2014.

[31] S. e. a. Nirkhi. Analysis of online messages for identity tracing in cybercrime investigation. pages 300–305, June 2012.

[32] Proposed Revisions to the Common Rule for the Protection of Human Subjects in the Behavioral and Social Sciences, 2014. http://www.nap.edu/catalog/18614/proposed-revisionstothecommon%-rulefortheprotectionofhumansubjectsinthebehavioral-andsocialsciences.

[33] S. Peddinti, A. Korolova, E. Bursztein, and G. Sampemane. Cloak and swagger: Understanding data sensitivity through the lens of user anonymity. In *Security and Privacy (SP), 2014 IEEE Symposium on*, pages 493–508, May 2014.

[34] Open source intelligence tools list. (http://www.subliminalhacking.net/2012/12/27/osint-tools-recommendations-list/ (For example, various open source online intelligence tools like Pastebin, The Harvester, Shodan, Jigsaw, NetworkX Python).

[35] A. Ramachandran, L. Singh, E. Porter, and F. Nagle. Exploring re-identification risks in public domains. In *2012 Tenth Annual International Conference on Privacy, Security and Trust (PST)*, pages 35–42, July 2012.

[36] D. Sanger, December 2014. "U.S. Said to Find North Korea Ordered Cyberattack on Sony." http://www.nytimes.com/2014/12/18/world/asia/us-linksnorthkoreatosonyhacking.html.

[37] H. Sarvari, E. Abozinadah, A. Mbaziira, and D. McCoy. Constructing and analyzing criminal networks. 2014.

[38] Scottish Health Informatics Program. http://www.scotshiptoolkit.org.uk/.

[39] L. Sweeney. Simple demographics often identify people uniquely., 2000. Carnegie Mellon University, Data Privacy Working Paper 3. http://dataprivacylab.org/projects/identifiability/paper1.pdf.

[40] O. Tene and J. Polonetsky. Big data for all: Privacy and user control in the age of analytics, 2013.

[41] W. Wineberg. www.exfiltrated.com (a website showing security data related to ICS / scada systems, smart grid devices, and medical devices).

[42] K. Zetter. AT&T hacker "Weev" sentenced to 3.5 years in prison, March 2013. http://www.wired.com/threatlevel/2013/03/att-hacker-gets-3-years/.

[43] M. Zimmer. But the data is already public? on the ethics of research in Facebook. *Ethics and Information Technology*, 12(4):313–325, 2010.