Privacy and Security with Big Data

Large-scale information and communications technology (ICT) systems create new opportunities for social scientists, but realizing this potential requires strong partnerships with computer scientists.

Simson L. Garfinkel

ABSTRACT

Social science is transitioning from working with "made data" to "found data." Made data is the traditional of social science—data collected through experiments and surveys. Today the shift is to found data, such as administrative data not collected for research purposes, data collected from online social networks, and other so-called "big data" modalities. To use found data in a way that is both fair and accurate, social scientists need to involve computer scientists and, more broadly, information and communications technology (ICT) professionals to obtain, transfer, wrangle, organize, and store massive amounts of found data so that it can be used as a basis of objective research. But the shift to found data requires significant methodological innovations, and not just because of its size. Because of its diversity, special efforts must be taken to make found data findable by the broad range of potential users. In some cases, advanced formal privacy techniques such as differential privacy and secure multi-party computation are needed to work with found data in a manner that is ethically and logistically permissible. Efforts are also required to make studies involving found data transparent and replicable.

Beyond found data, our society's growing reliance on ICT is creating opportunities for social scientists to create made data on scales never before possible. Today there are technical infrastructures that make it possible (and common) for a single investigator to engage thousands or even hundreds of thousands of individuals in an experiment. Looking forward, there are opportunities to work with made data at the scale of 10^6 to 10^8 individuals—for example, by making deliberate changes to the technical infrastructure on which these individuals rely. Experiments have already been conducted showing that privileged social science investigators can covertly manipulate the emotions of people and change the outcome of elections. These experiments will require close partnerships with ICT professionals to assure technical accuracy and scientific validity. Moving forward, social scientists and ICT professionals must develop both appropriate technical controls and ethical frameworks to minimize the risk of these experiments to both the research participants and society at large.

For nearly a century social scientists have worked with data collected directly by their own surveys and using statistics produced by official statistics agencies. In both cases these data were typically collected under a promise of confidentiality—a promise mandated by law in the case of government agencies[1]—and the results subject to disclosure limitation[2] before they were made publicly available.

[1] For example, the confidentiality of information collected by the US Census is protected by Title 13 of the U.S. Code, which provides that information identifying individuals or businesses may not be published or otherwise disclosed, nor may be information be used "for any purpose other than the statistical purpose for which it is supplied." (13 USC §9 and §214) Information collected for statistical purposes by the Census Bureau, the Bureau of Labor Statistics, and other US government agencies is also protected by the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002 (Title V of Public Law

Increasingly social scientists are supplementing these traditional "made data" sources with "found data" such as administrative records and data that are organically generated and collected as part of our increasingly digitized society. For example, the Longitudinal Employer-Household Dynamics Program[3] operated by the US Census Bureau combines traditional census data products and surveys with administrative records supplied by states about workers and employers to create high-resolution data products that were previously impossible to conceive. The LEHD online tool OnTheMap[4] contains detailed information regarding commuting patterns of workers throughout the 50 states and the District of Columbia, allowing transportation planners to understand the source of traffic and congestion and to make accurate predictions about the impact of new transit options. The online tool OnTheMap for Emergency Management[5] "shows potential impact on jobs/workers and population for hurricanes, tropical storms, fires, floods, snow and freezing rain probability and disaster declaration areas. Real-time geographic data of disaster events are automatically updated."[6]

OnTheMap isn't just a management information system that collects data and displays it with an easy-to-use interface. Because the data that it displays are based on the data from individuals, the results must go through a disclosure limitation process involving the infusion of dynamically consistent noise[7] so that the contributions to the displayed data of an individual person's or company's contributions cannot be discerned.

Tools like OnTheMap show that important policy questions can increasingly answered with carefully designed tools that use a combination of made survey data and found administrative data. Looking forward, there's also a growing opportunity to use operational data as well. "Operational data" means the highly granular data that are used to operate systems. For example, whereas OnTheMap currently considers administrative data such as payroll tax data, commuting patterns could also be inferred by using real-time trip data collected by watching the movement of electronic toll payment systems (e.g. E-ZPass[8]) or Internet-enabled smart

---

107-347, "E-Government Act.") Data collected by federal agencies is also protected by the Privacy Act of 1974. In Europe, the confidentiality of data collected for official statistics is protected by Eurostat Principle 5, "Statistical confidentiality," which derives its authority from the Regulation 223/2009 on European Statistics, the European "statistical law." See http://ec.europa.eu/eurostat/.

[2] For information on statistical disclosure limitation within the US Government, see Statistical Policy Working Paper 22 (Second version, 2005), Report on Statistical Disclosure Limitation Methodology, Federal Committee on Statistical Methodology, December 2005. https://fcsm.sites.usa.gov/reports/policy-wp/. Current methods of statistical disclosure control are summarized well by George T. Duncan, Mark Elliot, Gonzalez Juan Jose Salazar. *Statistical Confidentiality: Principles and Practice*; Springer Science 2011. For a history of Statistical Disclosure Limitation, see *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*, George T. Duncan, Thomas B. Jabine, and Virginia A. de Wolf, Editors; Panel on Confidentiality and Data Access, National Research Council, ISBN: 0-309-57611-3, 288 pages. http://www.nap.edu/catalog/2122/

[3] LEHD Data, Center for Economic Studies (CES), US Census Bureau. https://www.census.gov/ces/dataproducts/lehddata.html

[4] http://onthemap.ces.census.gov/

[5] https://onthemap.ces.census.gov/em/

[6] LED (Local Employment Dynamics) New Data from the States and the U.S. Census Bureau, 2015. http://lehd.ces.census.gov/doc/LEDonepager_2015.pdf

[7] Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. 2008. "Privacy: Theory Meets Practice on the Map." In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. Red Hook, N.Y.: Curran Associates.

[8] Kashmir Hill, "E-ZPasses Get Read All Over New York (Not Just At Toll Booths)," Forbes, Sept. 12, 2013. http://www.forbes.com/sites/kashmirhill/2013/09/12/e-zpasses-get-read-all-over-new-york-not-just-at-toll-booths/

phones.[9,10] Indeed, operational data from both E-ZPass and smart phones are already being collected and used on a large-scale basis to detect surface traffic flow patterns.

Using collected administrative or operational data for social science research and other secondary purposes is fraught with ethical and logistic complications, but it is generally legal to do so. Indeed, such secondary uses typically happen without notice to or consent of the data subjects. Instead of getting consent, data brokers remove identifying information from the data before they are sold for use in research or for other purposes. For example, the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule[11] specifically states that de-identified health information is not considered protected health information (PHI) and, as such, is not protected by the privacy rule.[12] Likewise, the Family Educational Rights and Privacy Act (FERPA) specifically allows de-identified data to be "shared without the consent required by FERPA (34 CFR §99.30) with any party for any purpose, including parents, general public, and researchers (34 CFR §99.31(b)(1))."[13]

In recent years, de-identification techniques have come under attack by computer science researchers who have shown that many datasets that had been de-identified could be readily re-identified—that is, the individual data records could be matched back up with the true identities of the data subjects.[14] In most of these cases the data that had been provided to researchers had not been properly de-identified initially. But in other cases, researchers could re-identify the de-identified data records because the actual privacy afforded by the de-identification technique was simply not up to the task. Most often, this is because many de-identification techniques that have been employed have been heuristic techniques that focused on removing the appearance of personal information, but which lacked a theoretical basis—there was no proof that the technique provided the privacy protection that its users were expecting.

As social science researchers transition from working with "made data" to "found data," they need to partner with computer scientists to operationalize new approaches for working with data in ways that protect the sensitivity of these datasets. Fortunately, computer science is up to the task, with a wide range of techniques that have been developed over the past decade that can protect data while unlocking its potential.

---

[9] Jungkeun Yoon, Brian Noble, and Mingyan Liu. 2007. Surface street traffic estimation. In *Proceedings of the 5th international conference on Mobile systems, applications and services* (MobiSys '07). ACM, New York, NY, USA, 220-232. DOI=http://dx.doi.org.proxy.library.georgetown.edu/10.1145/1247660.1247686

[10] Mingqi Lv, Ling Chen, Gencai Chen, and Daqiang Zhang. 2014. Detecting traffic congestions using cell phone accelerometers. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication* (UbiComp '14 Adjunct). ACM, New York, NY, USA, 107-110. DOI=http://dx.doi.org.proxy.library.georgetown.edu/10.1145/2638728.2638744

[11] 45 CFR 160 and 45 CFR 164 subparts A and E; see also https://www.hhs.gov/hipaa/for-professionals/privacy/.

[12] See also "How Can Covered Entities Use and Disclose Protected Health Information for Research and Comply with the Privacy Rule?" US Department of Health and Human Services National Institutes of Health, https://privacyruleandresearch.nih.gov/pr_08.asp. Last updated February 2, 2007.

[13] "Data De-Identification: An Overview of Basic Terms," Privacy Technology Assistance Center, US Department of Education http://ptac.ed.gov/sites/default/files/data_deidentification_terms.pdf, last updated May 2013.

[14] For a discussion of several high-profile attacks on de-identified data sets, see Simson L. Garfinkel, NISTIR 8053: De-Identification of Personal Information, National Institute of Standards and Technology Interagency Report 8053, October 2015. http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf

Keeping Data Confidential

In 1986, John Diebold, one of the pioneers of the computer age wrote of how transactional data containing place and time information could become far more sensitive than first apparent. The case involved a bank that, in 1979 "had recently installed an automatic teller machine network and noticed 'that an unusual number of withdrawals were being made every night between midnight and 2:00 a.m.' . . . Suspecting foul play, the bank hired detectives to look into the matter. It turns out that many of the late- night customers were withdrawing cash on their way to a local red light district!"[15] An article about the incident that appeared in the Knight News Service observed: "there's a bank someplace in America that knows which of its customers paid a hooker last night."[16]

Compared with traditional social science data, administrative and operational data poses challenges resulting from both sensitivity and scale. With traditional surveys and laboratory experiments, respondents who don't want to take place in a research project have many means for opting out—for example, by not filling in a survey. But when it comes to administrative and operational data, opting out is harder. It's unlikely that a person will opt out of using their cellphone or credit cards on the chance that the resulting data might be used in as part of some economist's research project. Therefore, researchers who have access to such data need to adopt strong controls for preserving data confidentiality.

Countless news reports over the past two decades has shown that operational and administrative data is a powerful magnet for computer hackers, criminals, and even espionage. These data remain attractive when they are removed from their originating systems and transported to research systems for analysis. As such, researchers have an obligation to protect the data using mechanisms that are appropriate to the data sensitivity. Researchers who have been trained and become accustomed to working with publicly available datasets may require significant retraining and retooling when they start to work with datasets containing potentially sensitive information.

At a minimum, sensitive data should be stored on devices featuring full disk encryption to minimize the chance of data compromise when equipment is decommissioned[17] and to protect against the risk of compromise in the event of equipment theft. If encryption is not used, there should be written, audited procedures to control physical assets such as computers, hard drives, and portable storage media, to assure that these devices are appropriately sanitized or physically destroyed when they are retired from service.

But device- or file-level encryption alone is not sufficient to assure that confidently of sensitive data. Researchers that have sensitive data may become the targets of organized hacking efforts by criminals, hacktivists and even foreign governments. Motives for attack may include financial gain, embarrassing the organization that provided the data—or the researchers!—and obtaining information for attacking the data subjects themselves. There are many approaches to protecting against these threats, with different approaches being appropriate for different threat levels. At minimum, researchers should be using computers running up-to-date operating systems, with up-to-date anti-virus/anti-malware systems installed, and researchers should have mandatory and ongoing cybersecurity training. Systems should be professionally managed, and should have network connections that are monitored by a competent managed security

---

[15] John Diebold, in James Finn and Leonard R. Sussman, eds. *Today's American: How Free?* New York: Freedom House, 1986. p. 111

[16] ibid.

[17] S. L. Garfinkel and A. Shelat, "Remembrance of data passed: a study of disk sanitization practices," in *IEEE Security & Privacy*, vol. 1, no. 1, pp. 17-27, Jan.-Feb. 2003.

provider. Systems holding confidential data should only be accessible by those using two-factor authentication. In some cases, the data may be so sensitive that it should only be allowed on systems that are physically disconnected from the Internet—for example, a stand-alone laptop or an "air-gapped" network. Unfortunately, building and maintaining an air-gapped network is complicated,[18] and there are very few resources available with step-by-step instructions on how to create and maintain them.[19] A workable middle ground for many kinds of data is to build a "secure enclave" in which the data resides on secure servers that can only be reached through low-bandwidth interconnections. Users can log in to the secure enclave over the Internet, but all they can do is run queries and view results—they can't export the data.

Another kind of security that's needed are techniques that protect the data from the researchers themselves.

As Diebold's story demonstrates, sometimes it is the very artifacts that make data interesting that also make the data sensitive. Minor inconsistencies and variations can attract the attention of an inquisitive mind; drilling down reveals secrets. When the "secrets" are generalizable to a relevant population, we call this process social and economic *research*. Such inquisitiveness should be rewarded: the hallmark of a good researcher is following where the data takes you.

Because it is collected at large scale and for purposes other than furthering scientific research, administrative and operational data can take researchers to places where they don't belong, to secrets that are outside the scope of legitimate scientific inquiry, and cross into the realm of the private or even the prurient. Sometimes such temptations can be difficult to resist. In 2013, The Wall Street Journal featured an article detailing how analysts had used national security collection systems to spy on their love interests.[20] Similar abuses of official information systems occur at the state and local level: in 2012 Anne Rasmusson, a former police officer, filed a federal invasion-of-privacy lawsuit against the cities of Minneapolis and St. Paul, after 104 police offers illegally accessed her photo and other driver's license data.[21]

Strong internal security measures can protect data subjects from these kinds of inappropriate accesses to their data, and serve as a deterrent to help keep researchers on the correct side of proper ethical conduct. These security measures include column-level encryption,[22] so that some

---

[18] Eric Byres. 2013. The air gap: SCADA's enduring security myth. Commun. ACM 56, 8 (August 2013), 29-31. DOI=http://dx.doi.org/10.1145/2492007.2492018

[19] See, for example, Bruce Schneier, "Air Gaps," Schneier on Security, https://www.schneier.com/blog/archives/2013/10/air_gaps.html. One of the main challenges in operating an air-gapped network is installing updates to software packages, see "Setting up OS Deployment in an air-gapped network," IBM Knowledge Center, http://www.ibm.com/support/knowledgecenter/SS63NW_9.2.0/com.ibm.tem.life.doc_9.2/Lifecycle_Man/OSD_Users_Guide/c_osd_setup_airgap.html

[20] Siobhan Gorman, "NSA Officers Spy on Love Interests," *The Wall Street Journal,* Aug. 23, 2013. http://blogs.wsj.com/washwire/2013/08/23/nsa-officers-sometimes-spy-on-love-interests/. See also Andrea Peterson, "LOVEINT: When NSA officers use their spying power on love interests," *The Washington Post*, August 24, 2013. https://www.washingtonpost.com/news/the-switch/wp/2013/08/24/loveint-when-nsa-officers-use-their-spying-power-on-love-interests/?utm_term=.7a02a2ebdf1d

[21] Kim Zetter, "Female Cop Gets $1 Million After Colleagues Trolled Database to Peek at Her Pic," Wired, November 5, 2012. https://www.wired.com/2012/11/payout-for-cop-database-abuse/. See also Jessica Lussenhop, "Is Anne Marie Rasmusson too hot to have a driver's license?", City Pages, February 22, 2012. http://www.citypages.com/news/is-anne-marie-rasmusson-too-hot-to-have-a-drivers-license-6755567

[22] For example, T. Ge and S. Zdonik, "Fast, Secure Encryption for Indexing in a Column-Oriented DBMS," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, 2007, pp. 676-685. A

fields of the database are algorithmically locked to prevent unrestricted access, and secure audit logs to detect inappropriate data searches and browsing.[23]

Provenance

To use found data in a way that is both fair and accurate requires more than good security measures: it also requires a principled approach to data management and data processing.

Administrative and operational data can be messy. Data from production systems frequently contains format inconsistencies—errors that look like typos, for example. Depending on the data source, the data may contain actual typos. The data may contain transmission errors. And the data formats may change over time, as software changes in the field. Failure to handle these errors may result in systematic bias being introduced into the dataset, which may prime the unwary experimenter for false discovery.

As the amount of data increases, data management becomes steadily more important. By its very nature, operational data has the tendency to grow without limit. Researchers have to decide what data to keep and what to discard. Uncomfortable when placed in the position of having to make a decision about data retention, some researchers punt and keep it all. As a result, data tends to increase over time. And because researchers need multiple copies of data to protect against operator error and silent data corruption,[24] storage requirements tend to grow geometrically.

Researchers purchasing or developing large scale storage systems should investigate or develop approaches for automatically capturing provenance and storing it as part of their metadata. This includes the source of data, the techniques that were used to clean it, and the software responsible for creating statistical results. Collecting and analyzing provenance can also have unexpected benefits, such as helping to understand and improve database performance.[25]

---

general overview for less-technical readers can be found in Ashvin Kamaraju, "Database encryption demystified: Four common misconceptions," ZDNet February 9, 2012. http://www.zdnet.com/article/database-encryption-demystified-four-common-misconceptions/

[23] For example, see Gunnar Hartung. 2016. Secure Audit Logs with Verifiable Excerpts. In *Proceedings of the RSA Conference on Topics in Cryptology - CT-RSA 2016 - Volume 9610*, Kazue Sako (Ed.), Vol. 9610. Springer-Verlag New York, Inc., New York, NY, USA, 183-199. DOI: http://dx.doi.org/10.1007/978-3-319-29485-8_11; Di Ma and Gene Tsudik. 2009. A new approach to secure logging. *Trans. Storage* 5, 1, Article 2 (March 2009), 21 pages. DOI=http://dx.doi.org/10.1145/1502777.1502779.

[24] For information on silent data corruption, see Sumit Narayan, John A. Chandy, Samuel Lang, Philip Carns, and Robert Ross. 2009. Uncovering errors: the cost of detecting silent data corruption. In *Proceedings of the 4th Annual Workshop on Petascale Data Storage* (PDSW '09). ACM, New York, NY, USA, 37-41. DOI=http://dx.doi.org/10.1145/1713072.1713083; David Fiala, Frank Mueller, Christian Engelmann, Rolf Riesen, Kurt Ferreira, and Ron Brightwell. 2012. Detection and correction of silent data corruption for large-scale high-performance computing. In *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (SC '12). IEEE Computer Society Press, Los Alamitos, CA, USA, Article 78 , 12 pages; and Leonardo Bautista Gomez and Franck Cappello. 2014. Detecting silent data corruption through data dynamic monitoring for scientific applications. In *Proceedings of the 19th ACM SIGPLAN symposium on Principles and practice of parallel programming* (PPoPP '14). ACM, New York, NY, USA, 381-382. DOI=http://dx.doi.org/10.1145/2555243.2555279;

[25] Peter Macko, Daniel Margo, and Margo Seltzer. 2013. Performance introspection of graph databases. In *Proceedings of the 6th International Systems and Storage Conference* (SYSTOR '13). ACM, New York, NY, USA, , Article 18 , 10 pages. DOI=10.1145/2485732.2485750 http://doi.acm.org/10.1145/2485732.2485750

Users of commercial statistics packages tend to underestimate the amount of provenance that is required to reconstruct a finding. For example, it may be necessary to capture both the user's scripts (e.g. a Stata "do-file"), the version of the statistics package, the version of the host operating system, the model number of the computer's microprocessor (CPU), the time and date that the software was run, the time zone, the amount of random access memory (RAM) installed in the computer, and other seemingly benign information as well. Having all this information captured, stored and permanently recorded can be vital years later when one is trying understand why results can't be readily replicated. Tools such as the provenance aware storage system[26] can record provenance automatically as part of operating system operations, while toolkits like Digital Forensics XML (DFXML)[27] can store provenance at the application level. Approaches such as these need to be incorporated into workflows. One way to accomplish this would be to store provenance as user-defined attributes in Hierarchical Data Format Version 5 (HDF5) files.[28] Sadly, at the present time there is no native implementation that allows HDF5 files to be read from Hadoop or Apache Spark, but approaches have been proposed.[29]

Provenance that is automatically collected and indexed is also critical for making data findable in a big data research environment. With multiple researchers at multiple institutions, provenance can feed a search engine and respond to queries such as *find data collected between 2010 and 2014 that was processed with R to produce a table of household income vs. immunization rates*. Of course, in order to be findable, the search interface must be accessible to would-be downstream users. One way to assure this is through a technique called *federated search,*[30] in which multiple search engines are tied together so that a single search request can answered by dozens or hundreds of independent but federated search engines. Federated search can result in results that more both more relevant and more diverse[31] than queries to a single search engine, and approaches to address privacy issues have been developed.[32] Implementing

---

[26] See Kiran-Kumar Muniswamy-Reddy, David A. Holland, Uri Braun, and Margo Seltzer. 2006. Provenance-aware storage systems. In *Proceedings of the annual conference on USENIX '06 Annual Technical Conference* (ATEC '06). USENIX Association, Berkeley, CA, USA; and Kiran-Kumar Muniswamy-Reddy and Margo Seltzer. 2010. Provenance as first class cloud data. *SIGOPS Oper. Syst. Rev.* 43, 4 (January 2010), 11-16. DOI=http://dx.doi.org/10.1145/1713254.1713258

[27] S. L. Garfinkel, "Automating Disk Forensic Processing with SleuthKit, XML and Python," *2009 Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering*, Berkeley, CA, 2009, pp. 73-84. doi: 10.1109/SADFE.2009.12

[28] Mike Folk, Gerd Heber, Quincey Koziol, Elena Pourmal, and Dana Robinson. 2011. An overview of the HDF5 technology suite and its applications. In *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases* (AD '11), Peter Baumann, Bill Howe, Kjell Orsborn, and Silvia Stefanova (Eds.). ACM, New York, NY, USA, 36-47. DOI=http://dx.doi.org/10.1145/1966895.1966900

[29] Herd Heber, "From HDF5 Datasets to Apache Spark RDDs," The HDF Group, March 12, 2015. https://hdfgroup.org/wp/2015/03/from-hdf5-datasets-to-apache-spark-rdds/

[30] Dong Nguyen, Thomas Demeester, Dolf Trieschnigg, and Djoerd Hiemstra. 2012. Federated search in the wild: the combined power of over a hundred search engines. In *Proceedings of the 21st ACM international conference on Information and knowledge management* (CIKM '12). ACM, New York, NY, USA, 1874-1878. DOI=http://dx.doi.org/10.1145/2396761.2398535

[31] Dzung Hong and Luo Si. 2013. Search result diversification in resource selection for federated search. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '13). ACM, New York, NY, USA, 613-622. DOI=http://dx.doi.org/10.1145/2484028.2484091

[32] Wei Jiang, Luo Si, and Jing Li. 2007. Protecting source privacy in federated search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '07). ACM, New York, NY, USA, 761-762. DOI=http://dx.doi.org/10.1145/1277741.1277896

these concepts into an operational system can be assured through the use of the *privacy by design*[33] methodology.

Privacy Preserving Data Collection, Processing and Publishing

So far this article has been concerned with techniques developed by computer scientists for securely storing sensitive information and making the results of data processing findable. But computer scientists have also developed techniques over the past three decades that can be used for collecting, processing, and publishing sensitive information in a manner that preserves privacy.

Many of these techniques find their intellectual heritage in the work of Andrew Yao, a Chinese cryptographer who in 1982 introduced the concept of secure two-party computation, also called secure function evaluation. Yao developed a solution to the Millionaires' problem, in which two millionaires, Alice and Bob, engage in a two-person mathematical protocol that lets the them determine who is the richer of the two without reveal their wealth to each other or to a trusted third party.[34] Although the solution for two-party protocols is too complicated to present here, there is a simple three-party protocol that is presented below.

Consider the problem of Alice, Bob and Carol who wish to compute their average salary without revealing to each other how much each person earns. That is, Alice's salary is A, Bob's is B and Carol's is C, and the three wish to compute the value of (A+B+C)/3 without revealing A, B or to each other.

The solution to this problem is easy to understand. Alice, Bob and Carol each chose a random number—R_A, R_B and R_C, respectively. Alice sends to Bob the number A-R_A and to Carol the number R_A.) Neither Bob nor Carol have enough information to reconstruct Alice's number A. Bob, meanwhile, sends to Alice the number B-R_B and Carol the number R_B. Carol sends the number C-R_C to Alice and R_C to Bob. Finally, each of the players add together the numbers that they have received from the other two and write their sum on a whiteboard at the front of the room. Three numbers are now written on the board:

Alice wrote a single number AA that is (B-R_B) + (C-R_C).

Bob wrote a single number BB that is (A-R_A) + R_C.

Carol wrote a single number CC that is (R_A) + R_B.

---

[33] Privacy by design was introduced by Cavoukian, A. 201 Privacy by Design in Law, Policy and Practice. Information and Privacy Commissioner, Ontario, Canada; see also Stuart S. Shapiro. 2012. The state and evolution of privacy by design. In *Proceedings of the 2012 ACM conference on Computer and communications security* (CCS '12). ACM, New York, NY, USA, 1053-1053. DOI: http://dx.doi.org/10.1145/2382196.2382324 and Luke Stark, Jen King, Xinru Page, Airi Lampinen, Jessica Vitak, Pamela Wisniewski, Tara Whalen, and Nathaniel Good. 2016. Bridging the Gap between Privacy by Design and Privacy in Practice. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '16). ACM, New York, NY, USA, 3415-3422. DOI: http://dx.doi.org/10.1145/2851581.2856503. Other approaches for designing in privacy are explored in Sarah Spiekermann and Lorrie Faith Cranor. 2009. Engineering Privacy. *IEEE Trans. Softw. Eng.* 35, 1 (January 2009), 67-82. DOI=http://dx.doi.org/10.1109/TSE.2008.88.

[34] Yao, Andrew C. (November 1982). "Protocols for secure computations". *FOCS*. 23rd Annual Symposium on Foundations of Computer Science (FOCS 1982): 160–164. doi:10.1109/SFCS.1982.88.

The three numbers are now added and divided by three. This number, (AA+BB+CC)/3 is equal to $((B-R_B) + (C-R_C) + (A-R_A) + R_C + (R_A) + R_B) / 3 = (A + B + C) / 3$, which is the value that was to be computed! Thus, Alice, Bob and Carol have computed their average salary without revealing their individual salaries to each other.

This simple multi-party protocol depends on each player being honest and in not sharing their intermediate results with each other. This protocol is also not tolerant of accidental errors. But what the protocol does is enable a computation that would be otherwise impossible without a trusted third party.

Since Yao's discovery, many protocols and procedures for various kinds of security and privacy preserving computations have been discovered:

- *Private information retrieval*[35] is a family of procedures and protocols that allow for a user to retrieve data from a server without revealing which item is retrieved. The first scheme was introduced in 1997 by Kushilevitz and Ostrovsky; recently Yi, Paulet and Bertino published a survey of many techniques.[36]
- *Search on Encrypted Data* are techniques that allow for a client to conduct searches on an encrypted database, without revealing the contents of the encrypted documents or the search terms. Early work was done by Song, Wagner and Perrig;[37] follow-up work by others has discovered approaches for searches that are tolerant of minor misspellings,[38] searches that can rank keywords;[39] searching medical imagery;[40] and many other variants.
- *Oblivious RAM (ORAM)*, in which information is stored and retrieved from a remote server, but the server is unable to determine what is stored, what is retrieved, or the update patterns.[41]
- *Cryptographic Voting Protocols*,[42] in which participants vote and votes are tallied, but for which it is impossible to determine the vote of any specific voter. Many voting protocols have additional properties, such as the ability of voters to verify that their votes

---

[35] Kushilevitz, Eyal; Ostrovsky, Rafail (1997). "Replication is not needed: single database, computationally-private information retrieval". *Proceedings of the 38th Annual Symposium on Foundations of Computer Science*. Miami Beach, Florida, USA: IEEE Computer Society. pp. 364–373. ISBN 0-8186-8197-7.

[36] Xun Yi; Russell Paulet; Elisa Bertino, "Private Information Retrieval," in *Private Information Retrieval* , 1, Morgan & Claypool, 2013, pp.114; doi: 10.2200/S00524ED1V01Y201307SPT005

[37] Dawn Xiaoding Song, D. Wagner and A. Perrig, "Practical techniques for searches on encrypted data," *Proceeding 2000 IEEE Symposium on Security and Privacy. S&P 2000*, Berkeley, CA, 2000, pp. 44-55.

[38] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren and W. Lou, "Fuzzy Keyword Search over Encrypted Data in Cloud Computing," *2010 Proceedings IEEE INFOCOM*, San Diego, CA, 2010, pp. 1-5.

[39] C. Wang, N. Cao, J. Li, K. Ren and W. Lou, "Secure Ranked Keyword Search over Encrypted Cloud Data," *2010 IEEE 30th International Conference on Distributed Computing Systems*, Genoa, Italy, 2010, pp. 253-262.

[40] J. Yuan, S. Yu and L. Guo, "SEISA: Secure and efficient encrypted image search with access control," *2015 IEEE Conference on Computer Communications (INFOCOM)*, Kowloon, 2015, pp. 2083-2091.

[41] Oded Goldreich. 1987. Towards a theory of software protection and simulation by oblivious RAMs. In Proceedings of the nineteenth annual ACM symposium on Theory of computing (STOC '87), Alfred V. Aho (Ed.). ACM, New York, NY, USA, 182-194.

[42] Chris Karlof, Naveen Sastry, David Wagner, "Cryptographic Voting Protocols: A Systems Perspective," 14th USENIX Security Symposium, 2005.
https://www.usenix.org/legacy/event/sec05/tech/full_papers/karlof/karlof.pdf

were actually counted, or the ability to mathematically prove that the votes were properly counted.

- *Differentially Private Algorithms*, in which a queries executed over a dataset (for example, the average salary of workers in particular ZIP code) is reported in a way that any specific individual's contribution cannot be inferred with a degree of certainty. Differential Privacy was invented by Dwork, McSherry, Nissim and Smith in 2006;[43] since then, many algorithms have been developed for a wide range of queries and applications.[44] Most differentially private algorithms are based on the addition of Laplace noise to the results of queries: by carefully controlling the amount of noise added, these algorithms can produce results that are both reasonable accurate and reasonably privacy preserving. Although Differential Privacy was originally developed to support online query systems, it has also been expanded to include algorithms that generate synthetic data such that the contribution of specific individual's cannot be derived from the synthetic dataset.

Collectively, these approaches can be called *formal privacy techniques*, because the privacy assumptions and guarantees are formally stated and privacy loss or protection is mathematically proven. When using a formal privacy technique, is it important to understand how words like *privacy* and *security* are formally defined, since the defined meanings may be subtly different than the colloquial ones.

To date there have been few uses of secure multiparty computation techniques and formal privacy methods in practice. Key success stories include:

- On January 14, 2008, 1200 Danish sugarbeet farmers used secure multiparty computation in an auction that was used to set the price of sugarbeets. The auction lasted 30 minutes, and resulted in 25 thousand tons of production rights changing owner,[45] without the need for a trusted intermediary or auctioneer.
- In 2014, Google implemented a system called "Randomized Aggregatable Privacy-Preserving Ordinal Response" (RAPPOR) into the Chrome web browser. The system uses randomized response to collect statistics from users running Chrome web browsers, while maintaining user privacy with formal privacy guarentees.[46]
- As previously noted, The Census Bureau's On The Map[47] project uses noise infusion, as proscribed with a mathematical framework similar to that of Differential Privacy, to

---

[43] Calibrating Noise to Sensitivity in Private Data Analysis by Cynthia Dwork, Frank McSherry, Kobbi Nissim, Adam Smith In Theory of Cryptography Conference (TCC), Springer, 2006. DOI=10.1007/11681878_14

[44] The Algorithmic Foundations of Differential Privacy by Cynthia Dwork and Aaron Roth. Foundations and Trends in Theoretical Computer Science. Vol. 9, no. 3–4, pp. 211-407, Aug. 2014. DOI=10.1561/0400000042

[45] Peter Bogetoft, Dan Lund Christensen, Ivan Damgård, Martin Geisler, Thomas Jakobsen, Mikkel Krøigaard, Janus Dam Nielsen, Jesper Buus Nielsen, Kurt Nielsen, Jakob Pagter, Michael Schwartzbach, and Tomas Toft. 2009. Secure Multiparty Computation Goes Live. In *Financial Cryptography and Data Security*, Roger Dingledine and Philippe Golle (Eds.). Lecture Notes In Computer Science, Vol. 5628. Springer-Verlag, Berlin, Heidelberg 325-343. DOI=http://dx.doi.org/10.1007/978-3-642-03549-4_20

[46] Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security* (CCS '14). ACM, New York, NY, USA, 1054-1067. DOI: http://dx.doi.org/10.1145/2660267.2660348

[47] Machanavajjhala, A., D. Kifer, J. Abowd, J. Gehrke, and L. Vilhuber. 2008. "Privacy: Theory Meets Practice on the Map." In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering*. Red Hook, N.Y.: Curran Associates.

create a web-based query interface to highly confidential census data without the risk of disclosing sensitive information for individual census respondents.

Formal privacy techniques can (and should!) be combined with the other privacy protecting techniques discussed earlier in this article. For example, an organization that has sensitive data within a data enclave could use an algorithm like Ullman's Private Multiplicative Weights[48] to produce a differentially private synthetic dataset that is publicly distributed. Researchers could use this dataset to develop queries and to perform their initial data analysis. Once the code is working, the researchers could then provide their code to operators of the secure data enclave, who would then run the code on the actual data and review the results for inappropriate disclosures prior to making the results available to the researchers.

Surprisingly, differential privacy might also help solve the problem of false discovery when the same dataset is re-analyzed multiple times by different researchers. The problem occurs when a single publicly available dataset is used by different researchers for both exploration and hypothesis testing. A solution is to use the noise-infusion mechanism of differential privacy to allow the same holdout data set to be reused many times during data analysis.[49]

Experimenting on Thousands, Millions or even Billions from your Livingroom

Although most of this article has focused on found data, increasingly information and communications technology (ICT) is making it possible for individuals to carry out large-scale sociological experiments on thousands, millions or even billions of people. These capabilities have the potential to cause significant disruption to the way that many people do science.

Only a few short years ago, recruiting 500 or 1000 people for a survey or study could be a painstaking task. Today there are numerous services such as Mechanical Turk, CrowdFlower and Prolific Academic that let alone researcher create a web-based experiment or survey, ship it out to "workers," and get a response within a matter of hours. Prices for tasks can be so low—perhaps 20 cents per task—that student or even an enthusiastic amateur can fund an experiment on hundreds or thousands of people with their own pocket change—that is, without having to rely on institutional funds.

Experimenting with ones' own funds is problematic, because the primary mechanism by which human subjects research is regulated in the United States, the Institutional Review Board (IRB) system as defined by the National Research Act[50] and the Common Rule,[51] only apply to research that is federally funded. Although many universities require that all research involving human subjects performed under the auspices of a university be approved by the university's IRB, there is no general requirement that experiments on human subjects, even medical experiments, be approved by a disinterested, objective body.

---

[48] Jonathan Ullman. 2015. Private Multiplicative Weights Beyond Linear Queries. In *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems* (PODS '15). ACM, New York, NY, USA, 303-312. DOI=http://dx.doi.org/10.1145/2745754.2745755

[49] The reusable holdout: Preserving validity in adaptive data analysis
By Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, Aaron Roth, *SCIENCE* 07 AUG 2015 : 636-638

[50] The National Research Service Award Act of 1974, Public Law 93-348

[51] 45 CFR 46

Individuals who carry out social science research without oversight have the potential to "poison the well" of good will—and even societal tolerance—for social science research in general, just as so-called *push polls,*[52] also known as advocacy polls, have damaged the credibility, effectiveness, and even the legitimacy of traditional public opinion polls. But whereas sending out a push poll to several thousands of homes might costs thousands of dollars, sending out a survey by Mechanical Turk to thousands of respondents might cost less than $100.

Another way to experiment on thousands of people, or even more, is to package the experiment into an app and upload it to Apple's App Store or Google Play. An ethical experimenter might clearly disclose the purpose of the experiment, the protocol, and obtain the user's permission to proceed—a kind of informed consent. Alternatively, the app might simply offer some sort of compelling feature to its users, and perform the experiment covertly. Alternatively, subjects can be recruited through electronic mail. For a truly covert experiment, the email could employ deception and diversion.

Yet another way to experiment on people—thousands to millions—is to embed the experiment in a web service that's already being widely used. For example, the experiment could be installed within a popular social media platform, or packaged into a web-based advertisement that is show to thousands or millions of people.

In fact, all of these different options have been explored in recent memory, with many of them resulting in both published academic papers and resulting righteous indignation on the part of academics, commentators, and bloggers when the resulting experimental methodology became known:

- In 2005, researchers at Indiana University sent targeted email messages to 921 members of the University community who became experimental subjects *who did not know that they were part of an experiment*. The email, which used email addresses that were spoofed from another 810 community members, directed the recipients to a website where their Indiana University username and password was requested. After the password was verified, the community members were told that they had been successfully "phished" and were directed to another website where they were provided with security training. Even though the study had been approved by the university's IRB and its computer security team, the timing of the experiment at the end of the semester and the fact that subjects were involved without their permission resulted in substantial negative publicity both at the University and in the greater academic community.[53]
- In 2010, Facebook conducted a study on an astounding 61 million of its users to see if it could influence the outcome of an election by selectively mobilizing different segments of its user population to vote. Specifically, Facebook's researchers theorized that by showing clickable buttons in a user's newsfeed a clickable button saying "I Voted," a user's friends would be incentivized to vote because they knew that their trusted friends had voted. After it conduced its research, Facebook concluded that the answer was "yes," Facebook could influence the outcome of an election, by selectively showing the "I Voted" button or elements of the news feed to the specific subgroups that it chose to

---

[52] Marjorie Connley, "Push Polls, Defined," The New York Times, June 18, 2014. https://www.nytimes.com/2014/06/19/upshot/push-polls-defined.html

[53] For a writeup of the experiment, see Tom N. Jagatic, Nathaniel A. Johnson, Markus Jakobsson, and Filippo Menczer. 2007. Social phishing. *Commun. ACM* 50, 10 (October 2007), 94-100. DOI=http://dx.doi.org/10.1145/1290958.1290968. See also P. Finn and M. Jakobsson, "Designing ethical phishing experiments," in *IEEE Technology and Society Magazine*, vol. 26, no. 1, pp. 46-58, Spring 2007.

influence in specific elections where it was known that the public opinion was closely split.[54]

- In 2012, researchers at Facebook manipulated the "News Feed" feature of 689,003 Facebook users to see if such manipulations could alter the affective outlook of the Facebook users. The researchers wanted to see if they could facilitate the transference of an emotional state from one user to another. In fact, Facebook could. There was significant negative publicity after the study was published.[55]
- Between May 2014 and January 2015, researchers at Georgia Tech and Princeton purchased online advertisements that delivered an experimental payload 141,626 times to 88,260 distinct IP addresses in 170 countries, resulting in more than 1,000 measurements each in China, India, the United Kingdom and Brazil, and more than 100 measurements each in Egypt, South Korea, Iran, Pakistan, Turkey and Saudi Arabia. The purpose of the experiments was to measure the pervasiveness of web censorship on a country-by-country basis, and this measurement was performed by instructing the unwitting users' web browsers to attempt to visit sensitive or controversial web content that was known to be blocked for moral or political reasons in some of these countries.[56] The research received considerable notoriety, as some of the researcher's colleagues believed that initiating such requests from web browsers in some countries might expose the users to undue risk from their governments. This was all the more problematic because the users did not know that they were participating in a US-sponsored research experiment. This study was so controversial that the editors of the conference proceedings in which the article appeared felt compelled to include a boxed disclaimer statement on the first page of the article (see Fig. 1)

Statement from the SIGCOMM 2015 Program Committee: The SIGCOMM 2015 PC appreciated the technical contributions made in this paper, but found the paper controversial because some of the experiments the authors conducted raise ethical concerns. The controversy arose in large part because the networking research community does not yet have widely accepted guidelines or rules for the ethics of experiments that measure online censorship. In accordance with the published submission guidelines for SIGCOMM 2015, had the authors not engaged with their Institutional Review Boards (IRBs) or had their IRBs determined that their research was unethical, the PC would have rejected the paper without review. But the authors did engage with their IRBs, which did not flag the research as unethical. The PC hopes that discussion of the ethical concerns these experiments raise will advance the development of ethical guidelines in this area. It is the PC's view that future guidelines should include as a core principle that researchers should not engage in experiments that subject users to an appreciable risk of substantial harm absent informed

---

[54] Robert M. Bond, Christopher J. Fariss, Jason J. Jones, Adam D. I. Kramer, Cameron Marlow, Jaime E. Settle & James H. Fowler, A 61-million-person experiment in social influence and political mobilization, *Nature*, Vol. 489, p. 295, September 13, 2012. http://fowler.ucsd.edu/massive_turnout.pdf

[55] The original study was published at Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8788-8790. A comprehensive analysis of the event can be found in Jukka Jouhki, Epp Lauk, Maija Penttinen, Niina Sormanen and Turo Uskali, Facebook's Emotional Contagion Experiment as a Challenge to Research Ethics, *Media and Communication*, Vol 4, No 4 (2016) http://dx.doi.org/10.17645/mac.v4i4.579

[56] Sam Burnett and Nick Feamster. 2015. Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests. In *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication* (SIGCOMM '15). ACM, New York, NY, USA, 653-667. DOI: http://dx.doi.org/10.1145/2785956.2787485

consent. The PC endorses neither the use of the experimental techniques this paper describes nor the experiments the authors conducted.

Figure 1: Statement that appeared on the first page of "Encore: Lightweight Measurement of Web Censorship with Cross-Origin Requests," by Burnett and Feamster, 2015.

In 1979 the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research published the *Belmont Report*[57], named after the Belmont Conference Center in Elkridge, Maryland where the report was drafted. Written largely in reaction to the public disclosure of the Tuskegee Syphilis Study,[58] The *Belmont Report* laid the ethical groundwork for what became the National Research Act and the Common Rule. Today all researchers in the United States who perform human subjects research must be trained and tested on the report's three fundamental ethical principles when working with human subjects: Respect for persons; Beneficence; and Justice.

In recognition that research information ICT might require additional ethical principles, between 2010 and 2012 the US Department of Homeland Security convened a series of workshops with leading experts in the computing field to create a new report regarding the ethical conduct of research in the cyber age. Called *The Menlo Report*,[59] the new guide expanded the Belmont Report's original three ethical principles to include a fourth principle, "Respect for Law and Public Interest." This principle was added in recognition of the fact that experimentation on modern computer systems could result in significant damage to both those systems and the greater society.[60]

Conclusion

In the past, social scientists largely confined their work to datasets that they either made themselves or that they acquired from official statistics agencies. Today there is a growing interest in using data that results from administrative or operational systems. These datasets promise visibility into aspects of the economy and society that were never before available to social scientists, and offer resolution that was unimaginable until now. But these datasets also come with the risk of significant privacy violations and unknown data quality.

Techniques developed by computer scientists over the past four decades offer the promise of being able to work with these new datasets in a manner that is ethically appropriate and scientifically defensible. But these techniques, although they have been mathematically demonstrated and published in the peer reviewed literature, have not been operationalized to

---

[57] National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, Department of Health, Education and Welfare (DHEW) (30 September 1978). *The Belmont Report* (PDF). Washington, DC: United States Government Printing Office.

[58] "Tuskegee Study - Timeline". *NCHHSTP*. CDC. Page last reviewed December 22, 2015. Page last updated December 8, 2016. Page regtrieved December 17, 2016.

[59] D. Dittrich and E. Kenneally, "The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research", Tech. rep., U.S. Department of Homeland Security, Aug 2012. https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/

[60] For example, in 1988 Robert Tappan Morris created a computer worm that was designed to assess the size of the Internet, but resulted instead in disabling more than 6,000 Internet-connected computers, roughly 10% of the computers that were connected to the Internet at the time. See Cyberpunk: Outlaws and Hackers on the Computer Frontier, by Katie Hafner and John Markoff, Simon and Schuster, Nov 1, 1995. 396 pages.

the point that they can be trivially incorporated into production systems. Instead, social scientists need to work with computer scientists to move these techniques from the laboratory into practice. The benefits of doing so will be clear.