CYBERSECURITY

Saving Big Data from Itself

A three-step plan for using data right in an age of government overreach

By Alex "Sandy" Pentland

For the first few decades of its existence, the National Security Agency was a quiet department with one primary job: keeping an eye on the Soviet Union. Its enemy was well defined and monolithic. Its principal tools were phone taps, spy planes and hidden microphones.

After the attacks of September 11, all of that changed. The NSA's chief enemy became a diffuse network of individual terrorists. Anyone in the world could be a legitimate target for spying. The nature of spying itself changed as new digital communication channels proliferated. The exponential growth of Internet-connected mobile devices was just beginning. The NSA's old tools apparently no longer seemed sufficient.

In response, the agency adopted a new strategy: collect everything. As former NSA director Keith Alexander once put it, when you are looking for a needle in a haystack, you need the whole haystack. The NSA began collecting bulk phone call records from virtually every person in the U.S.; soon it was gathering data on bulk Internet traffic from virtually everyone outside of the U.S. Before long, the NSA was collecting an amount of data every two hours equivalent to the U.S. Census.

The natural place for the NSA to store this immense new haystack was the same place it had always stored intelligence assets: in the agency's own secure facilities. Yet such concentration of data had consequences. The private, personal information of nearly all people worldwide was suddenly a keystroke away from any NSA analyst who cared to look. Data hoarding also made the NSA more vulnerable than ever to leaks. Out-

IN BRIEF

Data about human behavior have always been essential for both government and industry to function. But how do we enable institutions to collect and analyze data without abusing that information?

We can start by embracing some basic principles. The NSA and other government organizations should leave big data resources spread across functionally separate databases overseen by separate organizations. Everyone who holds or shares personal data, including citizens, must safeguard transmission and storage through encryption.

In the digital era, we must also realize that existing policy and tradition will not suffice. Constant, transparent experimentation with big data procedures is the only way to find out what works.

Illustration by Bill Mayer

raged by the scope of the NSA's secret data-collection activities, then NSA contractor Edward Snowden managed to download thousands of secret files from a server in Hawaii, hop on a flight to Hong Kong and hand the documents over to the press.

Data about human behavior, such as census information, have always been essential for both government and industry to function. But a secretive agency collecting data on entire populations, storing those data in clandestine server farms and operating on them with little or no oversight is qualitatively different from anything that has come before. No surprise, then, that Snowden's disclosures ignited such a furious public debate.

So far much of the commentary on the NSA's data-collection activities has focused on the moral and political dimensions. Less attention has been paid to the structural and technical aspects of the NSA debacle. Not only are government policies for collecting and using big data inadequate, but the process of making and evaluating those policies also needs to move faster. Government practices must adapt as quickly as the technology evolves. There is no simple answer, but a few basic principles will get us on track.

Alex "Sandy" Pentland directs the M.I.T. Human Dynamics Laboratory and co-leads the World Economic Forum's big data and personal data initiatives. His latest book, *Social Physics*, was published in January by Penguin Press.

STEP

1

SCATTER THE HAYSTACK

ALEXANDER WAS WRONG about searching for needles in haystacks. You do not need the entire stack—only the ability to examine any part of it. Not only is it unnecessary to store huge amounts of data in one place, it is dangerous both for the spies and for the spied on. For governments, it makes devastating leaks that much more likely. For individuals, it creates the potential for unprecedented violations of privacy.

The Snowden disclosures made clear that in government hands, information has become far too concentrated. The NSA and other government organizations should leave big data resources in place, overseen by the organization that created the database, with different encryption schemes. Different kinds of data should be stored separately: financial data in one physical database, health records in another, and so on. Information about individuals should be stored and overseen separately from other sorts of information. The NSA or any other entity that has good, legal reason to do so will still be able to examine any part of this far-flung haystack. It simply will not hold the entire stack in a single server farm.

The easiest way to accomplish this disaggregation is to stop the hoarding. Let the telecoms and Internet companies retain their records. There need be no rush to destroy the NSA's current data stores, because both the content of those records and the software associated with them will quickly become ancient history.

It might be hard to imagine the NSA giving up its data-collection activities—and realistically, it will not happen without legislation or an executive order—but doing so would be in the agency's own interest. The NSA seems to know this, too. Speaking at the Aspen Security Forum in Colorado last summer, Ashton B. Carter, then deputy secretary of defense, diagnosed the source of the NSA's troubles. The "failure [of the Snowden leaks] originated from two practices that we need to reverse…. There was an enormous amount of information concentrated in one place. That's a mistake." And second, "you had an individual who was given very substantial authority to access that information and move that information. That ought not to be the case, either." Distributed, encrypted databases running on different computer systems would not only make a Snowden-style leak more difficult but would also protect against cyberattacks from the outside. Any single exploit would likely result in access to only a limited part of the entire database. Even authoritarian governments should have an interest in distributing data: concentrated data could make it easier for insiders to stage a coup.

How does distributing data help protect individual privacy? The answer is that it makes it possible to track patterns of communication between databases and human operators. Each category of data-analysis operation, whether it is searching for a particular item or computing some statistic, has its own characteristic pattern of communication—its own signature web of links and transmissions among databases. These signatures, metadata about metadata, can be used to keep an eye on the overall patterns of otherwise private communications.

Consider an analogy: When patterns of communication among different departments in a company are visible (as with physical mail), then the patterns of normal operations are visible to employees even though the content of the operations (the content of the pieces of mail) remains hidden. If, say, the person responsible for maintaining employee health records sees that an unusual number of these private records are suddenly being accessed by the financial records office, he or she can ask why. In the same way, structuring big data operations so that they generate metadata about metadata makes oversight possible. Telecommunications companies can track what is happening to them. Independent civic entities, as well as the press, could use such data to serve as an NSA watchdog. With metadata about metadata, we can do to the NSA what the NSA does to everyone else.

STEP

2

HARDEN OUR TRANSMISSION LINES

ELIMINATING the NSA's massive data stores is only one step toward guaranteeing privacy in a data-rich world. Safeguarding the transmission and storage of our information through encryption is perhaps just as important. Without such safeguards, data can be siphoned off without anyone knowing. This form of protection is particularly urgent in a world with increasing levels of cybercrime and threats of cyberwar.

Everyone who uses personal data, be they a government, a private entity or an individual, should follow a few basic security rules. External data sharing should take place only between data systems that have similar security standards. Every data operation should require a reliable chain of identity credentials so we can know where the data come from and where they go. All entities should be subject to metadata monitoring and investigative auditing, similar to how credit cards are monitored for fraud today.

A good model is what is called a trust network. Trust networks combine a computer network that keeps track of user permissions for each piece of data within a legal framework that specifies what can and cannot be done with the data—and what happens if there is a violation of the permissions. By maintaining a tamper-proof history of provenance and permissions, trust networks can be automatically audited to ensure that data-usage agreements are being honored.

Long-standing versions of trust networks have proved to be both secure and robust. The best known is the Society for Worldwide Interbank Financial Telecommunication (SWIFT) network, which some 10,000 banks and other organizations use to transfer money. SWIFT's most distinguishing feature is that it has never been hacked (as far as we know). When asked why he robbed banks, mastermind Willie Sutton allegedly said, "Because that's where the money is." Today SWIFT is where the money is. Trillions of dollars move through the network every day. Because of its built-in metadata monitoring, automated auditing systems and joint liability, this trust network has not only kept the robbers away, it has also made sure the money reliably goes where it is supposed to go.

Trust networks used to be complex and expensive to run, but the decreasing cost of computing power has brought them within the reach of smaller organizations and even individuals. My research group at the Massachusetts Institute of Technology, in partnership with the Institute for Data Driven Design, has helped build openPDS (open Personal Data Store), a consumer version of this type of system. The idea behind the software, which we are now testing with a variety of industry and government partners, is to democratize SWIFT-level data security so that businesses, local governments and individuals can safely share sensitive data—including health and financial records. Several state governments in the U.S. are beginning to evaluate this architecture for both internal and external data-analysis services. As the use of trust networks becomes more widespread, it will become safer for individuals and organizations to transmit data among themselves, making it that much easier to implement secure, distributed data-storage architectures that protect both individuals and organizations from the misuse of big data.

STEP

3

NEVER STOP EXPERIMENTING

THE FINAL and perhaps most important step is for us to admit that we do not have all the answers, and, indeed, there are no final answers. All we know for sure is that as technology changes, so must our regulatory structures. This digital era is something entirely new; we cannot only rely on existing policy or tradition. Instead we must constantly try new ideas in the real world to see what works and what does not.

Pressure from other countries, citizens and tech companies has already caused the White House to propose some limits on NSA surveillance. Tech companies are suing for the right to release information about requests from the NSA—metadata about metadata—in an effort to restore trust. And in May the House of Representatives passed the USA Freedom Act; though considered weak by many privacy advocates, the bill would begin to restrict bulk data collection and introduce some transparency into the process. (At press time, it is pending for the Senate.)

Those are all steps in the right direction. Yet any changes we make right now will only be a short-term fix for a long-term problem. Technology is continually evolving, and the rate of innovation in government processes must catch up. Ultimately, the most important change that we could make is to continuously experiment and to conduct small-scale tests and project deployments to figure out what works, keep what does and throw out what does not. SA

With metadata about metadata, we can do to the NSA what the NSA does to everyone else.

MORE TO EXPLORE

Personal Data: The Emergence of a New Asset Class. World Economic Forum, January 2011. www.weforum.org/reports/personal-data-emergence-new-asset-class

Social Physics: How Good Ideas Spread—The Lessons from a New Science. Alex Pentland. Penguin Press, 2014.

FROM OUR ARCHIVES

The Data-Driven Society. Alex "Sandy" Pentland; October 2013.

scientificamerican.com/magazine/sa

SCIENTIFIC AMERICAN ONLINE For more on big data and security, go to ScientificAmerican.com/aug2014/big-data