The darker side of data science

By: Sharon Weinberger – June 20, 2013

In 1975, the US Congress held a series of hearings to discuss a frightening new surveillance technology they feared would take away individual liberties. “Technological developments are arriving so rapidly and are changing the nature of our society so fundamentally that we are in danger of losing the capacity to shape our own destiny,” said John Tunney, a senator from California, during the opening of the first hearing. Lawmakers expressed fears that the new system would enable the federal government to amass large amounts of data on private citizens, cross-linking traffic violations with tax returns.

What was it that NBC News described as a “secret electronic intelligence network that gives the White House, the CIA, and the Defense Department Instant access to computer files on millions of Americans?” Arpanet, the then nascent computer network developed by the Pentagon’s research agency, which eventually led to the modern day Internet.

Now, almost four decades later, the US government is facing a new wave of criticism over a controversial programme that connects the data dots. Some of the issues being raised echo those heard in 1975; people are again asking whether the ability to collect massive amounts of data on human society compromises privacy.

At issue are recent revelations by the Guardian newspaper that the US government was collecting wide-ranging “telephony metadata” on its customers, including numbers called, time and duration of calls, and other identifying information. That was quickly followed by further revelations that the National Security Agency and FBI are also mining Internet traffic – everything from email to video chats.

The idea behind collecting such information is that modern data science can navigate a sea of information to alert law enforcement and intelligence agencies to any potential terrorist networks. But the debate is also about a broader issue: the emerging science of “big data,” which allows you to see connections and trends in large masses of information, and perhaps even predict types of behaviour.

“Big data is very useful, it can be incredibly revealing,” says Julian Sanchez, a fellow specialising in privacy and technology at the Washington DC-based Cato Institute. But the flip side, he says, is that big data can also be very revealing when it comes to what we think is private information.

The government’s argument is that the data it’s collecting is largely anonymised, and thus does not encroach on individual privacy. But that presumption rests on the idea that anonymised data truly is anonymous. Recent studies show that this may not be the case.

In a study published in Nature earlier this year, a group of researchers looked at the phone records of some 1.5 million mobile phone users in an undisclosed small European country, and found it took only four different data points on the time and location of a call to identify 95% of the people. “Hence, even coarse datasets provide little anonymity,” they concluded.

As soon as you have some piece of information, such as gender, language, or zip code, you can start to identify people, according Vincent Blondel, a professor at the Université catholique de Louvain, in Belgium, and one of the study’s authors. “It’s very hard to make things anonymous,” he says.

Full disclosure

If current projections are correct, there’s soon going to be much, much more data available. Abe Usher, chief innovation officer at US-based data and cybersecurity firm HumanGeo Group, points out that there will be more mobile devices than people by 2014, which creates what he call a “human sensor network.” While such information can be used by the private sector to target consumers, such as to up-sell someone a particular computer, Usher says “the real value is in understanding the aggregates and macrotrends” – the kind of information that could tell a large retailer where to place a store to maximise profits.

Big data is also growing in terms of public-sector support. Last year, the Obama administrationannounced the Big Data Research and Development Initiative, with $200 million going towards new ways to crunch numbers. Obama has also been surrounded by big data supporters; Rayid Ghani, the chief scientist on Obama’s campaign, is now taking big data to the nonprofit sector, where he hopes it might help with disaster relief, according to a profile in Technology Review. And recently, BusinessWeek revealed that Google’s Eric Schmidt is investing a company founded by Obama’s “big data gurus”.

How much of the government surveillance side of this is really new is not altogether clear, since much of the work, other than the recent leaks, is still a tightly held secret. “We can't know for sure, but this could well be a direct descendant, if not a continuation, of similar efforts in the Bush administration,” says Steve Aftergood, who writes the Secrecy News blog for the Federation of American Scientists. “There seems to be a more or less well-founded belief, shared by both administrations, that large volumes of data lend themselves to exploitation in new and productive ways.”

Whether all of this data is useful in the realm of national security is beside the point, argues Aftergood. Programmes like telephony collection are used for more than just investigating potential terrorist plots. “This is not some specific investigation that must be kept confidential, it's the cornerstone of an entire apparatus of surveillance,” he says.

The issue is making sure the debate over big data and privacy keeps up with the science. Yves-Alexandre de Montjoye, one of the authors of the Nature article, says that the ability to cross-link data, such as matching the identity of someone reading a news article to posts that person makes on Twitter, fundamentally changes the idea of privacy and anonymity.

“What we think is important to highlight is that it might be time for modern issues, such as digital privacy, open data, and net neutrality, to jump to the mainstream of politics,” de Montjoye says. “Ultimately, we need to know where politicians stand with respect to important 21st Century issues like this one.”