Alice E. Marwick
amarwick@gmail.com

Big Data, Data-Mining, and the Social Web

Talk for the New York Review of Books Event: Privacy, Power & the Internet
October 30, 2013

While recent revelations regarding the NSA's role in the collection and mining of the personal information and digital activities of millions of people across the world have garnered immense media attention and public outcry, there are equally troubling and equally opaque systems run by advertising, marketing and data-mining firms which have not attracted as much attention. Using techniques ranging from supermarket loyalty cards to targeted Facebook advertising, private companies systematically collect very personal information, from who you are, to what you do, to what you buy. Data about your online and offline behavior is combined, analyzed, and sold to marketers, corporations, governments, and even criminals. The scope of this collection, aggregation, and brokering of information is similar to, if not larger than, that of the NSA, yet it is almost entirely unregulated and many of the activities of data-mining and digital marketing firms creep under the radar.

Today I want to talk about two things: the *involuntary*, or passive, collecting of data by private corporations; and the voluntary, or active, collection and aggregation of *their own* personal data by individuals. While I think it is the former that we should be more concerned with, the latter poses the question of whether it is possible for us to take full advantage of social media without this playing into larger corporate interests.

Part One: Database Marketing

The industry of collecting, aggregating, and brokering personal data is known as *database marketing.* The second-largest company in this area, Axiom, has 23,000 computer servers which process more than 50 trillion data transactions per year, according to the *New York Times.*[1] It claims to have records on approximately 500 million Americans, including 1.1 billion browser cookies, 200 million mobile profiles, and an average of 1,500 pieces of data per consumer. This data includes information gleaned from publicly available records like home valuation and vehicle ownership, information about online behavior tracked through cookies, browser

[1] Natasha Singer, "Acxiom, the Quiet Giant of Consumer Database Marketing," *The New York Times*, June 16, 2012, sec. Technology, http://www.nytimes.com/2012/06/17/technology/acxiom-the-quiet-giant-of-consumer-database-marketing.html.

advertising, and the like, data from customer surveys, and "offline" buying behavior. The CEO, Scott Howe, says, "Our digital reach will soon approach nearly every Internet user in the US."[2]

Visiting virtually any website places a digital cookie, or small text file, on your computer. "First-party" cookies are placed by the site itself, such as Gmail saving your password so that you don't have to log in every time you visit the site. "Third party cookies" persist across sites, tracking what sites you visit, in what order. Google, Chrome, and Firefox sync this browsing history across devices, combining what you do on your iPad with your iPhone with your laptop. This is used to deliver advertising. For example, a few nights ago I was browsing LLBean.com for winter boots on my iPhone. A few days later, LLBean.com ads showed up on a news blog I was reading on my iPad. This "behavioral targeting" is falling out of fashion in favor of "predictive targeting," which uses sophisticated data mining techniques to predict whether or not I am likely to purchase something upon seeing an LLBean.com ad.

Axciom provides, quote, "premium proprietary behavioral insights" that "number in the thousands and cover consumer interests ranging from brand and channel affinities to product usage and purchase timing." In other words, Axciom creates a profile, or digital dossier, based on the 1,500 points of data it claims to have. This data might include your education level; how many children you have; the type of car you drive; your stock portfolio; your recent purchases; and your race, age, and education level. This data is combined across sources to determine whether you fit into a number of pre-defined categories such as "McMansions and MiniVans" or "adult with wealthy parent."[3] Acxiom is then able to sell these consumer profiles to their customers, who include 12 of the top 15 credit card issuers, seven of the top 10 retail banks, eight of the top 10 telecom/media companies, and nine of the top 10 property and casualty insurers.[4]

Axciom may be one of the largest data brokers, but they represent a very large shift in the way that personal information is handled online. The movement towards "Big Data," which uses computational techniques to find social insights in very large data sets, is rapidly transforming industries from health care to electoral politics. Big Data has a great deal of positive potential. But Big Data also poses new challenges to privacy on an unprecedented level and scale. Big data is made up of "little data," and this little data may be deeply personal.

Alone, the fact that you purchased a bottle of cocoa butter lotion is unremarkable. Target, on the other hand, assigns each customer a single Guest ID number, linked to their credit card number,

---

[2] Judith Aquino, "Acxiom Prepares New 'Audience Operating System' Amid Wobbly Earnings," *AdExchanger.com*, August 1, 2013, http://www.adexchanger.com/analytics/acxiom-prepares-new-audience-operating-system-amid-wobbly-earnings/.
[3] Natasha Singer, "A Data Broker Offers a Peek Behind the Curtain," *The New York Times*, August 31, 2013, sec. Business Day, http://www.nytimes.com/2013/09/01/business/a-data-broker-offers-a-peek-behind-the-curtain.html.
[4] Crunchbase, "Acxiom | CrunchBase Profile," October 25, 2013, http://www.crunchbase.com/company/acxiom.

email address, or name.[5] Every purchase and interaction you have with Target is then linked to your Guest ID, including the cocoa butter.

Now, Target has spent a great deal of time figuring out how to market to people about to have a baby. While most people remain fairly constant in their shopping habits—buying toilet paper here, socks there-- the birth of a child is a life change that brings immense upheaval. Since birth records are public, new parents are bombarded with marketing and advertising offers. So Target's goal was to identify people *before* the baby was born. The chief statistician for Target, Andrew Pole, said "We knew that if we could identify [new parents] in their second trimester, there's a good chance we could capture them for years."[6] So Pole had been mining immense amounts of data about the shopping habits of pregnant women and new parents. He found that women purchased certain things during their pregnancy, such as cocoa butter, calcium tablets, and large purses that could double as diaper bags.

Target then began sending targeted mail to women during their pregnancy. Unfortunately, this backfired. Women found it *creepy*—how did Target know they were pregnant? In one famous case, the father of a teenage girl called Target to complain that they were encouraging teen pregnancy by mailing her coupons for car seats and diapers. A week later, he called back and apologized; she hadn't told her father yet that she was pregnant.

So Target changed their tactics. They mixed in coupons for wine and lawnmowers with those for pacifiers and Baby Wipes. Pregnant women could use the coupons without realizing that Target knew they were pregnant. As Pole told the *New York Times*, "Even if you're following the law, you can do things where people get queasy."

These same techniques were used to great effect by the 2012 Obama campaign. Famously, the campaign recruited the best and brightest young minds in analytics and behavioral science, and put them in a room called "The Cave" for 16 hours a day.[7] The chief data scientist for the campaign, for example, was a former analyst who mined Big Data to improve supermarket promotions. This geek "Dream Team" was able to deliver micro-targeted demographics to Obama— they could predict *exactly* how much money they would get back from each fundraising email. When the team discovered that East Coast women 30-40 were not donating to the best of their ability, they offered a chance to have dinner with Sarah Jessica Parker as an

[5] "How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did," *Forbes*, accessed October 28, 2013, http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/.
[6] Charles Duhigg, "How Companies Learn Your Secrets," *The New York Times*, February 16, 2012, sec. Magazine, http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.
[7] Jim Rutenberg, "The Obama Campaign's Digital Masterminds Cash In," *The New York Times*, June 20, 2013, sec. Magazine, http://www.nytimes.com/2013/06/23/magazine/the-obama-campaigns-digital-masterminds-cash-in.html.

incentive.[8] Every evening, the campaign ran 66,000 simulations to model the state of the election. The Dream Team was not only using cutting-edge database marketing techniques, they were *developing* techniques that were far beyond the state of the art.

The Obama campaign's tactics illuminate something that is often missed in our discussions of datamining and marketing—the fact that governments are major clients of marketing agencies and databrokers. For instance, the campaign bought data on the television watching habits of Ohioans from a company called FourthWallMedia. Each household was assigned a persistent number, but the names of those in the household were not revealed. The Obama campaign, however, was able to combine lists of voters with lists of cable subscribers, which they could then coordinate with the supposedly anonymous ID numbers used to track the usage patterns of television set-top boxes.[9] They could then target campaign ads to the exact times that certain voters were watching television. As a result, the campaign bought airtime during unconventional programming, like *Sons of Anarchy, The Walking Dead* and *Don't Trust the B—- in Apt. 23,* rather than the conventional wisdom of local news programming.[10]

The "cave dwellers" were even able to match voter lists with Facebook information, using "Facebook Connect," Facebook's sign-on technology which powers many sign-ups and commenting systems online. Knowing that these users were Obama supporters, the campaign could figure out how to use them to persuade their perhaps less-motivated friends to vote. Crawling lists of Facebook friends and comparing them with tagged photos, the campaign matched these "friends" with lists of persuadable voters and then mobilized Obama supporters to convince their "real-life" friends to vote.

Part Two: Social Media

This brings me to the second part of my talk today: given these unbelievably sophisticated data-mining and analyzing techniques, is there any way we can use social media—or the internet itself—without adding to our profiles collected by companies like Axciom, Experian, or Epsilon?

Social media allows us to collect and track data about ourselves. For instance, I have been using a website called Last.fm since 2005 to track every piece of digital music I have listened to using iTunes or Spotify. As a result, I have a fascinating picture into how my musical tastes have changed over time, and Last.fm is able to recommend obscure bands to me based on this extensive listening history.

---

[8] Michael Scherer, "Inside the Secret World of the Data Crunchers Who Helped Obama Win," *Time*, November 7, 2012, http://swampland.time.com/2012/11/07/inside-the-secret-world-of-quants-and-data-crunchers-who-helped-obama-win/.

[9] Lois Beckett ProPublica et al., "Everything We Know (So Far) About Obama's Big Data Tactics," *ProPublica*, accessed October 28, 2013, http://www.propublica.org/article/everything-we-know-so-far-about-obamas-big-data-operation.

[10] Scherer, "Inside the Secret World of the Data Crunchers Who Helped Obama Win."

Using social media allows us to connect with friends; to learn more about ourselves; even to improve our lives. The Quantified Self movement, which builds on techniques used by women for decades such as counting calories, evangelizes the use of personal data for self-knowledge. Measuring your sleep cycles over time, for instance, can help you learn to avoid caffeine after 4pm, or realize that you can't use the internet for an hour before bedtime.

But this data is insanely beneficial to data brokers. Imagine how a health insurer might react to viewing your caloric intake on MyFitnessPal, the number of steps you walk per day tracked by FitBit, how often you check in to your local gym using Foursquare, and what you eat based on the pictures you post on Instagram. Each piece of information, by itself, may be inconsequential. But the aggregation of this information creates a larger picture that may be more than the sum of its parts. You probably don't remember what you ate for breakfast last Wednesday. But if you had tweeted about that, or posted a digital picture of it, this *digital instantiation* of formerly ephemeral information means that it can be combined with other pieces of information to create a bigger picture of yourself. And since we are perfectly capable of forming these impressions of others based on their social media activities, imagine what happens once data brokers gain access to this information and combine it with their already-extant databases.

I close with three larger concerns that I think deserve more attention.

The first is data discrimination. Once customers are sliced and diced into segmented demographic categories, they can be *sorted.* An Acxiom presentation to the Consumer Marketing Organization in 2013 placed customers into "customer value segments" and noted that the top 30% of customers add 500% of value, the bottom 20% actually *cost* 400% of value. In other words, it behooves companies to shower their top customers with attention, while ignoring the latter 20%, who may spend "too much" time on customer service calls, cost companies in returns or coupons, or otherwise *cost* more than they *provide.* These "low-value targets" are known in industry parlance as "waste." Joseph Turow, a University of Pennsylvania professor in Communications who studies niche marketing, asks what happens to those people who fall into the categories of "waste," entirely without their knowledge or any notification. Do they suffer price discrimination? Poor service? Do they miss out on the offers given to others? This discrimination is still more insidious because it is entirely invisible.

Second, we may be more concerned with the government than with marketers or data brokers collecting personal information, but this ignores the fact that government regularly *purchases* data from these companies. ChoicePoint, now owned by Elsevier, was an enormous data aggregator that combined personal data sourced from public and private databases, including social security numbers, credit reports, and criminal records. It maintained 17 *billion* records on businesses and individuals, which it sold to approximately 100,000 clients, which included 35

government agencies[11] and 7,000 federal, state and local law enforcement agencies.[12] For instance, the State Department purchased records on millions of Latin American citizens which was then checked against immigration databases. Choicepoint was also investigated for selling 145,000 personal records to an identity theft ring. More recently, Experian, one of the three major credit bureaus, mistakenly sold personal records to a Vietnamese hacker. Scammers refer to these records, which include social security numbers and mother's maiden name, as "fullz," f-u-l-l-z, because they contain enough personal information to apply for credit cards or take out loans.

Finally, a few years ago, I toured the experimental lab of a large advertising agency. They showed me the cutting edge of consumer monitoring technologies. Someday, not too far in the future, if you're at Duane Reade, aimlessly staring at a giant shelf of shampoo trying to figure out which shampoo to buy, the shelf will track your eye movements and which bottles you pick up and examine in more detail. Using this data, Duane Reade can algorithmically generate a coupon for a particular brand of shampoo, which you can print from the shelf. I watched an experimental application which tracks individual movements through a mall, based on the unique identifiers, or MAC addresses, of their cellphones, kept in purses or pockets, but available to wireless tracking devices. Again, in all of these cases, the individual is unaware that they are being tracked. This information may be hidden at the end of a byzantine privacy policy, or written on a notice next to a CCTV camera. While it may not be technically illegal, it feels ethically dubious.

While the easy answer to these problems is to opt-out of loyalty cards, internet use, or social media, this is hardly realistic. In fact, it is impossible to live life, online or offline, without being tracked without resorting to extreme paranoia. Cities track car movements; RFID tags are in clothing and dry cleaning; CCTV cameras are in every store.[13] The technology is developing far more rapidly than our consumer protection laws, which in many cases are out of date and difficult to apply to our networked world. The FTC and the Commerce Committee are currently investigating data brokers and calling for more transparency in the collection and dissemination of personal information. Those of us concerned with privacy must continue calling for checks and balances on these private corporations. I also encourage you to investigate the various opt-out tools, ad-blockers, and plugins that are available for most platforms. While the scrutiny of the NSA is necessary and needed, we must apply equal pressure to private corporations to ensure that seemingly harmless targeted mail campaigns and advertisements do not give way to far more insidious and dangerous personal privacy violations.

---

[11] Electronic Privacy Information Center, "Choicepoint," *EPIC.org Choicepoint*, August 19, 2013, http://epic.org/privacy/choicepoint/.
[12] "ChoicePoint," *Wikipedia, the Free Encyclopedia*, October 5, 2013, http://en.wikipedia.org/w/index.php?title=ChoicePoint&oldid=572910159.
[13] Sarah Kessler, "Think You Can Live Offline Without Being Tracked? Here's What It Takes," *Fast Company*, accessed October 28, 2013, http://www.fastcompany.com/3019847/think-you-can-live-offline-without-being-tracked-heres-what-it-takes.

Thank you.

Citation: Marwick, A. (2013). "Big Data, Data-Mining, and the Social Web." Governments, Corporations and Hackers: the Internet and Threats to the Privacy and Dignity of the Citizen. Power, Privacy and the Internet. New York Review of Books conference, Scandinavia House, New York, NY, October 30-31, 2013.