Inferences & Connections

by Alex Rosenblat, Tamara Kneese, and danah boyd

A workshop primer produced for:
The Social, Cultural & Ethical Dimensions of "Big Data"
March 17, 2014 - New York, NY
http://www.datasociety.net/initiatives/2014-0317/

Brief Description

Data-oriented systems are inferring relationships between people based on genetic material, behavioral patterns (e.g., shared geography imputed by phone carriers), and performed associations (e.g., "friends" online or shared photographs). What responsibilities do entities who collect data that imputes connections have to those who are implicated by association? For example, as DNA and other biological materials are collected outside of medicine (e.g., at point of arrest, by informatics services like 23andme, for scientific inquiry), what rights do relatives (living, dead, and not-yet-born) have? In what contexts is it acceptable to act based on inferred associations and in which contexts is it not?

Detailed Topic Description:

The appeal of "big data" comes from the supposed ability of data analysts to infer meaningful connections from data that has a high degree of volume, velocity, and variety. Reports on such analysis often downplay how challenging it can be to make sense of the available data. Size means little when the raw data are extraordinarily messy. Additionally, more data doesn't necessarily lead to more rational decisions or inferences, and data can be manipulated so that certain inferences are easy to make while others are less visible or available. Where is the human element in the inference-making process? What mechanisms could be in place to give more context to data collections and uses, such that the connections made through data analysis have a recognizable bias? How do we account for the accuracy of inferences when we rely on the data to reveal connections? When are data-driven connections spurious, and how do we correct for that?

Consider the challenge of determining the strength of interpersonal ties. Although sociologists have long calculated social tie strength based on surveys and observations, the availability of other sources of data is considered a tremendous boon to research. And yet, neither the articulated lists of shared connections on social media nor the communication and geographic traces available through cell phone records convey the full picture of our interpersonal connections. Although this information often provides valuable insight into people's lives, its meaning is not always easy to discern. For instance, just because teens defriend their parents on Instagram doesn't mean that their parents aren't important to

them. And just because employees communicate at phenomenal rates and share countless hours in close proximity to their bosses does not mean that that their work relationships are more significant than their other relationships, such as with their spouses, children, or friends.

In addition to the difficulties of *interpreting* data about interpersonal connections, the very *collection* of data is fraught with ethical challenges. For instance, when a database of genomic information is compiled by the police for law enforcement purposes, it has different implications than a similar database compiled with the intent of medical research. The Supreme Court recently ruled that police can collect the DNA of people they arrest, and that information can be entered into a local or national database, regardless of whether the arrested person is convicted of the crime they are suspected of committing. Particularly in light of the racialized imbalances in arrest rates, the collection of genetic data in a law enforcement context may amplify civil rights concerns about targeted or racial profiling for groups or communities that are surveilled or policed more heavily. Moreover, having familial records on criminal behavior through DNA collections can potentially lead to negative inferences about people who are related to those that are arrested. Does a familial record of suspected or proven criminal behavior resonate differently than a familial history of diabetes or Alzheimer's?

Large compilations of genomic or health data can advance medical research, but they can also pose serious personal consequences for the relatives -- living, dead, or not-yet-born -- of the individuals from whom data are collected. The family of Henrietta Lacks, whose HeLa cells are used for vast amounts of genetics research on cancer, Parkinson's, and other diseases, recently came to an agreement with the National Institute of Health. They enacted a policy that gives the Lacks family some control over third-party access to the full genome sequence data from HeLa cells. The basis for that policy originates with the publication of medical research that revealed the family's risks for developing certain diseases, to which they objected. How should access to genomic databases be controlled? What discrimination can result from data on a family's genetic predispositions? How can we control for potential privacy harms that may be immediate, or may occur several decades later?

Geo-locational data can also serve as the basis for inferred associations. Phone carriers can use their geo-location data to infer who is physically together, how often, and where they are; in addition, they know who calls whom, how frequently, etc. Other geo-location data result from voluntary broadcasting (for example, through sites like Rally, Twitter, or Foursquare) and is used by advertisers to market goods or discounts to people based on the stores they are passing by or at the venues they check-in to, either alone or with friends they are meeting with. Are performed associations like 'checking-in' different than the behavioral patterns or relationships between people that are inferred from shared geographical associations made evident from cell-phone carriers' data? Tracked location data and inferred associations can optimize the services that businesses can offer, the way that the application-based transportation service, Uber, anticipates where and when the greatest demand for its cars will be, and directs drivers to those areas, using its geo-location

data. In the context of law enforcement, using geo-location data to track or infer associations between people can have more tangible consequences than in a marketing or a service industry context. Does what AT&T, law enforcement, Facebook, or a fitness-tracking app knows about your location data have different implications for user privacy, or for ethical considerations? How important is contextual privacy to the collection or sharing of data from which inferences and associations are made?

Finally, inferred associations can be made about individuals based on their web use -- search queries, health apps, Facebook comments, and many other sources. Websites and applications that both collect user data and promote spaces for social exchanges mediate information flows not only between individuals and big organizations, but also on a peer-to-peer scale. A Facebook Newsfeed algorithm, for example, can affect who sees which friends' status updates, who 'likes' what, what commentaries they make, what links they post, and other relational paraphernalia that inform and direct how people socialize both online and offline. Individuals as well as big data aggregators can infer and assign connections to people based on the activities they see generated in the forms of 'likes' or other signals, which can sometimes be misleading. Facebook has come under scrutiny by irate users for artificially 'liking' items on users' behalves, and without their consent, and advertising that 'like' to users' friends. What responsibilities do entities that collect data, and which impute connections to people, have to those who are associatively implicated? Are data available in a forum like Facebook considered 'public', and is it ethical to use that data for research without the explicit consent of the people whose data is being used? Is accountability for the impact of these inferences and connections negotiated differently in a health care, law enforcement, or social media context?

Not only are we more visibly connected to broader swathes of networks now through social media, genetic information databases, DNA records, purchasing histories, etc., but individual control over what is shared, and with whom, is a subject ripe for debate. How can we collectively come to terms with what is known about us due to data-oriented practices? How much do we trust the platforms we interface with to draw the right sorts of connections or inferences about us? Who gets a say over how the data is used, or what the consequences or benefits of inferences are? What are the power dynamics to consider when data-driven decisions are made? How should different domains, like health, marketing, education, or criminal justice, evaluate the risks and benefits of drawing inferences from the data they have? What does "big data" replace in decision-making processes?

Case Study 1: Opting Out of Inferences

Most likely, even if you've chosen not to join Facebook, Facebook knows who you are, who you know, and some of your interests. Your friends may have uploaded their address books, including your email addresses, to the Facebook servers. They may have uploaded photos in which you were present, tagging you in the hopes that you'll join later. (Unfortunately, if you're not on the site, you can't untag yourself.) Not only that, but because Facebook tries to predict who's in a photo, it will learn about you from the photos

that you're present in when one of your friends shares it, and inferring other relational connections that your friend didn't explicitly identify. Facebook also tries to guess if you might be associated with a particular university or if you have particular tastes. If you eventually do decide to join the site, Facebook will try to make the onboarding process easier by asking you to confirm what they've already ascertained.

The inferences made by early social media sites use to be quite inaccurate. For example, it was quite common for participants on Friendster to receive recommendations to "friend" their exes. Algorithmically, the reason was logical. Exes are inevitably people you haven't friended, but with whom you have a lot of shared connections. Although these algorithms have improved, people still encounter uncomfortable situations of this sort.

People use the multiplicity of data they gather from indirect and direct social contact to determine the bounds of their relationships; and how people socialize is deeply affected by their data practices, and privacy controls. For example, a person who does not share their geo-location data with a set of friends may be excluded from spontaneous events organized and attended by the others. Other times, individuals may have private information exposed as a function of being plugged into multiple online communication forums. One transgender woman was recently outed to a colleague after Google integrated Google+ into its Android operating system, and shared information from her Google+ profile with her SMS contacts.

The types of information that is collected about individuals have implications for the groups and networks they are connected to, by varying degrees of separation. Algorithmic inferences are a cornerstone of the "big data" phenomenon, but they go against the mantra of being able to "opt out" as a mechanism for achieving personal control of one's data. Indeed, it is very difficult to opt out of inference-based algorithms, let alone the explicit sharing of one's friends. More often than not, users don't even know what produced the association, let alone how to make it stop. When individual control is not viable, who is responsible for holding those making the inferences accountable?

Although most associations are harmless, inferences can be particularly dicey for those from more marginalized backgrounds. In theory, in a democratic country with a strong commitment to presumed innocence, guilt through association should not be a legal problem. Yet, in reality, many people find themselves judged based on those that they know, both formally and informally. How can and should people protect themselves from problematic inferences, particularly when such inferences are not necessarily visible or obvious to the individuals who are affected by them?

Case Study 2: Genetic Connections

In Maryland v. King, the Supreme Court ruled that DNA samples are a legitimate and reasonable part of an arrest under the 4th Amendment, equivalent to taking someone's wallet, fingerprints, or photographs. These practices are considered legitimate in the context of a criminal arrest because they act as identifiers. Indeed, the practice of collecting

fingerprints during an arrest dates back to antiquity when fingerprints were also used to sign and seal documents.

At first blush, one might see DNA as an extension of fingerprints because genetic material can serve as a (mostly) unique identifier. (Of course, this is only somewhat accurate; identical twins may have identical DNA sequences and a person's DNA can change over time.) But collecting genetic information implicates more than that individual because people's genetic codes are similar to those to whom they are related. By collecting DNA, police databases effectively gather information about a person's parents, siblings, and not-yet-born children and grandchildren. What rights do these people have over the data collection that is taking place, or the future use of that data?

Given racialized imbalances in arrest rates, does the collection and aggregation of such data affect or exacerbate racialized approaches to predictive policing? Given that blacks and Latinos are more likely to be arrested, are other blacks and Latinos more likely to find that they are connected to people who have been previously arrested? How are these data used and what inferences are drawn when arrested individuals are connected with others in the database? What are the implications of associating genetic information with possible criminal activity?

Case Study 3: Educational Inferences

Data can be used to support inferences about individuals' propensity for future behavior, based on their algorithmic association with other people. These associations can have real consequences to those individuals: for instance, schools may use tracking systems to identify students for certain types of treatment, based on their characteristics and/or previous course of behavior.

Tracking methods have been employed by specific school systems in order to lessen the dropout rate, attempting to target the most vulnerable students. IBM used analytic models to reduce the student dropout rate in Mobile County, Alabama. Tracking methods helped to red-flag struggling students, notifying educators when they should step in to formulate individualized plans of action for addressing the needs of students who might otherwise fall through the cracks. Academic information was combined with demographic data, allowing researchers to determine a model for predicting dropout risk. While not legally bound to do so, Mobile County Public Schools helped to alleviate privacy concerns by requiring parental consent to use student information and by allowing parents to access the data through home computers or handheld devices. Researchers took pains to protect the data and were aware that findings could be misconstrued or that students could be falsely identified as dropout risks. IBM and the school system also addressed the potential problems with placing students in particular educational tracks, advising administrators and teachers on the limits of predictive models and on the need for educated interpretation.

Although these teams took significant measures to reduce problematic inferences, we can imagine that a number of issues can arise as data analytics proliferate in the educational

sector. The implementation of data analytics in education has potential benefits, but may also peg individual students as "problems" early on and affect what classes they are allowed to take in the long term. Giant databases containing student information have raised privacy concerns among officials, parents, and educators alike. In New York State, a celebrated high school principal argued that the "wall of privacy" promised by data anonymization had been shattered. InBloom required students' names, addresses, email addresses, and phone numbers to the system while also asking schools to submit student attendance records and codes. Because of this, details regarding student illness or disciplinary action would be revealed and associated with individual students. As a result of these problems, many educators and parents are protesting systems such as inBloom, prompting the Electronic Privacy Information Center to sue the U.S. Department of Education over the matter. While tracking methods appeared to reduce dropout rates in Mobile, what are the boundaries? What happens when truancy and suspension data circulates outside of school walls? How will this information be used and by whom? If data are used to target individual students or in any way jeopardize their academic, personal, or work-related futures, what recourse do parents, students and concerned educators have? Can data analytics inadvertently penalize already marginalized students and school systems?

Questions to Consider

- What are the major social, cultural, and ethical tensions that emerge when connections are inferred? What needs to be better understood to address what's happening?
- What conflicting values and tradeoffs are at stake? How do we understand relevant actors, stakeholders, and "camps"?
- How are the opportunities and challenges of inferred associations different in different domains (e.g., social media vs. healthcare vs. criminal justice)?
- What are additional salient case studies that highlight the tensions, tradeoffs, and issues?
- Who should be holding algorithms accountable when associations are inferred? What is the role of the government? Of data providers? Of technologies and tools? Of educational institutions? Of media institutions?
- What kinds of privacy harms are attached to imputed connections?
- When can or should an individual make a reasonable claim against the collection or aggregation of their data for a 'collective' purpose, like progressing medical research? What allows them to do so effectively?
- How are not-yet-born individuals affected by the data collected on them or their families? What rights should they have?
- Which organizations have the right to act on inferred associations, and what kind of data should they use to make those inferences?