THE CHRONICLE OF HIGHER EDUCATION

TECHNOLOGY

Harvard Researchers Accused of Breaching Students' Privacy

Social-network project shows promise and peril of doing social science online

Jon Chase, Harvard U.

Jason Kaufman, of Harvard's Berkman Center for Internet & Society, says critics of his research on student Facebook profiles are acting like "academic paparazzi."

By Marc Parry | JULY 10, 2011

In 2006, Harvard sociologists struck a mother lode of social-science data, offering a new way to answer big questions about how race and cultural tastes affect relationships.

The source: some 1,700 Facebook profiles, downloaded from an entire class of students at an "anonymous" university, that could reveal how friendships and interests evolve over time.

It was the kind of collection that hundreds of scholars would find interesting. And in 2008, the Harvard team began to realize that potential by publicly releasing part of its archive.

But today the data-sharing venture has collapsed. The Facebook archive is more like plutonium than gold—its contents yanked offline, its future release uncertain, its creators scolded by some scholars for downloading the profiles without

students' knowledge and for failing to protect their privacy. Those students have been identified as Harvard College's Class of 2009.

The story of that collapse shines a light on emerging ethical challenges faced by scholars researching social networks and other online environments.

The Harvard sociologists argue that the data pulled from students' Facebook profiles could lead to great scientific benefits, and that substantial efforts have been made to protect the students. Jason Kaufman, the project's principal investigator and a research fellow at Harvard's Berkman Center for Internet & Society, points out that data were redacted to minimize the risk of identification. No student seems to have suffered any harm. Mr. Kaufman accuses his critics of acting like "academic paparazzi."

Adding to the complications, researchers like Mr. Kaufman are being asked to safeguard privacy in an era when grant-making agencies increasingly request that data be shared—as the National Science Foundation did as a condition for backing Harvard's Facebook study.

The Facebook project began to unravel in 2008, when a privacy scholar at the University of Wisconsin at Milwaukee, Michael Zimmer, showed that the "anonymous" data of Mr. Kaufman and his colleagues could be cracked to identify the source as Harvard undergraduates.

"The steps that they tried to take to engage in innovative research, to me fell short," says Mr. Zimmer, an assistant professor at Milwaukee's School of Information Studies and co-director of its Center for Information Policy Research. "It just shows that we have a lot of work to do to make sure that we're doing this kind of research correctly and in ways that don't jeopardize the subjects that we're studying."

The controversy over the Harvard data set, known as "Tastes, Ties, and Time," comes amid growing interest in social-network research across disciplines, including sociology, communications, history, geography, linguistics, business, computer science, and psychology. The daily minutiae of our digital lives are so culturally valuable that the Library of Congress is on the eve of opening a research archive of public tweets.

"If you had to dream of research content, it would be sending out a diary and having people record their thoughts at the moment," says Alex Halavais, an associate professor of communications at Quinnipiac University and soon-to-be president of the Association of Internet Researchers. "That's like a social scientist's wet dream, right? And here it has kind of fallen on our lap, these ephemeral recordings that we would not have otherwise gotten."

But that boon brings new pitfalls. Researchers must navigate the shifting privacy standards of social networks and their users. And the committees set up to protect research subjects—institutional review boards, or IRB's—lack experience with Web-based research, Mr. Zimmer says. Most tend to focus on evaluating biomedical studies or traditional, survey-based social science. He has pointed to the Harvard case in urging the federal government to do more to educate IRB's about Web research.

'Complete Social Universe'

The project at the center of this dispute dates to Facebook's younger days. Even then the Harvard-born network was on its way to conquering American higher education. In 2006, with clearance from Harvard's IRB and Facebook, Mr. Kaufman's team began dipping into the profiles of one class to build a data archive for social-science research.

The researchers downloaded each student's gender, home state, major, political views, network of friends, and romantic tastes. To determine race and ethnicity,

they examined photographs and club affiliations. They recorded who appeared in students' photo albums. And they culled cultural tastes like books, music, and movies (top film: *Lord of the Rings*).

The archive was built to feed a team of five sociologists—four from Harvard, one from the University of California at Los Angeles—whose research interests include culture, race, and public health. Their push to vacuum up so many Facebook profiles helped overcome a big obstacle to social-network research: getting enough data. Typically researchers conduct such studies through external surveys of social-network users, Mr. Zimmer says. Or they'll do an ethnography of a smaller group. That means the available data can be soiled because of self-reporting biases and errors, he says. Or it may not truly represent the population. Not only had Mr. Kaufman's team amassed an ample data set, but they had improved it by collecting information from the same class over four years. The data, as Mr. Kaufman puts it, amount to "a complete social universe."

But here's where things get sketchy. Mr. Kaufman apparently used Harvard students as research assistants to download the data. That's important, because they had access to profiles that students might have set to be visible to Harvard's Facebook network but not to the whole world, Mr. Zimmer argues in a 2010 paper about the case published in *Ethics and Information Technology*. The assistants' potentially privileged access "should have triggered an ethical concern over whether each student truly intended to have their profile data publicly visible and accessible for downloading," Mr. Zimmer says in an e-mail.

In an interview, Mr. Kaufman declined to discuss who helped collect the data. But the sociologist did concede in a videotaped 2008 Berkman talk that the assistants created "an interesting wrinkle to this, from a legal point of view."

"We faced a dilemma as researchers," Mr. Kaufman said on tape. "What happens if a student has a privacy setting that says, 'You can't see me unless you're my friend,' and our undergraduate research assistant who is downloading the data is a friend of that person? Then can we include them in our data?"

He left that question unanswered at the time. But Mr. Kaufman talks openly about another controversial piece of his data gathering: Students were not informed of it. He discussed this with the institutional review board. Alerting students risked "frightening people unnecessarily," he says.

"We all agreed that it was not necessary, either legally or ethically," Mr. Kaufman says.

Muddled Online Ethics

The Harvard case reflects how the Internet is changing the relationship between researchers and their subjects, sometimes creating what Elizabeth A. Buchanan, director of the Center for Applied Ethics, at the University of Wisconsin-Stout, calls a "strange distance" between the two. Researchers may grab content posted online without interacting with the people who wrote it or considering them "human subjects." But they may be aggregating data that can be traced to individuals, says Ms. Buchanan.

The fundamental question is how best to protect subjects, she says, "and sometimes in Internet research ... those issues get muddled."

For example, Quinnipiac's Mr. Halavais did a Twitter study focused on protests surrounding the Group of 20 summit in Pittsburgh. But something unanticipated happened: Some people were arrested for using Twitter to help demonstrators evade police. After that, one of the key people in the study deleted his Twitter account. What the subject didn't know was that researchers had collected his tweets in an archive and planned to publish papers about the data.

Mr. Halavais didn't seek approval from his review board—as he sees it, studying Twitter is like studying newspapers. "We did not predict that the very act of tweeting something might be considered a criminal offense," he says. "I don't think an IRB would have been able to predict that any better than we would."

A rule of thumb holds that if an online community requires a password to enter, then researchers must seek IRB approval to study its members. But some scholars go further, Mr. Halavais says, arguing that researchers should seek approval to study open publishing platforms like blogs and Twitter.

Attitudes toward privacy are also evolving, among both researchers and companies. Fred Stutzman, a postdoctoral fellow at Carnegie Mellon University who studies privacy in social networks, used to harvest Facebook data that students made public on his university network. He isn't sure he'd do that today.

"This is the nature of these systems," says Mr. Stutzman, who has criticized the Library of Congress's Twitter project. "Maybe in three years, we'll look at public tweets and say, Oh, my God, those weren't public. A lot of people that are using Twitter nowadays may actually want to go back and delete their accounts or take those things out of the public at a later date, and they no longer can."

Twitter recently alarmed researchers by saying that collecting tweets and making them openly available violates the terms of service, a blow to academics who want to share data.

Facebook, too, has taken a stricter approach to research as the company has matured and weathered several privacy controversies. Cameron Marlow, its head of data science and "in-house sociologist," has built up a small but tightly controlled program for external research since joining Facebook, in 2007.

Asked about the Harvard sociologists' project, Mr. Marlow says things would be different had it begun now: "We would have been much more involved with the researchers who are doing the data collection."

All work would be done on Facebook servers, for example. And releasing data? Unlikely.

"We tend to not release any data, for the fact that it's almost impossible to anonymize social-network data," he says.

What's the Danger?

Mr. Zimmer proved him right. Within days of Harvard's release from the data archive, he zeroed in on the institution without even downloading the profiles. Most of what he needed was in the archive's code book—a lengthy document, at the time easily available online (it has since been restricted), that described in detail how the data set was collected and what it contains. The size of the class, uniquely titled majors like "organismic and evolutionary biology," and Harvard's particular housing system all clued Mr. Zimmer in to the source of the Facebook information.

Mr. Kaufman, for his part, won't comment on whether Harvard is, in fact, the source of his data.

But assuming that Mr. Zimmer is correct, why does it matter? What's the danger?

One issue, Mr. Zimmer says, is that someone might be able to figure out individual students' identities. People with unique characteristics could be discovered on the basis of what the Harvard group published about them. (For example, the original code book lists just three students from Utah.) Their

information could be absorbed by online aggregators, like Pipl. A prospective employer might Google a student and use the resulting information to discriminate against him or her, Mr. Zimmer says.

"These bits and pieces of our personal identities could potentially have reputational harm," he says.

He's right about how easy it is to identify people who are presumably part of the data set. By searching a Facebook group of Harvard's Class of 2009, a *Chronicle* reporter quickly tracked down one of those three Utah students. Her name is Sarah M. Ashburn. The 24-year-old is in Haiti working for a foundation that helps AIDS victims.

The Facebook-data controversy was news to her. In a telephone interview, Ms. Ashburn says her main qualm with the project is its use of students who may have had privileged access to data that was supposed to be shared only with friends, or friends of friends. Because of that, she feels that the researchers should have informed the class about their project.

Still, she isn't concerned about the possibility that her own data is out there.

"Anything that's put on Facebook somehow will make it out into the general public, no matter what you attempt to do," she says. "So I never have anything on my Facebook profile that I wouldn't want employers, my grandmother, like anyone in the world to be able to see."

The Biggest Victim

In their defense, the Harvard sociologists stress that researchers outside their own group had to apply for access to download the data and agree not to share it or identify people within it. Distribution was halted immediately after privacy

concerns were raised, says Kevin Lewis, a Ph.D. candidate who is part of the research team. By that time, he says, fewer than 20 researchers had access. Each presumably still has a copy.

After the initial release, the researchers took additional steps to protect the students' identities. For example, a revised code book substituted general regions, like "mountain" and "Pacific," for students' home states, and general major categories, like "humanities" and "life sciences," for their academic backgrounds.

As for the criticism of Harvard's institutional review board, the university seems to agree on the need for greater guidance. A spokesman, Jeff A. Neal, notes that "current federal human-subjects regulations were written well before the Internet age, and there is still little published guidance for IRB's on the implications of new and emerging technologies and potential risks." He adds, "Federal regulators, professional associations, and IRB's are all working to understand these risks and to develop guidelines."

The biggest victim in this case may be scholarship.

The controversy has tainted Harvard's data. And "once a data set has been clearly de-anonymized, it becomes a little bit like kryptonite," says Mr. Halavais. "People will touch it, but you're putting your own ethical stance at risk if you do."

There may never be another chance to touch it. The Harvard sociologists are still using the data for their own research. But they haven't settled on a secure way of publicly sharing it again.

Since the public release ceased, in 2008, Mr. Lewis has received more than 200 requests for the data from researchers. He still gets one or two inquiries each week.