Jennifer Golbeck: The curly fry conundrum – Why social media “likes” say more than you might think.

0:11

If you remember that first decade of the web, it was really a static place. You could go online, you could look at pages, and they were put up either by organizations who had teams to do it or by individuals who were really tech-savvy for the time. And with the rise of social media and social networks in the early 2000s, the web was completely changed to a place where now the vast majority of content we interact with is put up by average users, either in YouTube videos or blog posts or product reviews or social media postings. And it's also become a much more interactive place, where people are interacting with others, they're commenting, they're sharing, they're not just reading.

0:53

So Facebook is not the only place you can do this, but it's the biggest, and it serves to illustrate the numbers. Facebook has 1.2 billion users per month. So half the Earth's Internet population is using Facebook. They are a site, along with others, that has allowed people to create an online persona with very little technical skill, and people responded by putting huge amounts of personal data online. So the result is that we have behavioral, preference, demographic data for hundreds of millions of people, which is unprecedented in history. And as a computer scientist, what this means is that I've been able to build models that can predict all sorts of hidden attributes for all of you that you don't even know you're sharing information about. As scientists, we use that to help the way people interact online, but there's less altruistic applications, and there's a problem in that users don't really understand these techniques and how they work, and even if they did, they don't have a lot of control over it. So what I want to talk to you about today is some of these things that we're able to do, and then give us some ideas of how we might go forward to move some control back into the hands of users.

2:01

So this is Target, the company. I didn't just put that logo on this poor, pregnant woman's belly. You may have seen this anecdote that was printed in Forbes magazine where Target sent a flyer to this 15-year-old girl with advertisements and coupons for baby bottles and diapers and cribs two weeks before she told her parents that she was pregnant. Yeah, the dad was really upset. He said, "How did Target figure out that this high school girl was pregnant before she told her parents?" It turns out that they have the purchase history for hundreds of thousands of customers and they compute what they call a pregnancy score, which is not just whether or not a woman's pregnant, but what her due date is. And they compute that not by looking at the obvious things, like, she's buying a crib or baby clothes, but things like, she bought more vitamins than she normally had, or she bought a handbag that's big enough to hold diapers. And by themselves, those purchases don't seem like they might reveal a lot, but it's a pattern of behavior that, when you take it in the context of thousands of other people, starts to actually reveal some insights. So that's the kind of thing that we do when we're predicting stuff about you on social media. We're looking for little patterns of behavior that, when you detect them among millions of people, lets us find out all kinds of things.

3:18

So in my lab and with colleagues, we've developed mechanisms where we can quite accurately predict things like your political preference, your personality score, gender, sexual orientation, religion, age, intelligence, along with things like how much you trust the people you know and how strong those relationships are. We can do all of this really well. And again, it doesn't come from what you might think of as obvious information.

3:43

So my favorite example is from this study that was published this year in the Proceedings of the National Academies. If you Google this, you'll find it. It's four pages, easy to read. And they looked at just people's Facebook likes, so just the things you like on Facebook, and used that to predict all these attributes, along with some other ones. And in their paper they listed the five likes that were most indicative of high intelligence. And among those was liking a page for curly fries. (Laughter) Curly fries are delicious, but liking them does not necessarily mean that you're smarter than the average person. So how is it that one of the strongest indicators of your intelligence is liking this page when the content is totally irrelevant to the attribute that's being predicted? And it turns out that we have to look at a whole bunch of underlying theories to see why we're able to do this. One of them is a sociological theory called homophily, which basically says people are friends with people like them. So if you're smart, you tend to be friends with smart people, and if you're young, you tend to be friends with young people, and this is well established for hundreds of years. We also know a lot about how information spreads through networks. It turns out things like viral videos or Facebook likes or other information spreads in exactly the same way that diseases spread through social networks. So this is something we've studied for a long time. We have good models of it. And so you can put those things together and start seeing why things like this happen. So if I were to give you a hypothesis, it would be that a smart guy started this page, or maybe one of the first people who liked it would have scored high on that test. And they liked it, and their friends saw it, and by homophily, we know that he probably had smart friends, and so it spread to them, and some of them liked it, and they had smart friends, and so it spread to them, and so it propagated through the network to a host of smart people, so that by the end, the action of liking the curly fries page is indicative of high intelligence, not because of the content, but because the actual action of liking reflects back the common attributes of other people who have done it.

5:47

So this is pretty complicated stuff, right? It's a hard thing to sit down and explain to an average user, and even if you do, what can the average user do about it? How do you know that you've liked something that indicates a trait for you that's totally irrelevant to the content of what you've liked? There's a lot of power that users don't have to control how this data is used. And I see that as a real problem going forward.

6:12

So I think there's a couple paths that we want to look at if we want to give users some control over how this data is used, because it's not always going to be used for their benefit. An example I often give is that, if I ever get bored being a professor, I'm going to go start a company that predicts all of these attributes and things like how well you work in teams and if you're a drug user, if you're an alcoholic. We know how to predict all that. And I'm going to sell reports to H.R. companies and big businesses that want to hire you. We totally can do that now. I could start that business tomorrow, and you would have absolutely no control over me using your data like that. That seems to me to be a problem.

6:49

So one of the paths we can go down is the policy and law path. And in some respects, I think that that would be most effective, but the problem is we'd actually have to do it. Observing our political process in action makes me think it's highly unlikely that we're going to get a bunch of representatives to sit down, learn about this, and then enact sweeping changes to intellectual property law in the U.S. so users control their data.

7:15

We could go the policy route, where social media companies say, you know what? You own your data. You have total control over how it's used. The problem is that the revenue models for most social media companies rely on sharing or exploiting users' data in some way. It's sometimes said of Facebook that the users aren't the customer, they're the product. And so how do you get a company to cede control of their main asset back to the users? It's possible, but I don't think it's something that we're going to see change quickly.

7:44

So I think the other path that we can go down that's going to be more effective is one of more science. It's doing science that allowed us to develop all these mechanisms for computing this personal data in the first place. And it's actually very similar research that we'd have to do if we want to develop mechanisms that can say to a user, "Here's the risk of that action you just took." By liking that Facebook page, or by sharing this piece of personal information, you've now improved my ability to predict whether or not you're using drugs or whether or not you get along well in the workplace. And that, I think, can affect whether or not people want to share something, keep it private, or just keep it offline altogether. We can also look at things like allowing people to encrypt data that they upload, so it's kind of invisible and worthless to sites like Facebook or third party services that access it, but that select users who the person who posted it want to see it have access to see it. This is all super exciting research from an intellectual perspective, and so scientists are going to be willing to do it. So that gives us an advantage over the law side.

8:48

One of the problems that people bring up when I talk about this is, they say, you know, if people start keeping all this data private, all those methods that you've been developing to predict their traits are going to fail. And I say, absolutely, and for me, that's success, because as a scientist, my goal is not to infer information about users, it's to improve the way people interact online. And sometimes that involves inferring things about them, but if users don't want me to use that data, I think they should have the right to do that. I want users to be informed and consenting users of the tools that we develop.

9:23

And so I think encouraging this kind of science and supporting researchers who want to cede some of that control back to users and away from the social media companies means that going forward, as these tools evolve and advance, means that we're going to have an educated and empowered user base, and I think all of us can agree that that's a pretty ideal way to go forward.