FTC Big Data: A Tool for Inclusion or Exclusion? Workshop
September 15, 2014
Segment 4
Transcript

SPEAKER 1: Time to find your spot without your food or beverage. Great. Thank you everyone for joining us this afternoon.

The afternoon session is going to be-- Commissioner Brill it's going to give the opening remarks to the afternoon session. And so without further ado, here is Commissioner Brill, who needs no introduction.

[APPLAUSE]

COMMISSIONER BRILL: Thanks everybody. Before I begin, let me just say thank you so much to Catherine, to Tiffany, to Patrick Eagan-Van Meter.

And to all the other-- and Katherine Worthman, if I didn't say that. To all the folks at the FTC who have been working so hard on this workshop.

I think that the quality of the panels this morning, the quality of the panels this afternoon show you how much work they put in to organizing this event.

So can we just have a quick round of applause for the FTC staff?

[APPLAUSE]

And thanks to all of you who are watching by webcast. And those of you who made it here today.

Our presenters and panelists are providing us with details about the current and emerging uses of big data to categorize consumers, the surrounding legal issues, and possible best practices for big data analytics providers.

I'd like to provide a more general and also, perhaps, personal perspective that makes, I hope, a simple point. Providing transparency into big data algorithms that categorize consumers has been done before.

It has put some concerns to rest. And companies and consumers have been better off as a result.

Now as I've said on one or two other occasions, those of you who've read some of my speeches or perhaps attended them, I believe that big data analytics can bring significant benefits to consumers. And to society.

But we must endow the big data ecosystem with appropriate privacy and data security protections in order to achieve these benefits.

Today, I'd like to focus on three of the more challenging issues at the intersection of big data and consumer protections that pertain to this workshop.

I'd also like to offer some suggestions. Some specific suggestions about what industry can do right now to address these concerns.

Consumer trust is critical here. And transparency and accountability are key to building it.

Now the first challenge involves traditional credit scores derived from credit reports. And alternative scoring models.

In this realm, as in many others, past is prologue. The origins of the Fair Credit Reporting Act have something to teach us about our current environment.

The FCRA was our nation's first big data law. The seeds for it were planted in the growing economy after World War II. Businesses formed cooperatives to enable quicker and more accurate decisions about credit worthiness by sharing information about consumers who were in default or delinquent on loans.

Over time, these agencies combined paving the way for consumers to gain access to credit, insurance, and jobs.

As credit bureaus increased their ability to draw inferences and makes correlations through ever larger databases, unease about the amount of information the credit bureaus held, as well as its accuracy and use, also increased.

Congress passed the Fair Credit Reporting Act in 1970 to address these concerns. The FCRA governs the use of information to make decisions about consumer credit, insurance, employment, housing, and other transactions initiated by consumers.

It covers not only credit bureaus, but also, importantly, their sources and their clients. The FCRA gives consumers important rights.

For instance, consumers are entitled to have access to their data, to challenge its accuracy, to have irrelevant data removed, and to be notified when they are denied credit, or get a loan at less than favorable rates because of negative information in their files.

The use of credit scores has thrived under the FCRA's rights of notice, access, correction, relevancy, and accuracy.

And the FCRA has enabled the credit reporting enterprise to serve a purpose useful not only to the credit reporting agencies and their clients, but also to consumers.

The credit scores that first emerged from analysis of consumers' credit files broadened access to credit. And they made determinations of a particular consumer's credit worthiness more efficient and more objective than the case was with prior more subjective determinations.

Now as scoring models began to proliferate and enter into new types of decisions, including employment, insurance, and mortgage lending, consumers and regulators grew concerned about what exactly was going on within these models.

Some of the most important questions were whether credit related scores were using variables that act as proxies for race, ethnicity, age, and other protected categories.

In 2003, Congress directed the Federal Trade Commission and the Federal Reserve to study these questions in the context of credit based insurance scores and traditional credit scores.

After extensive and rigorous studies, both agencies found that the scores they examined largely did not serve as proxies for race or ethnicity.

The FTC and Federal Reserve reports shed a lot of light on traditional credit scores and assuaged some important concerns. Which was good for everyone involved.

Consumers, credit bureaus, and credit score users.

Now let's fast forward to today. We're now seeing a proliferation of other types of scores being used to make FCRA covered eligibility determinations.

While these scores are subject or many of them are subject to the same obligations of access, accuracy, security, and other requirements imposed by the FCRA, they haven't yet been subject to the same kind of scrutiny that Congress and the federal agencies brought to bear on traditional credit scores.

The use of new sources of information, including information that goes beyond traditional credit files to score consumers raises fresh questions about whether these alternate scores may have disparate impacts along racial, ethnic, or other lines that the law protects. Or that should be addressed.

Those questions are likely to linger and grow more urgent unless and until the companies that developed these alternate scores go further to demonstrate that their models do not contain racial, ethnic, or other prohibited biases.

These companies may learn that their models have unforeseen inappropriate impacts on certain populations. Or they might simply find their algorithms should eliminate or demote the importance of certain types of data.

Because their predictive value is questionable as FICA recently discovered with respect to paid off collection agency accounts and medical collections.

Just as we did a decade ago, the FTC and other appropriate federal agencies should once again devote serious resources to studying the real world impact of alternate scoring models.

But industries shouldn't wait for federal agencies or for Congress, for that matter, to get involved to review their own scoring models.

Companies can begin this work right now. And provide us all with greater insight into and greater assurances about their models.

The second big data challenge I'd like to discuss comes from the unregulated world of data brokers.

As outlined in the Commission's recent report, and as discussed this morning, data brokers' profiles combine massive amounts of data from online and offline sources into profiles about nearly all of us.

Data brokers' clients use these profiles for purposes that range from marketing, to helping companies determine whether, and on what terms, they should do business with us as individual consumers.

Now the main data broker issue that I'd like to highlight today concerns data broker segments that track sensitive characteristics. Including race, religion, ethnicity, sexual orientation, income, children, and health conditions.

As I noted when the FTC released its landmark report on data brokers, I see a clear potential for these profiles, ethnic second city struggle or urban scrambler, to harm low income and other vulnerable consumers.

In an ideal world, a data broker's products that identify consumers who traditionally have been underserved by the banking community can be used to help make these consumers aware of useful opportunities for credit and other services.

However, these same products could be used to make these consumers more vulnerable to high interest payday loans and other products that might lead to further economic distress.

It all depends on how these products are actually used. Importantly, our recent data broker report did not attempt to analyze the harms that could potentially come from the uses of consumer segmentation of poor or minority communities.

Now one of the reasons I support legislation to create greater transparency and accountability for data brokers, as well as their sources and customers, is so we can all begin to understand how these profiles are being used, in fact.

And whether and under what circumstances they are harming vulnerable populations. In the meantime, the data broker industry should take stronger, proactive steps right now to address the

potential impact of their products that profile consumers by race, ethnicity, or other sensitive characteristics.

Or that are proxies for such sensitive classifications.

Here's what I'd like to see data brokers do. They should find out how their clients are using these products. They should tell the rest of us what they learn about their actual uses.

They should take steps to ensure any inappropriate uses cease immediately. And they should develop systems to protect against such inappropriate uses in the future.

Now, the third challenge I want to mention relates to companies that use their own data and analyze their own data about their customers.

Companies understandably are eager to determine what makes their customers happy. And how they can more efficiently service these customers.

As they dive into their own treasure trove of customer data, in order to offer perks, or better deals to loyal customers, companies may also find that these common practices disadvantage certain groups of individuals.

Thereby, in the words of the White House's big data recent report, exacerbating existing socioeconomic disparities.

Back in January, the Harvard Business Review asked companies to think deeply about where value added personalization and segmentation ends. And harmful discrimination begins.

Now I want to emphasize that all of these industry players, traditional credit reporting agencies and their newfangled progeny using alternate scoring models, data brokers, and the companies that use their products, and companies engaged in analysis of their own customer data, all of these players can take steps right now to address concerns about the potential discriminatory impact of their use of algorithms.

I'm hopeful that the same reservoirs of data that create the concerns I outlined will also lead to ways to get them under control. I encourage all members of industry to look for ways that the data in their hands could be used to identify disparate treatment along racial, ethnic, gender, or other inappropriate lines.

And to correct such treatment to the extent it exists. Thank you very much.

[APPLAUSE]

SPEAKER 1: Thank you very much, Commissioner Brill. Now the next part of our afternoon agenda, before we get the next panel, is going to be a presentation digging into the data.

And I'd like to introduce Latanya Sweeney who's been the Chief Technologist at the FTC. And Jinyan Zang, a research fellow in technology and data governance.

So I'll leave you with the clicker. Excellent.

[APPLAUSE]

LATANYA SWEENEY: So it's great to be here. My name got mentioned a couple of times. So I feel like I don't need any other introduction.

[SLIGHT LAUGHTER]

But I do want to thank Tiffany, and Catherine, and Katherine, and Patrick, and Maneesha, and DPIP for organizing this. And for allowing us this opportunity to present our work.

Assuming I can get the clicker to work. Because after all, I'm the technologist, right?

[LAUGHING]

So one of things I wanted to also let you know is we started a summer research program under the guidance and leadership of Chairwoman Ramirez, who you met this morning.

The idea was to bring in some of the best and brightest students and have them do research during the summer on areas of interest to the FTC.

Today, we're going to report on one such project. And we worked as a team. So all of the fellows kind of contributed to all of the efforts. But Jinyan and I primarily did the one that we're going to talk about today.

Christa and Jim couldn't be here, but Paul is here. I'll just have him stand up. And the work that's coming out from the other fellows will be coming over the next weeks.

So the Pittsburgh Courier was once the country's most widely circulated black newspaper. It had a circulation of about 200,000.

If you worked for the Courier or if you were to interview their staff back in 1911, they would say that your clicker doesn't work.

[LAUGHTER]

They would say that when an ad appeared in their newspaper, they would review that ad. They had to review that ad because they didn't want to run the risk of alienating, isolating, or insulting the audience that they served.

Today, the Pittsburgh Courier-- the clicker still doesn't work. The Pittsburgh Courier has an online website.

And their ads are delivered through an online network for which no staff member actually reviews the ad. Instead, it's a big data analytic engine that delivers their ad.

Now we all know the promise. And we've heard a lot about the advantages of big data analytics. And online advertising is no exception.

It's not that you want just any only ad showing up anywhere. You want the ads organized so that the fisherman sees the fishermen ads. And the young mother sees the baby products. And so that's the promise.

But in order to deliver that, there's a lot that happened to get that Macy's ad on that Pittsburgh Courier page, there are a lot of parties and a lot of different ways that can happen.

So let me just blow it up and introduce some of the ways. So there's groups that will help you put together your ad campaign and your ad copy. Help you find platforms on which to sell it.

There are data brokers that are involved to taking the outside data-- is it the battery? Or is it I just don't know how to push the button?

[LAUGHTER]

Data brokers taking outside data, bringing it into the online network, figuring out when it is to offer, or what kind of offer, or which ad would be the right one to target directly to you. And make that connection from end to end through that kind of network.

And so that's normally called targeted advertising. But we're not going to talk about targeted advertising right now.

Let's talk something simpler. Where there's only one party that's going to go from end to end. Such as the Google network.

Google delivers more than 30 billion ads a day. And every ad is delivered in the time it takes to load a webpage. I'm a computer scientist, that is awesome. That's really awesome.

Well, how do they do this? Well, we're not going to get into specifics. And I'm not sure everyone actually knows the specifics outside of Google.

But we do know that there are billions of ads on one side. And what an ad bid is basically the copy. The ad copy.

The key words of the audience that they would like to show that ad to. And how much money they'll pay either to get that ad put in front of the audience. Or for someone to click on it.

On the other side are these publishers who will basically take an ad. And so Google gets to make the decision as to which ad is going to show up when.

We're very interested in how Google goes about doing that. Not so much about ripping open that cloud, that blue cloud, but understanding what effects might be on the outside.

So one of the things we did was we turned to Mix Rank. Mix Rank is a service whose whole business is about capturing online ads. So they survey the internet constantly, record every ad they encounter, where they encountered it, the data that was encountered.

And so then you can look at the data through the eyes of the publishing site or through the advertisers. So this is an example.

One of the things they do is they get rid of behavioral effects. And retargeting effects. So this is nice for our study.

Because now we're looking at it with the assumption that that blue cloud doesn't know anything more about you than it would know about anyone else. Under those circumstances, how does the blue cloud perform?

So we found this website, Omega Psi Phi. Now Omega Psi Phi had its 100th anniversary in 2011. Set up a special domain just for the site.

It's a fraternity that is very popular in the United States among black men in colleges. It sports many outstanding black men among its members. Including Congressman Clyburn, Bill Cosby, Shaquille O'Neal.

And we became interested in what kind of advertisement showed up on that site. Well, there are lots of ads about graduate degree programs. Which, of course, seems incredibly appropriate given that it's an undergraduate fraternity. And a clicker that doesn't work.

[LAUGHTER]

What is it with this? I'm going to win.

[LAUGHTER]

There are also advertisements about luxury vacations. And other kinds of opportunities like that. And then there are also these kinds of advertisements such as this one.

Click here to view your arrest record now. Now there has been much said about Instant Checkmate. And this is an Instant Checkmate ad.

I did earlier work about the suggestive nature of arrest ads around Instant Checkmate. But I think it's very clear to see that this actual ad is not showing up the way it regularly showed up.

It actually shows up with flashing colors. So it has kind of a neon effect. Flashing your arrest record would be a presumption that this particular audience would not appreciate. It wasn't the only ad though that made that kind of presumption.

There was also ads for a criminal lawyer. And there were ads for credit cards. Now it turns out that the financial industry is the number one marketer on online.

So they're the number one industry that's advertising online. And given what we have just seen of Omega Psi Phi, we became very interested in what kind of credit card ad is that?

And what are credit card ad experiences? I hope you have better luck with the clicker.

JINYAN ZANG: All right. So going more generally from the Omega Psi Phi anecdote, we first started looking for what are lists of quality cards versus ones that are more harshly criticized online.

So here you can see a list of the top 25 most harshly criticized cards or the most highly praised cards that we were able to find.

And from Omega Psi Phi, they actually had two of the ads from the harshly criticized list show up on their site. Including First Premier card and the Centennial card. None of the ads from the highly praised cards list actually showed up on their site.

And, in fact, for the highly praised cards list, it's not necessarily those cards are all just high credit score, really luxury cards. In fact, you have secured cards that were highly praised as well. Like the CapitalOne Secure card.

But digging back into the comparison of the two cards, what we saw was if you looked at the most popular ad that ran for first Premier card. Which is one of the most often criticized cards if you go online.

And compare that to the most popular ad that was ran by American Express for the blue card, the sites that those card ads appeared on do look very different. And one theme that quickly jumps out at you, especially for the American Express Blue card, is around higher education.

Where you had sites such as harvardmagazine.com, or yellowalumnimagazine.com, or like theheismanwinners.com, as sites that American Express was advertising on.

On the other hand, for First Premier's card, there didn't seem to be as much of a cohesive scheme that we picked up.

LATANYA SWEENEY: So we wanted to dig further. Like what is the nature of these cards? Where are they appearing generally? And is it somehow related, perhaps, to the popularity of the website?

So if you think about popularity of websites, there are a few websites that are highly popular. Almost everyone goes to. They're on the top of everyone's top 10 list.

And then the popularity of the website drops as you go further out. Alexa is a company that ranks the traffic to and from domains. And so we used them to rank all of the publishers of all of the credit card ads' deliveries that were made of the praised cards and the criticized cards.

And what we learned was that the criticized cards appear completely across the entire spectrum in increasing order as the popularity of the domain drops. So it's a curve that's going this way.

And in every segment of the popularity zones, there are, in fact, credit card ads for the criticized ads. The highest number though were in those ads whose popularity ranks were above a billion.

Now to be above the billion, you probably aren't getting much traffic. That would be the issue with respect to your popularity. Those ads that are close to the left are highly popular.

Those are very curated. And there's a lot of information that exists about the audience. So you could actually look up on services like Quantcast to find out the demographic makeup of those websites.

But when you're way out in the billions and millions, that kind of information doesn't exist. The other thing to note though is where were the praised cards?

They didn't follow the same pattern. Instead, they were heavily generated in the middle. Around the 100,000 to the one billion. So these ads are showing up on different popularity. On websites and domains with different kinds of popularity profiles.

JINYAN ZANG: And another perspective that we took to look at the type of sites that these card ads are running on was from the perspective of understanding that different websites do attract different types of audiences.

And that there are websites out there that are more exclusive to an audience of one demographic group than other demographic groups. So we took the approach are analyzing Comscore's data on the browsing behavior of 46,000 American households in 2013.

And looked through the four million websites those households go to. To look for sites that are more commonly visited by households of certain demographic groups.

And so, for example, if we took a racial lens to demographics, we found that for Latino Americans, they're more likely to go to sites like Univision, or Toringa, or musica.com.

For African American households, they went to sites like worldstarhiphop, footlocker.com.

Now in this case, it doesn't necessarily mean that only African Americans go to footlocker.com. Footlocker.com could have lots of other visitors from other racial groups, as well.

But African American households are much more likely to go to footlocker.com. And so we looked at exclusivity from the lens of race, from age, from income, from the level of education in the household.

And also whether the household had children or not. And we are able to find for each of those different lenses sites that were exclusive to each of those groups.

And this raises a question for us of if there are sites that are out there that are more exclusive to certain groups, what is the advertising experience like on those sites? And could there be the potential for disparate impact if depending on the type of ads that are shown or the type of ads that are actually not shown on those sites?

LATANYA SWEENEY: So one of the things that we learned was that these groups are appearing almost evenly across the entire popularity of these domains.

That means that no matter which ad campaign you ran, whether one you were trying to focus on popular domains or less popular domains, you could easily encounter one of these domains for which there was an exclusive audience.

Because, in fact, they appeared in all the domains. So what we then became interested in was to what extent could we predict whether or not the ad would receive?

Or were their sites in the Comscore data that should have received these credit card ads? And were those sites part of these exclusive groups?

And we found that it's true. That around race and income and age, there were differences. And, in fact, there were praised ads. And these praise ads, for example, for Asians we saw CapitalOne secure card.

And CapitalOne, and Citibank, and Discover finding domains which were more exclusive to Asians. And you couldn't tell by the name or the key word of the page.

The domains are names like dealstobuy.com or visajourney.com. Discover did a very good job using seekingalpha.com to target people whose incomes are 100k or more.

That's an exclusive audience at seekingalpha.com. And it's a very popular site. And then we also found examples in age ranges. Discover with ages 18 to 20.

And CapitalOne Secure card found some domains that were somewhat exclusive to ages 25 to 29. Or ages 65 plus.

So-- why did you do that? So domains with exclusive audiences do exist. And ads are not exempt from being delivered to those sites.

So the lack of ads or the too much of another ad could lead to a disparate impact. And demographics could, therefore, infer what kind of advertising experience might you have.

We're going to stop here. If you want more information about the work, we'll have a blog post later with some of the details. And a paper to follow right after that.

I did want to leave the panel that's coming up next with three questions from this work. One of them is that by subscribing to an online ad network, a publisher may not have an opportunity to review ads anymore.

And if there is a problem, what is the publisher's rights and responsibilities? Another question that comes from this is when we look at Omega Psi Phi.

What are the sufficient and necessary circumstances for a community to experience adverse impact in this setting?

And the last question is that the kind of audience exclusivity measure that we use to find these audiences that had that type of exclusive nature to the audience is something that could actually be used inside of the big data engine in that same fraction of a second to realize that this ad probably shouldn't go to the site at this time.

If that's so, and it's that easy to do, should or how might a big data analytic engine be required to use it or an equivalent remedy?

So to find out more about the work, check out our Tech@FTC blog. Thank you.

[APPLAUSE]