NEW YORK UNIVERSITY
NYU STERN
LEONARD N. STERN
SCHOOL OF BUSINESS

DEPARTMENT OF INFORMATION, OPERATIONS & MANAGEMENT SCIENCES

HENRY KAUFMAN MANAGEMENT CENTER

FOSTER PROVOST
Professor
NEC Faculty Fellow
Paduano Fellow in Business Ethics (Emeritus)

February 18, 2011

Re: Comment on FTC report "Protecting Consumer Privacy in an Era of Rapid Change."

To Whom It May Concern:

I am submitting this comment on the request of Prof. Edward Felten of the FTC. My name is Foster Provost. Besides the credentials in the letterhead, I am a computer scientist, and just retired from being the Editor-in-Chief of the journal *Machine Learning*. My main non-academic professional interest is in designing and incubating data-science-based companies, currently mainly in advertising technology. I have at least two specific research interests related to the FTC report. First, one of my main lines of research lies in using data to build models to estimate things that will improve some aspect of business (or government). I have worked on such "data mining" research related to applications such as fraud detection, targeted marketing, counterterrorism, and many others. Over the past four years or so I have focused on data modeling for on-line advertising, including modeling for the effective targeting of display advertising. Another of my main interests is data "privacy". Putting the two together, I am particularly interested in whether and how data modeling and targeting can be done in a privacy-sensitive manner—currently focusing on on-line display advertising.[1] In this comment I will focus primarily on data and privacy in on-line advertising, because that is the specific topic about which I've done the most thinking. I believe the general principles apply to data practices more broadly.

I was quite impressed by the FTC report draft, and I am happy that the general notion of "privacy by design" is being taken seriously. To keep this comment as brief as possible, I will not elaborate all the good things in the report, but will focus on one area of concern.

Background

Let me give a brief background, lest my comments be grossly misinterpreted. I am in favor of more informed choice by consumers with respect to when and how data about them are stored and used. A problem I have with the current designs for on-line privacy solutions like "Do Not Track" (DNT) relates to the "informed" part of "informed choice." In order for us to give a

[1] NB: I would like to state clearly that I have invested in and advise an on-line advertising firm that I believe makes the best tradeoff of privacy and efficacy—which happens to use methods I designed and published (with others). While I try to be objective, this certainly colors my opinions.

consumer informed choices about when and how her data are being used, we need to be able to do (at least) two things. (1) We have to be able to describe the different choices clearly and in a not-too-technical way. I haven't seen that much progress in this direction, and much of what follows attempts to illustrate that there are subtle but very important differences from the privacy perspective. Using the examples I describe below, explicit-profile based targeting is quite different from doubly anonymized targeting. Will the user really understand the differences well enough to choose? If we believe these to be fundamentally different from the privacy standpoint, is it fair to paint them with the same brush? (2) It is not really informed choice if we do not inform the consumer about both the benefits and the drawbacks. Do consumers know that some of the most effective targeted ads inform people of special sales, and they may not receive these ads if they choose not to be "tracked"?[2] More problematic still, no one knows right now what the drawbacks of choosing not to be targeted will be. I will leave a deep analysis to my economics colleagues, but it certainly seems easily within the realm of possibility that if Do Not Track were presented as having only benefits, then many people would choose it. If then there is a substantial difference in the revenue to a publisher from those who have chosen not to receive targeted ads and those who have, then it makes sense (for example) that publishers might institute a multi-tiered system of access to content, based on privacy choices. Then, would consumers have been properly informed about the choices that they made? How could they if at this point we don't know what the response of different players will be?

Concern over unintended consequences

Because of this substantial uncertainty, it would make sense for the FTC to proceed with caution. Why not start by trying to provide options for the most egregious privacy-invasive practices, and at the same time giving incentives for more privacy-friendly solutions. In parallel, we can work to define consumers' choices in a clear and complete manner, continually improving the option design. A superficial treatment inevitably will favor some parties over others in unintended ways. I have a particular interest in small business, and I fear that when huge on-line ad serving companies also are browser makers and large-scale publishers and run advertising exchanges, we need to be very careful not to tilt the world further in their favor unless we actually intend to.

Concern that high-level framework may stifle innovation in privacy-friendly solutions

By my reading of the report and other treatments of the topic of consumer tracking, especially for the purpose of targeting on-line ads, I feel that there may not be a general understanding of the important nuances of data collection, tracking, modeling, and targeting. If consumer tracking is painted with too broad a brush, it will hinder or stymie the development of good "privacy by design" solutions, treating all of them as equally invasive. I have argued in my writing and speaking that there is a spectrum of possible solutions between two extremes. Almost all the debate occurs at the extremes, but neither of the extremes is particularly attractive: "you can't do anything with *my* data" and "we can do anything we want with whatever data we can get our hands on." I believe that we should promote the development of a variety of

[2] This introduces the important distinction between not being tracked and not being *targeted*. One might decide they don't want to be tracked in a semantically meaningful way (e.g., saving explicit data about their prior actions), but they would not mind being targeted in a privacy-friendly way. Someone else might decide that it is the targeting itself that is objectionable—in both cases, the consumer should understand clearly both the benefits and the drawbacks.

different privacy-by-design solutions, residing at intermediate points along this spectrum—and I'm not sure that we really will know what are the best privacy/efficacy tradeoffs until firms have the incentives to try seriously. I will present a couple alternatives as we go on, but there may be much better ones that just have not yet been designed. My fear with the whole FTC effort is that the result will actually reduce innovation on privacy-sensitive solutions, and possibly have a catastrophic effect on the on-line advertising industry or equally troubling, concentrate the power in the hands of a few extremely large players that engineer the on-line world in their own favor, at the expense of a market for innovation.

This comment

My main technical point for this comment is that there are a variety of different sorts of data that are/could be collected, different degrees of anonymization of these data, and different fundamental ways of targeting based on these data. In my view, they definitely are not all equally objectionable from a privacy/confidentiality standpoint. Moreover, not considering the breadth of possibilities can block meaningful debate, e.g., if you mean one thing by "tracking" and I mean another (see below). So, in what follows I will try to lay out some different sorts of data and different targeting practices, and give my own opinion of the elements that one might consider in a privacy-by-design approach. If you don't want to read through the whole thing, please note that this includes a design that in my opinion is quite sensitive to privacy concerns (predictive modeling based on doubly anonymized data)—a design that also has been shown to be quite effective both in research and in large-scale practice. I will include as an appendix by hyperlink a technical paper on the subject and the slides from a lecture that include very large-scale real-world targeting results.

Different methods of targeting and the associated data

There are several different ways that we might target on-line ads to browsers, each based on a different sort of data. The privacy implications for the different scenarios are not the same. Let me discuss the implications first for what I see to be the data targeting scenario that garners the most concern. I then will discuss some other approaches that are very important and very effective, but in what I have read and heard, seem to be largely overlooked (and are in danger of being thrown out with the bathwater).

Let me start a running example by, as I write this sentence, loading http://finance.yahoo.com. I receive display ads from Fidelity and from Scottrade. Not surprising—those are almost certainly premium ad purchases to a very high-traffic page. No reason to believe my seeing them has anything to do with personalized targeting.

Practice #1: Direct targeting based on an explicit profile (incl. "behavioral targeting")

To my knowledge the approach that garners the most privacy concern can be described roughly as "compile an explicit profile on each browser and target people based on one or more elements of that profile." Now, in our running example, Fidelity and Scottrade may be interested in advertising in the much (much!) less expensive non-premium ad market (where most on-line display ad slots are bought and sold)—both reducing their own costs and effectively pumping money into the rest of the internet economy besides the premium publishers. They may be

interested in advertising on other pages that are finance related, for which they might use "contextual targeting," which I won't discuss here.

Usually, an advertiser also is interested in targeting browsers who have a "profile" that fits its prior notions of the type of consumer who would be a good prospect for their product. These profiles could be based on buying socio-demographic data from various sources in the display ad ecosystem. They also could be based on viewing the browsing behavior of individuals—specifically, the types of pages that they visit. If they've seen that I tend to visit finance-oriented sites, they may want to target me wherever I go, not just when I happen to be on a finance-oriented page. This creating of explicit profiles and targeting based on them often is called "behavioral targeting."

From a privacy (confidentiality) standpoint, there are two important dimensions. First, does the data directly identify the person (let's call that "direct" PII)? A first step in protecting privacy is for firms to commit to a design that does not store any direct PII. This could include for example associating all data with a random key that is stored in the browser's cookie, and having the firm commit not to store a consumer's name, address, and anything else that is deemed to be direct PII. I don't want to seem to treat this lightly: there are important nuances that have to be taken seriously. And indirect PII is a concern as well. Do we consider IP address to be PII? Maybe, maybe not. If so, then a privacy-sensitive design should have an option not to use IP address. Is it possible to "reidentify" a user based on some other data that is stored? The likelihood depends on the specific design choices made, and also on policy choices. An explicit policy *not* to try to reidentify consumers possibly can go a long way. And there remains the concern of a data breach, to which I will return later.

The second privacy dimension of concern is the explicit profiles. Even if there were no association with your identity at all, perhaps you wouldn't want your browsing to be continually associated with an explicit profile of your interests. (There are various reasons, which I won't elaborate here.) This is a bit harder to deal with, because a key for advertisers is to be able to find the "Finance" people and target them (for example) on small-business web sites, where the costs are lower.[3] It seems to me that it would be possible to have an encrypted system (perhaps based on public-key encryption) such that firms could target particular categories, but the profile would not ever be viewable.[4] I'm not aware of this sort of a method being used in this context, but technically it certainly seems possible, and as a consumer I think I would prefer such a system.[5]

I am concerned with what I've read and heard about potential regulation by the FTC and others: there seems to be the potential actually to restrict the development of privacy-by-design

---

[3] Small business web sites commonly sell their advertising space via ad exchanges. Advertisers bid significantly higher for this space if they can know certain things about the individual visitor.

[4] And further, firm/policy decisions might restrict particularly sensitive categories completely, such as health-related, sex-related, etc.

[5] Some of my colleagues have introduced a different privacy-by-design approach that provides an alternative solution for direct targeting based on an explicit profile—store the profile in the browser itself! See http://crypto.stanford.edu/adnostic/ and http://crypto.stanford.edu/adnostic/adnostic-ndss.pdf .

solutions, by painting all these options with the same brush. *Even in a "Do Not Track" world, should the proposed anonymized, encrypted system be treated identically to a completely privacy in-sensitive system that tracks as much as it possibly can?* Is there a line beyond which the systems are not "tracking" in the spirit of the regulation? If so, should such systems be exempt?[6] I'm not an economist, but an exemption strategy seems to give incentives to design more privacy-sensitive systems and firms. It also may well reduce the fear in the industry of a catastrophe—especially if there were privacy-sensitive systems that were approximately as effective as the privacy-insensitive systems.

Practice #2: Retargeting

As I mentioned above, there are other data-oriented strategies for targeting ads in the non-premium display ad market, that are in danger of being painted with the same brush as the explicit profiling approaches. My main point is not to argue that these should necessarily be "exempt" (although I believe they should, if they are designed well). I rather would like to make sure that they are not overlooked in a careful analysis of how best to proceed.

A large portion of the on-line display ad industry believes that one of the most effective[7] methods of targeting display ads is "retargeting." In its simplest form, retargeting works like this: the advertiser, usually via its targeting agents, records that a particular browser has visited the advertiser's own site (possibly buying something, or just browsing). This browser then is targeted with ads from this advertiser elsewhere on the web, under the belief that someone who already has shown brand affinity is a very good candidate for advertising. So following our running example: if I were to visit Fidelity's site, the retargeting strategy would be to subsequently target me with Fidelity ads elsewhere on the web. (By bidding for my browser in the ad exchanges, for example.) Real targeting results consistently show that retargeted browsers buy advertisers' products at a rate consistently higher than random browsers and higher than most other targeting methods.[8]

At first glance, retargeting looks an awful lot like behavioral profiling. However, from the privacy perspective there are two very important differences. First, retargeting does not require a broad, detailed, multivariate profile. It just requires one or a handful of closely related variables (did the browser visit our site at all? Did it buy something?).

The second difference is more subtle, but very important. The data being used to target an ad is the advertiser's own experience with that browser. This may be done by an agent (such as a demand-side platform, or DSP) on behalf of the advertiser, so the technical details and the sort of agency are important, but just about everything in the advertising world is done by an agent on

---

[6] The notion of a distinction from a privacy perspective between different sorts of data is not new. The Fair Information Practices Principles differentiate between personal information and non-personal information. It may be that the data ecosystem of the 21st century really requires further, careful thought about the different levels of confidentiality now possible.

[7] This belief is not universal. There exist both practitioners in firms with competing technologies and academics who have argued that retargeting may not be as effective as is generally believed. I have not seen incontrovertible evidence either way, and much depends on how one defines "effective." In any case, this is not critical to the present discussion.

[8] The drawback from the advertiser's perspective is that the reach is limited to those browsers who have been observed to visit the brand's site.

behalf of a brand.[9] The distinction is that there is a certain sense in which the fact of a browser visiting a particular site is data that is jointly "owned" by the browser and the site.[10] It seems quite different for a brand to target based on it's own prior experience with a browser, than to target based on an explicit profile obtained from who knows where.

Furthermore, retargeting could be very privacy friendly. It is easy to envision privacy-sensitive solutions where the visits to different brands' sites are encoded such that visits to brand X's site only are "knowable" in the context of a setting up a campaign for brand X.

The main point is that from a privacy-by-design standpoint, there are important differences in the sort of data collected based on browsers' visits to sites. Treating them all equally lumps together different practices with very different privacy implications. I have a hard time seeing why a well-designed (anonymization-based) retargeting system should not be exempt from a Do Not Track rule, if the intention is to find a good tradeoff between privacy and efficacy.

Practice #3: Targeting based on predictive modeling (incl. privacy-friendly social targeting)

A quite different approach to targeting might be called the "predictive modeling" approach. This is the approach that I've espoused in my own work as being particularly "privacy friendly" or privacy-preserving.[11] Instead of creating an explicit profile and then having advertisers choose particular sorts of browsers based on elements of the profile, the predictive modeling approach uses data science technologies (e.g., machine learning) to automatically build a model of the sort of browsers that would be good prospects, for example those who would buy a particular advertiser's products. Then in use, the system automatically targets browsers that receive high scores from the model.

I won't go into detail here about how exactly that would work. From the privacy perspective there are two very important differences from the explicit profiling case. First, no explicit, semantically meaningful[12] profile ever need be created. Second, and relatedly, and perhaps surprisingly, no semantically meaningful data at all need be stored about the browser, nor even any encrypted data that can ever be decoded (with one caveat that I'll return to below).

How can that be? Doesn't the predictive model need some data on a browser to score it? Of course it does, but the key is that because the modeling and scoring is all done completely automatically, by machine, the data can be completely anonymized through the whole process.

---

[9] The notion of "first-party" vs "third-party" marketing tends to be treated superficially in debates about data privacy. There are many different parties and agency relationships, as well as large holding companies both on the publisher and advertiser sides, and of course the mega-companies that play many roles. Well-designed, anonymous "third-party" tracking may be more privacy friendly than "first-party" tracking, when the first party actually knows who you are (e.g., you've bought things from them or have an email account with them).

[10] "Owned" in quotes, because the concept of data ownership is a quagmire that we don't need to step into right now.

[11] NB: The term privacy-friendly has different meanings by different people. The approach that I espouse tries to ensure privacy by design as much as possible, while still allowing effective targeting. This is described in detail in "Audience Selection for On-line Brand Advertising: Privacy-friendly Social Network Targeting." Provost, F., B. Dalessandro, R. Hook, X. Zhang, and A. Murray. In *Proceedings of the Fifteenth ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining (KDD 2009)*, the link to which I've included as an appendix.

[12] I'm not sure this is the best term. What I mean is a profile that a human can read and get meaning from.

This needs clarification, because the notion of anonymization is used in many claims these days. First, the browser can be anonymized, as described above. But more important, the data can be doubly anonymized. The resulting data can have all semantic meaning removed permanently; we can anonymize them *and throw away the key*. Returning to our example, say that one data point that happens to lead to better targeting of Scottrade ads is whether or not a browser in the past has visited http://finance.yahoo.com. Instead of saving any data on having visited that site, or having visited a finance site, or the words on the page, or whatever, the system can encode that data point immediately and irreversibly (say, as pa98wey#2se). The predictive modeling does not care; it will use pa98wey#2se identically to the original value. If the encoding is done irreversibly at the "outer wall", then it would take a substantial social engineering effort to figure out anything about any browser. This general approach, and a specific implementation, is described in detail in the paper mentioned in Footnote 11.

Above I mentioned that there is one caveat to the semantic irreversibility. In order to build ("train") the predictive model we need to have a "target" variable—that thing that is going to be predicted, like proclivity to purchase, or some surrogate thereof. Fortunately, the perfect information for training is the retargeting data discussed above. And following the same line of reasoning, this can be encrypted, only to be "knowable" in the context of the campaign for that particular brand—which as discussed above, if designed appropriately is quite benign from a privacy perspective.

I mentioned above that I would return to the possibility of a data breach. If the data are completely and irreversibly encrypted (as with the doubly anonymized solution), then there will be minimal fallout from a data breach as well. A stolen or mistakenly released data file will contain gibberish. What's more, having the semantic meaning irreversibly removed will limit the secondary uses to which the data can be put.

Results both in the lab and in large-scale practice show that predictive modeling approaches, based on doubly anonymized data such as this, can be remarkably effective (cf., the paper and talks slides in the appendix). By my understanding, from multiple sources, advertisers who evaluate this sort of technique in practice consistently rate it as being among the most effective approaches. The slides for the talk "Machine Learning for Display Advertising" that I've linked as an appendix show (page 13, "in vivo" performance) that in a large-scale real-life evaluation, the "privacy-sensitive" techniques for targeting browsers are many times better than non-targeted advertising, across scores of different major advertisers (these results are typical).

A curious but important nuance to privacy-by-design approaches is that for very well-designed systems, it may be difficult or impossible for firms to comply with requests to allow consumers to "see and manipulate" what "information" has been stored "about them."[13] If the data are truly anonymized, then as a consumer the best I could ask is what data are associated with this particular browser that I'm using right now. If the data are truly and irreversibly doubly anonymized, then the firm would have no semantically meaningful "information" about me to report to me. (NB: it still may be able to report that I am a strong candidate for such-and-such a brand.) Reporting encrypted gibberish to a consumer may be worse than nothing, if the

[13] Quotes from p. 63 of the draft report.

consumer does not understand the system's privacy design—it may seem that the firm is instead hiding something. If I don't understand what they're doing, I might decide just to opt out.

Finally, it is not clear in the report whether techniques such as these are included in the definition of "behavioral advertising" at all. In the industry, to my understanding "behavioral advertising" typically is used in the specific sense of advertising based on explicit profiles (as discussed above) that are aggregated into categories that are of interest to advertisers. However, we could think of extreme privacy-sensitive approaches also as being covered by the term "behavioral advertising." Do we want to lump these all together? If not, how can we design a system where it is realistic for consumers actually to take the time to understand the differences, in order to make an informed choice?

Closing remarks

Just because my own research led to one particular design, and I've invested in it, I want to be clear about my message. Of course I would prefer that the approaches and the business not be rendered irrelevant by regulation. And even more I would like regulation to reward R&D on privacy-by-design, and to reward businesses that have taken privacy seriously without being forced to. However, more to one main point of this comment, it's not clear whether this particular design that I happen to have worked on is the best. *We should do whatever we can to give incentives for the production of more, different, and even better privacy-sensitive designs.*

I don't want to treat this issue lightly by suggesting that I have a silver bullet. I think this itself is a design problem that should be approached with at least as much thought and care as the development of a privacy-sensitive system or business. However, the FTC report already contains a general mechanism that might be used (the devil's in the details, of course). Specifically, the report already contains the recommendation certain "commonly accepted" data practices be exempt. Should we have an exemption for accepted privacy-preserving data practices as well? That could be made flexible such that as new privacy-sensitive designs are invented, they could become "accepted" via some mechanism. My fear with such an approach is that the time required to obtain acceptance would stymie innovation—would I have the opportunity to test different designs in practice, or would I have to wait for approval for each? An alternative would be a mechanism that exempts everything <u>but</u> a specified set of practices, and the set could grow as specific practices are identified and deemed to need privacy protection. That might at the same time protect consumers, give regulatory flexibility, and spur innovation. The fear there would be that it gives firms the incentive simply to get around the existing rules, rather than to produce solutions that in the light of day are really better from a privacy standpoint.

In summary, it would be great if the industry itself really worked toward better and better privacy designs. I urge the FTC strongly to figure out how to put in place a framework that challenges the industry to do better.

When we take into account both privacy concerns and business concerns, a poorly designed approach could produce much more damage than benefit, especially if it results in large numbers of consumers opting out of essentially all targeted advertising, because they have been given a superficial "informed" choice. By my understanding, audience targeting tends to increase the

income most (relatively speaking) to small-scale web sites. Returning to our earlier example: rather than spending its marketing budget primarily on the likes of Yahoo! Finance, if Scottrade can target individuals well, it likely would be quite happy to spread its advertising spending across more small-business web sites. On the other hand, with a poor design it may be exactly the small-scale web businesses who are hit hardest; in a high-opt-out world, one may expect small businesses to have the least ability to offer differential services. So while the big boys adjust their systems to deal economically with the opt-outs, the small businesses just have to live with lower revenue for the same content.

Also, since I happen to be particularly interested in small businesses, let me reiterate a point I made above. The relationships among firms and technologies in the current online advertising ecosystem are very complicated, with certain players playing multiple roles. A superficial treatment of privacy design inevitably will favor some parties over others in unintended ways. For example, when huge on-line ad serving companies also are browser makers and large-scale publishers and run advertising exchanges, we need to be very careful not to tilt the world even further in their favor unless we actually intend to.

If the answer has to be "Do Not Track" regulation, I'd like to take this opportunity to be so bold as to recommend: please think carefully about the spirit of what practices we would like to avoid, and what practices we would like to keep or to encourage. I hope I have illustrated that in on-line ad targeting there are very different data-based practices and very different data, and that all should not be painted with the same brush.

If you have any questions, please don't hesitate to ask.

Sincerely,

Foster Provost
Professor
NEC Faculty Fellow
Paduano Fellow in Business Ethics (Emeritus)

Appendix

Audience Selection for On-line Brand Advertising: Privacy-friendly Social Network Targeting.
Provost, F., B. Dalessandro, R. Hook, X. Zhang, and A. Murray. In *Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (**KDD 2009**). http://pages.stern.nyu.edu/~fprovost/Papers/kdd_audience.pdf

Machine Learning for Display Advertising
Keynote talk discussing privacy and efficacy of targeting on-line ads
http://pages.stern.nyu.edu/~fprovost/Papers/MLOAD.pdf