Senate Judiciary Committee
Subcommittee on Privacy, Technology and the Law
Hearing on "What Facial Recognition Technology Means for Privacy and Civil Liberties"
July 18, 2012

Questions for the Record from U.S. Senator Al Franken
for Professor Alessandro Acquisti

1. **Your research showed that you could use off-the-shelf hardware and software to predict an incredible amount of information about someone you've never met before just by taking a photo of them. How hard and how expensive would it be for someone to commercialize your experiment—to basically create an app that lets people identify strangers?**

The costs and technical challenges of replicating our experimental results are, and will keep, falling. In fact, an application such as "Klik" (an app made available by Face.com before it was acquired by Facebook) supports the conclusions of our experiment, as it shows that real-time face recognition on mobile devices is already becoming a consumer product – notwithstanding the undeniable current limitations of these technologies.

One of the goals of our experiment was, in fact, to only use publicly available data and technologies that other third parties and end-users could also get access to: off-the-shelf face recognition technology,[i] limited computational power accessed through cloud computing services, and limited amounts of facial images made publicly available by end users on online social networks. In reality, today, both governmental and private sector entities have access to more powerful computational tools and much larger (and more accurate) repositories of digital photos than we had.

In other words, what we did can be replicated. Of course it would take time to re-create the procedure and mine the data necessary for face matching. Furthermore, the results we obtained are not yet scalable to the entire American population for a number of reasons: First, computational costs: we estimate that comparing the shot of a person's face to a database with mugshots of 280 million US residents aged 14 years or older, using the same hardware as in our experiments, would take over four

hours (rather than the few seconds that process took in our experiments). Second, "false positives:" when comparing millions of human faces, several individuals' faces will be similar to each other, and computers do not yet excel in separating a face of a person from a face of someone who looks very much like that person. Third, light conditions, facial hair, or non-frontal poses impair the accuracy of machine face recognizers. Fourth, photographic images (and in particular frontal mugshots) may not be available for the entire population.

Those hurdles, however, are being progressively overcome. They are transient, not systemic. First, as computers' processing power increases, and cloud computing costs decrease, it will become more efficient to run end-user applications on mobile devices similar to the one we developed, for mass-scale, automated, peer-to-peer face recognition. Second, false positives will likely keep reducing, as machine face recognition error rates are decreasing by one half about every two years. Third, researchers are actively working on improving computers' ability to recognize faces under varied conditions of light, facial hair, and poses. Fourth, entities such as online social networks are amassing some of the largest known databases of identified photos, from which increasingly accurate biometric models or "faceprints" of increasing portions of the US population can be, and are being, built.

2. **Based on your different experiments, can you tell us what you can more or less predict about a stranger using off-the-shelf facial recognition technology?**

In our experiment (summarized at www.heinz.cmu.edu/~acquisti/face-recognition-study-FAQ/) we demonstrated how to predict individuals' Social Security numbers (SSNs) starting from their face. This is possible through a process of *data accretion*[ii] that involves a chain of inferences. The process itself, however, is not limited to the prediction of SSNs; hence, a priori, there is no specific limit to the amount of personal data which can be linked to (and therefore can be inferred from) a person's face. In other words, once someone develops the appropriate chain of inferences, that person may infer, for instance, a stranger's sexual orientation, credit score, or current address.

By this, I do not mean that all sensitive inferences will be equally possible or likely. I am also not implying that all these inferences are possible right now: the accuracy of the predictions depend on a large number of factors, including the amount of data publicly available and the accuracy of face

recognizers. However, what I do mean is that the process itself is sufficiently universal to be applicable to a vast set of potentially sensitive inferences.

To explain, the process can be summarized in the following manner: First, face recognition links an unidentified subject (for instance, a face among many on the street) to a record in an identified database (such as an identified photo of the subject on Facebook, LinkedIn, Amazon, or in a state's DMV database). Once the link has been established, any online information associated with that record in the identified database (such as names and interests found in the subject's Facebook profile; or demographic data found on Spokeo.com - a social network and data aggregator) can in turn be probabilistically linked to the unidentified subject. Lastly, through data mining and statistical re-identification techniques, such online information can be used for additional, and much more sensitive inferences (such as sexual orientation,[iii] or Social Security numbers[iv]), which, in turn, can be linked back to the originally unidentified face. As I wrote in my testimony, sensitive data is therefore linked to an anonymous face through what we may refer to as a "transitive property" of (personal) information - a process that merely requires publicly available data. Sensitive information thus becomes "personally predictable" information.

This process (that we applied to SSNs) is generalizable for the following reason: In recent years, research in statistical re-identification and data mining has shown that you can predict sensitive data starting from non-sensitive information. While in 2009 we used dates of birth and states of birth to successful predict individuals SSNs, other researchers have identified medical conditions starting from demographic data, sexual orientation based on a Facebook users' network of friends, psychological profiles/psychiatric conditions based on their online social network profiles, credit scores based on someone's purchase history, and so forth. This is why I stated that, *in principle,* there is no ex ante barrier to what one can predict, starting from public data, since more and more individual data about individuals is being captured, and techniques for finding patterns in that data are getting more sophisticated.

Hence, a person's face becomes a veritable conduit between her different personas: it can link a person's online world (for instance, her social network presence) to her offline world (the person walking in the street). If an application can link a person's face to her name, and can link her name to public data about that person, then that application can also make sensitive inferences based on that data (as in the examples above, her SSN, your sexual orientation, or credit score) and link them back to the person's face.

[i] We used PittPatt, a face recognition application developed by former CMU researchers. Just a few days before our results were publicly presented, PittPatt was acquired by Google, thus the technology is no longer publicly available.

[ii] P. Ohm, 2010. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." UCLA Law Review, 57, 1701.

[iii] C. Jernigan and B.F.T. Mistree, 2009. "Gaydar: Facebook friendships expose sexual orientation." First Monday,14(10).

[iv] A. Acquisti and R. Gross, 2009. "Predicting Social Security Numbers From Public Data." Proceedings Of The National Academy Of Science, 106(27), 10975-10980.