

Digital Footprints: What Can be Learned from the Traces We Leave on Social Networks

Table of Contents

Carnegie Mellon University..... 4

SEI Copyright..... 4

Digital Footprints: What Can be Learned from the Traces We Leave on Social Networks 5

CS..... 10

What is Metadata?..... 12

What is Metadata?..... 13

“We don’t see any content.” 14

Well,... what might your postman infer? 15

What about this person? 16

And these residents? 17

Metadata is Data..... 18

Just 3 pieces of information..... 19

Boston, 1772 – A piece of the picture. 20

Boston, 1772 21

Boston, 1772 22

Explore your own Metadata 24

My Network, early 2010 26

My Network, late 2010 27

My Network, 2011 28

My Network, early 2012 28

My Network, late 2012	29
My Network, early 2012	30
My Network, late 2012	31
What if we only see you, and your likes?	32
October 2012	33
Kosinski, Stillwell, & Graepel (2012, PNAS)	34
October 2012	36
Kosinski, Stillwell, & Graepel (2012, PNAS)	40
Kosinski, Stillwell, & Graepel (2012, PNAS)	41
Kosinski, Stillwell, & Graepel (2012, PNAS)	42
Kosinski, Stillwell, & Graepel (2012, PNAS)	45
Anonymization is Easy to Break.....	46
Anonymization is easily broken	47
The US Census.....	48
The US Census.....	49
The Netflix Prize	50
The Netflix Prize	51
The record of movies you watch is a fingerprint	52
Health Records.....	55
Health Records.....	56
There are Real Consequences.....	57
There are real consequences.....	57
There are real consequences.....	61
Real Consequences	65

Personal Precautions	66
Cookies	67
First-Party Cookies	69
Third-Party Cookies.....	70
Third-Party Cookies.....	71
Incognito/Private Browsing	72
Disabling Third-Party Cookies	73
Browser fingerprinting.....	74
Browser fingerprinting.....	76
Fingerprinting and cookies example.....	77
Revoking access	79
Seeing what's known about you, and controlling what's shared.....	80
Takeaways.....	81
Digital Footprints: What Can be Learned from the Traces We Leave on Social Networks	85

Carnegie Mellon University

This video and all related information and materials (“materials”) are owned by Carnegie Mellon University. These materials are provided on an “as-is” “as available” basis without any warranties and solely for your personal viewing and use.

You agree that Carnegie Mellon is not liable with respect to any materials received by you as a result of viewing the video, or using referenced websites, and/or for any consequences or the use by you of such materials.

By viewing, downloading, and/or using this video and related materials, you agree that you have read and agree to our terms of use (www.sei.cmu.edu/legal/).

Distribution Statement A: Approved for Public Release; Distribution is Unlimited

SEI Copyright

Copyright 2018 Carnegie Mellon University. All Rights Reserved.

This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center.

The view, opinions, and/or findings contained in this material are those of the author(s) and should not be construed as an official Government position, policy, or decision, unless designated by other documentation.

References herein to any specific commercial product, process, or service by trade name, trade mark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by Carnegie Mellon University or its Software Engineering Institute.

NO WARRANTY. THIS CARNEGIE MELLON UNIVERSITY AND SOFTWARE ENGINEERING INSTITUTE MATERIAL IS FURNISHED ON AN “AS-IS” BASIS. CARNEGIE MELLON UNIVERSITY MAKES NO WARRANTIES OF ANY KIND, EITHER EXPRESSED OR IMPLIED, AS TO ANY MATTER INCLUDING, BUT NOT LIMITED TO, WARRANTY OF FITNESS FOR PURPOSE OR MERCHANTABILITY, EXCLUSIVITY, OR RESULTS OBTAINED FROM USE OF THE MATERIAL. CARNEGIE MELLON UNIVERSITY DOES NOT MAKE ANY WARRANTY OF ANY KIND WITH RESPECT TO FREEDOM FROM PATENT, TRADEMARK, OR COPYRIGHT INFRINGEMENT.

[DISTRIBUTION STATEMENT A] This material has been approved for public release and unlimited distribution. Please see Copyright notice for non-US Government use and distribution.

This material may be reproduced in its entirety, without modification, and freely distributed in written or electronic form without requesting formal permission. Permission is required for any other use. Requests for permission should be directed to the Software Engineering Institute at permission@sei.cmu.edu.

Carnegie Mellon® and CERT® are registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

DM18-0711

Digital Footprints: What Can be Learned from the Traces We Leave on Social Networks

Digital Footprints: What Can be Learned from the Traces We Leave on Social Networks

April Galyardt

Carson Sestili

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University
Software Engineering Institute

[Distribution statement A] Approved for public release and unlimited distribution.

**001 Speaker: And hello from the campus of Carnegie Mellon University in Pittsburgh, Pennsylvania. We welcome you to Virtual SEI. Our presentation today is Digital Footprints: What Can be Learned from the Traces We Leave on Social Networks. My name is Shane McGraw. I'll be your audience moderator for today's presentation. And I'd like to thank you for attending. We want to make today's event as interactive as possible. So, we will address questions throughout today's presentation and again at the close of the presentation. And you can ask those questions at any time, depending on what platform you're watching on, through the Q and A or chat tabs.

What else here? Also, we will have a survey. That survey tab will be in our chat window here in a second as your feedback is always greatly appreciated. So, please complete that upon exiting today's event.

And now, I'd like to introduce our two speakers for today. April Galyardt is a statistician and data scientist specializing in applications of statistical machine learning tools to cognitive science, learning analytics, and educational data mining. Welcome, April. Next, we have Carson Sestili. And Carson's a machine learning research scientist within our CERT division in their data science group where he uses data science, statistics, and machine learning for research and cybersecurity and intelligence. Carson, welcome.

Speaker: Thank you for having me.

Speaker: And now, we're going to turn it over to April. April, all yours.

Speaker: Thank you. So, it seems pretty obvious that one of the reasons we're talking about this right now is everything that's just come out recently about the Cambridge Analytica and Facebook association. And so, maybe we should start with what happened--

Speaker: Yeah, can you--

Speaker: With that.

Speaker: Remind me.

Speaker: So, Cambridge Analytica set up a survey to-- it looked like a scientific survey. And they actually used a scientific personality survey. And a lot of Facebook users went there taking the personality survey and answered a lot of other questions. And then they asked, through the Facebook API, to have access to the user's account through that API. And a lot of users said yes. But when they went-- when they shared their data, then they also shared a lot of their friend's data as well.

Speaker: Friends who had not used the service.

Speaker: Friends who had not taken the survey. Friends who gave no permission to let their data go out. And so, I don't have the numbers quite right off the top of my head, but it was something like thirty thousand people filled out the survey. And they shared more than a hundred thousand people's data through that, so just all those first network connections. So much data got out. And then that data was used for a lot of political purposes in the last election. And so, that's-- the fact that that happened, and data was used in ways that surprised people, I think has restarted a national conversation about how data gets used. And so, we wanted to contribute to that national conversation.

Speaker: I think even also what data even is, that's part of--

Speaker: Metadata--

Speaker: Right.

Speaker: And kinds of things, yeah.

Speaker: That's part of what we're about to talk about today as well is we're about to show a video clip in which a senator, during the Zuckerberg hearing, tries to ask a question about what kind of data is being released but doesn't quite have the language to--

Speaker: Right, and how Facebook uses the data. And he's trying to ask a really important question. It's one of the reasons I chose this particular clip is because Senator Schatz, the question he is trying to ask is really important, but he doesn't quite get that across. So, let's play that now.

Speaker: Because, both as a matter of practice and as a matter of not being able to decipher those terms of service in the privacy policy, is what exactly are you doing with the data, and do you draw a distinction between data collected in the process of utilizing the platform and that which we clearly volunteer to the public to present ourselves to other Facebook users?

Speaker: Senator, I'm not sure I fully understand this. In general, people come to Facebook to share content with other people. We use that in order to also inform how we rank services like newsfeed and ads to provide more relevant experiences.

Speaker: Let me try a couple of

specific examples. If I'm emailing within WhatsApp, does that ever inform your advertisers?

Speaker: No, we don't see any of the content in WhatsApp. It's fully encrypted.

Speaker: Right, but is there some algorithm that spits out some information to your ad platform. And then, let's say I'm emailing about "Black Panther" within WhatsApp, do I get a "Black Panther" banner ad?

Speaker: Senator, we don't-- Facebook systems do not see the content of messages being transferred over WhatsApp.

Speaker: Yeah, I know, but that's not what I'm asking. I'm asking about whether these systems talk to each other without a human being touching it.

Speaker: Senator, I think the answer to your specific question is if you message someone about "Black Panther" in WhatsApp, it would not inform any ads.

Speaker: Okay.

Speaker: Okay so, what Senator Shatz is really trying to ask is is there a distinction between the data we choose to share and what we know we're sharing and the data which is just vacuumed up, and how do you use the data that's vacuumed up. And the answer that Zuckerberg didn't give is, "Yeah, we use all of it."

Speaker: Yeah, there's so much that I-- there's so much information I know just about who you're talking to and not even about-- I don't have to know for sure the content of your conversation.

Speaker: Right, the content.

Speaker: But something about who I'm talking to can also tell me a lot about what we could be talking about.

Speaker: Right, and so, the we don't use that information for ads is-- that's a shortcut around the question that Senator Shatz was trying to ask. So, let's see if this works.

CS



IN CS, IT CAN BE HARD TO EXPLAIN THE DIFFERENCE BETWEEN THE EASY AND THE VIRTUALLY IMPOSSIBLE.

Today, we're gonna try.

<https://xkcd.com/1425/>
(CC BY-NC 2.5)

**003 So, I thought this xkcd cartoon-- this is-- this cartoon is actually three or four years old, which is why it's extra funny to me

now. It says, "When a user takes a photo the app should check whether they're in a national park. Oh sure, that's easy. And check whether the photo is of a bird. I'll need a research team and five years." Well, since this is three or four years old, yeah we can almost do that now.

Speaker: It's doing a pretty good job, yeah.

Speaker: Yeah, we're much better at-- we can tell whether there's probably a bird in that photo. But the point is that it can be really hard to explain the difference between what's easy and what's almost impossible. And of course, what's almost impossible is changing every day. So, today, we're going to try and explain at least some of the things that are super easy to do with the kinds of data that Facebook has, the stuff that a lot of the CS researchers and statisticians know. Oh yeah, we learned that in undergrad kind of things that a lot of people don't know are even possible.

Speaker: It's worth knowing if you're in the audience and you're thinking, "Well, I don't use Facebook. I'm fine," everything that we're saying today about Facebook can be used for many forms of communication. In fact, later in the talk, it will be even when you're not communicating with somebody, there's information that you give away about yourself when you do online activity.

Speaker: Right, Facebook's the reason we're talking about it, but this is--

Speaker: They're not the only guilty party.

Speaker: Everybody who's online, this is an issue.

Speaker: Right.

What is Metadata?

What is Metadata?

Metadata



Content



**004 Speaker: So, let's start with What is metadata? So, I tend to think of metadata as it's the outside of the envelope. You don't get to see what's in the package. The content is what you get to see if you open the package. The metadata is what's on the envelope.

What is Metadata?

What is Metadata?



**005 But if you think about that, you've got a recipient. You've got a sender. You've got a date. You know what kind of package it is. You know how big the package is. Was it insured? This one has eBay on the label. So, there's actually a lot of data that's outside of the envelope data.

“We don’t see any content.”



**“We don’t see
any content.”**

****006** So, when Zuckerberg says,
"We don't see any content," that's
kind of misleading because there's a
lot of information that's not content.

Well,... what might your postman infer?

Well,... what might your postman infer?



Carnegie Mellon University
Software Engineering Institute

Digital Footprints
© 2018 Carnegie Mellon University

Distribution Statement A Approved for public release and unlimited distribution.

7

****007** And really to drive this point home, if you are-- you see this in a mailbox, and the postman is delivering this mail, so actually paper mail analogy here, this gives you a pretty clear picture of who this person might be. We've got ACLU, NPR, Cook's Illustrated, a bank statement.

Speaker: Without opening any of the pieces.

Speaker: Without opening anything, you have a picture.

What about this person?

What about this person?



**008 And then this set of mail, you get a very different picture. Again, we haven't opened anything. This is just outside of the envelope.

And these residents?

And these residents?



**009 And now, we can think about this person. We've got a parenting magazine, I don't know, some sort of kid's box package thing, children's hospital bill. That's a lot of children's hospital bills. And then we see that. And again, we haven't opened any envelopes. But we know some of what's going on in this person's life. And we know it's not good.

Metadata is Data

****010** So, that's the first thing.
Metadata is data. The data on the
outside of the envelope is data, and it
can be used.

Speaker: Sure.

Speaker: And so, the we don't see
any content statement, I find quite
misleading.

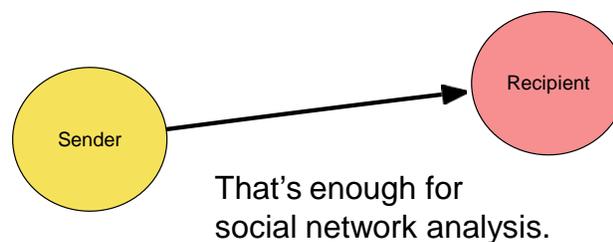
Speaker: And to go back to, for
instance, Zuckerberg's analogy. They
don't need to read the content of
your text to know that you're the
kind of person who might be
interested in seeing the movie "Black
Panther." There's in fact-- there are--
not that there's only one kind of
person, but there's going to be
people who are more interested in
that movie than they would be
interested in some other movies. And
they can absolutely use that content

to market to you, to profile you in certain ways, only using that information on the outside of the envelope.

Just 3 pieces of information

Just 3 pieces of information

- Sender
- Recipient
- Date



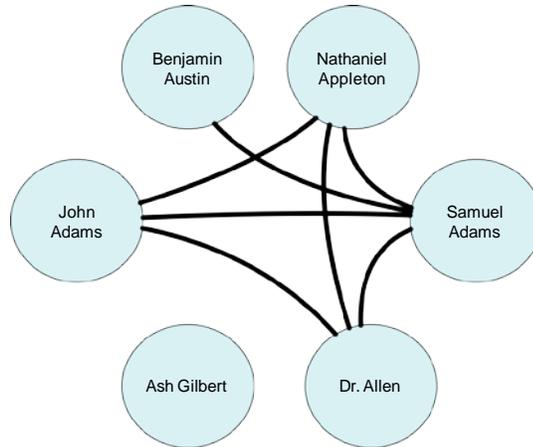
**011 So, just to kind of drive this home, if we've got just three pieces of information, so you have a sender, a recipient, and maybe a date the connection was made.

Speaker: This is just for like email?

Speaker: Yes, well this is an area of statistics, social network analysis. It's been done in the social sciences for a long time. But we've, over the last twenty years, figured out ways to make it really quantitative and really precise. And so, if you start with just three pieces of information, that's enough to start using some of these statistical methods.

Boston, 1772 – A piece of the picture.

Boston, 1772 – A piece of the picture.



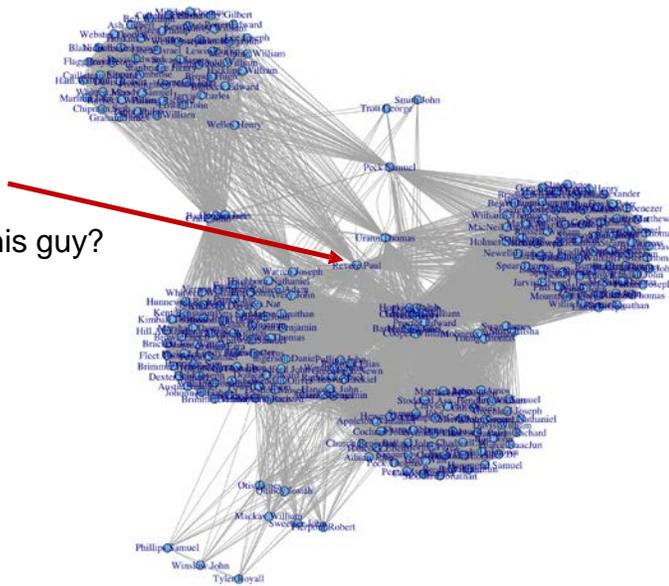
This example is due to Kieran Healy:
<https://kieranhealy.org/blog/archives/2013/06/09/using-metadata-to-find-paul-revere/>

**012 And this example is-- I'm probably going to mispronounce her name, but Kieran Healy, she's a social scientist who used some of this on data from Boston in 1772. So, a connection between two of these men indicates that they belonged to the same club. So, yes, that John Adams and yes, that Samuel Adams, they belong to a club together.

Boston, 1772

Boston, 1772

Wait, Who's this guy?



**013 And if you do this for all of the clubs in Boston in 1772, you get something like this. So, you get a lot of groups that okay, these guys all belong to the same two clubs. These guys all belong to the same three clubs.

Speaker: What does the distance in this visualization mean?

Speaker: So, this is done-- the distance is done algorithmically. So, the closer two points are, the more they're in the same cluster. It's kind of a spring-loaded thing.

Speaker: The more they're likely to be in the same club, for instance?

Speaker: Well, the more connections they have together, the

closer they're going to be to each other.
Speaker: Okay, sure.

Speaker: And so, if you look at this guy right here in the middle, just from looking at the picture, you can see he's in the middle of everything. And he is connected to everybody.

Boston, 1772

Boston, 1772



With 3 pieces of information and minimal calculations, we've identified a figure central in the "uprising" of 1776.

**014 So, that's Paul Revere. He is your connection between all of the men, the politically active men, in Boston in 1772. So, the mathematics to make the graph and calculate those distances, that's a little fancy, but to find the guy in the middle, that's like I can count.

Speaker: Sure, and a reminder that a gray line here, an edge, is just they

were in the same club

Speaker: They were in the same club.

Speaker: Okay, there was not even any knowledge about what they would have talked about or--

Speaker: Nope.

Speaker: When even. This is just they're in the same club.

Speaker: This is just they knew each other.

Speaker: Okay, yeah.

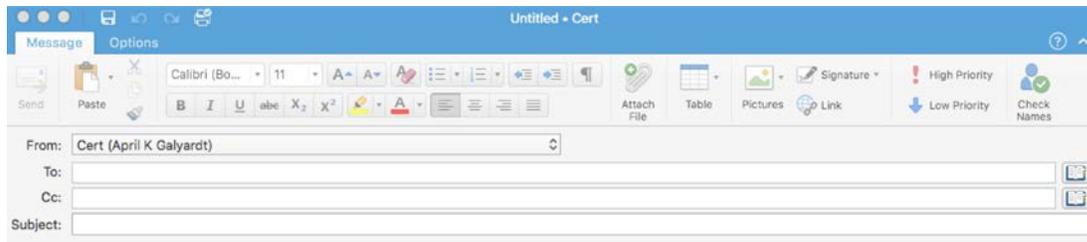
Speaker: They saw each other regularly.

Speaker: Got it.

Speaker: And so, if you calculate some of those basic measures of between-ness and centrality, Paul Revere's at the top of the list. And so, with three pieces of information per person, and basically the only calculation is addition, we've identified a central figure of the uprising of 1776.

Explore your own Metadata

Explore your own Metadata



Data Used:

From

To

CC

Timestamp

<https://immersion.media.mit.edu/>

**015 And so, you can-- there's a tool that was put together a few years ago by a group at MIT where you can explore your own metadata. And they will look at your Gmail account, and they really just use for this the from, the to, who was cc'd, and a timestamp. And so, I went and used this. And to use this app, you have to give them ridiculous permissions to so much of your Google account. So, I used it for twenty minutes, and then I deleted all the permissions. I deleted all the data. And if you use this, I recommend highly that you do the same. That when you are done, delete everything and remove the permissions.

Speaker: Yeah, and too, this is a little hint at what I'm about to talk about later, but they're very up front

with you.

Speaker: Right.

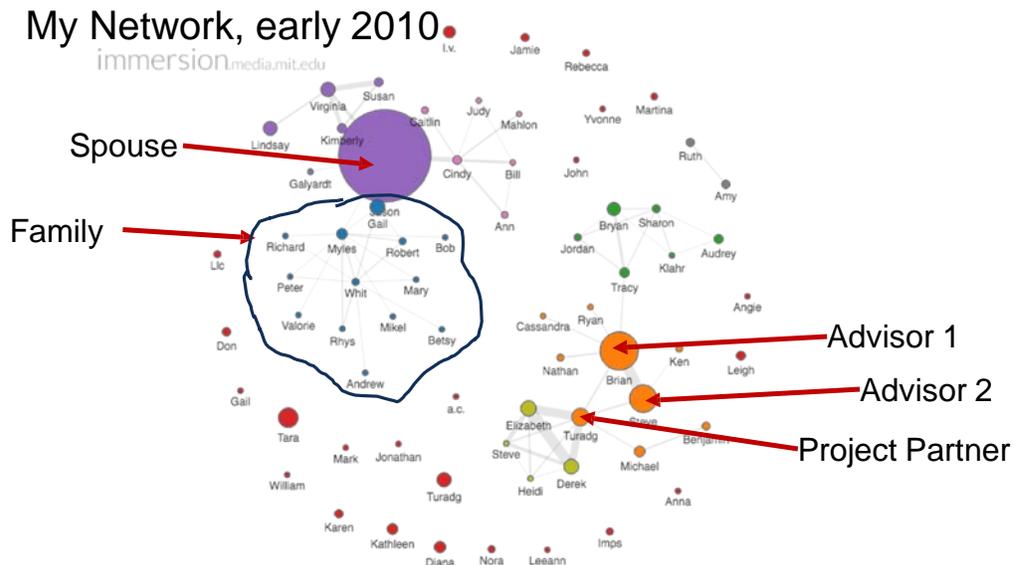
Speaker: They say, "If you give us access to your Gmail account, we can see everything. We promise not to use anything more than the from, the to, the cc, and the timestamp. And we also promise not to sell your data to everybody. But we could if we wanted to." And I think it's very important to understand whenever you're giving an app access like this, they will not always be as up front with you.

Speaker: Right, well these guys are researchers, and they're bound by their university's IRB and an approval process. And there are strong controls so that they are up front and not abusing that trust.

Speaker: Right.

Speaker: And a lot of advertisers are not bound by controls like that. So, this is what I did.

My Network, early 2010



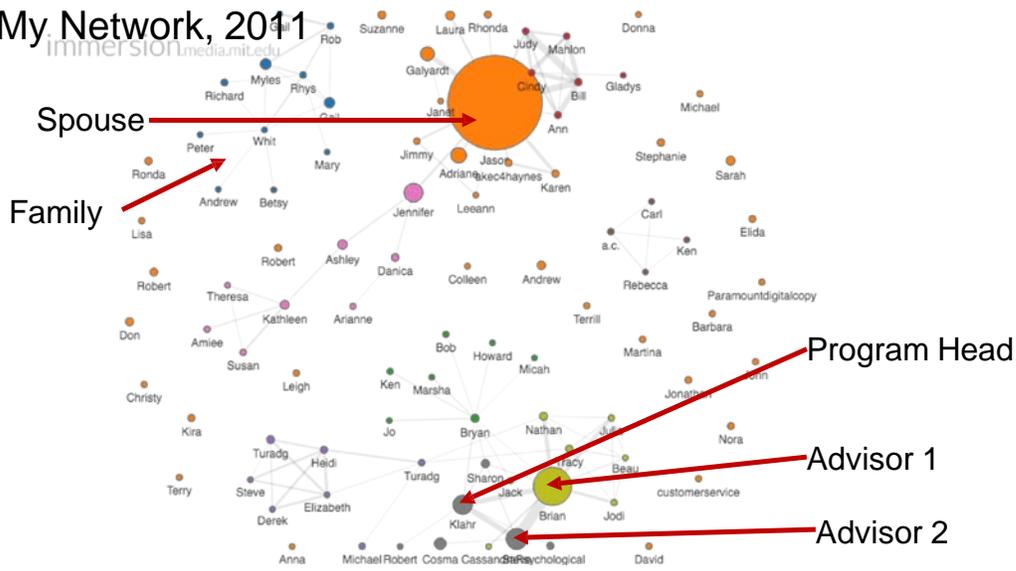
**016 And so this is my network from Gmail, people I emailed in 2010. And the big one, that's my spouse. In 2010, I was in graduate school. So, there's advisor number one.

Speaker: Size of circles is number of emails?

Speaker: Yeah, number of emails, and an edge between people indicates that they were cc'd. So, you can see a lot of lines between Brian and Steve. So, that-- yes, of course, I wrote a lot of emails to Brian and Steve because I was working with both of them. And then the kind of medium-sized one there, Turadg, he was a partner that I worked on a project with. And my connection with all of these people is already public. I have published papers with Turadg, and Elizabeth, and Derek. So, I feel comfortable sharing this because it is known that we work together. And

My Network, 2011

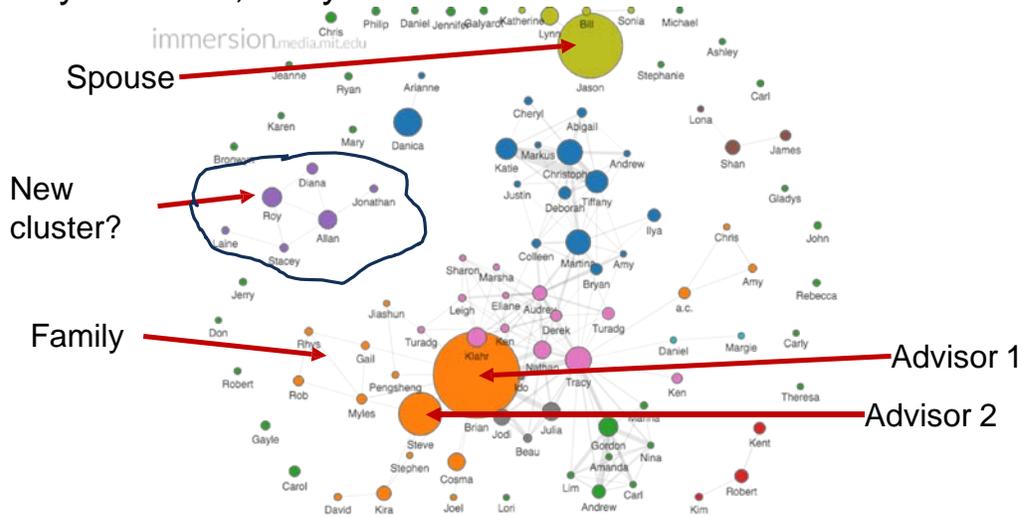
My Network, 2011



**018 Now, the program had shown up, and the family cluster has moved away a little bit.

My Network, early 2012

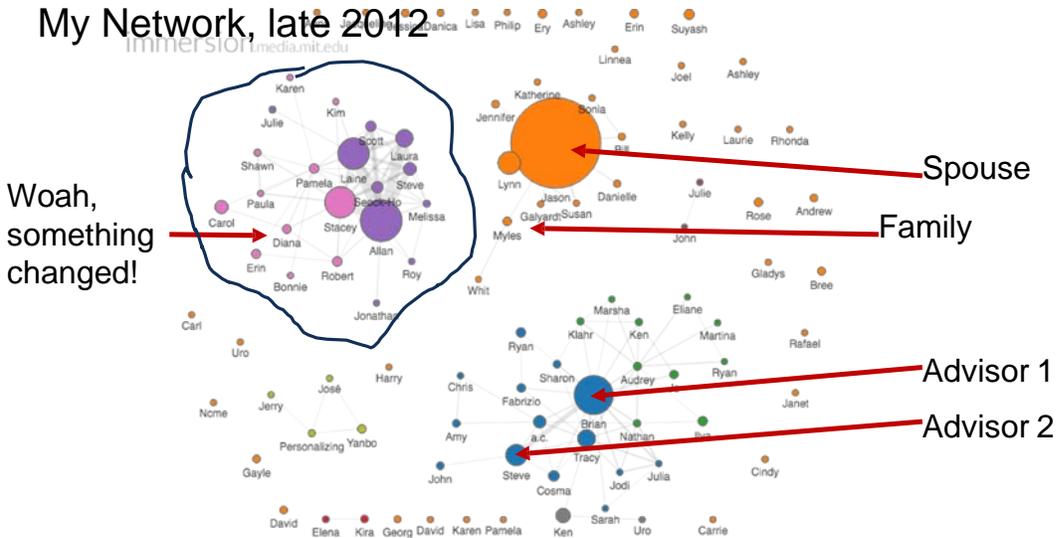
My Network, early 2012



**019 Now, early 2012, there's a new cluster forming.

And if you move to later 2012, there's a big change that's obvious.

My Network, late 2012



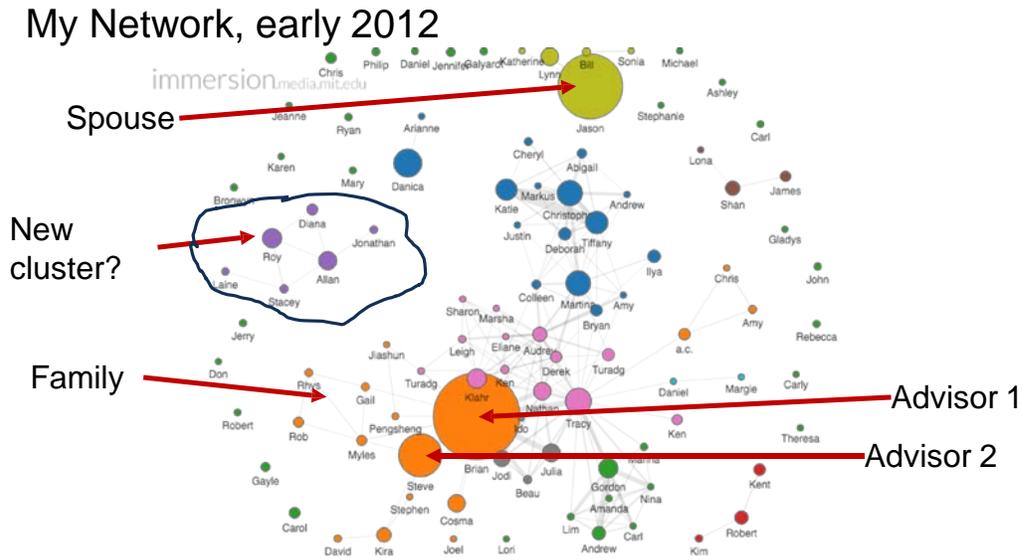
**020 There's this big new cluster that is not connected to anything that's been going on before. And I showed you so many of the early slides so you could see that it was a pretty stable network. But this is a big change. And so, if you're looking at-- if you're the NSA, or you're looking at terrorism suspects or criminal suspects, this would be an indication that something changed, and something is about to go down.

Speaker: Sure. You maybe joined a group.

Speaker: I maybe joined a group.

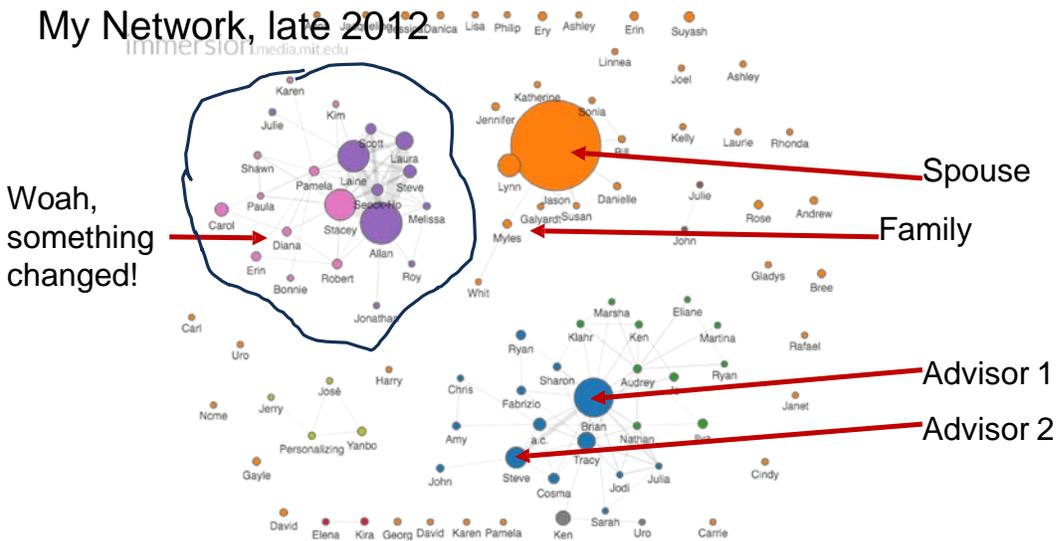
Speaker: Started talking to some new people.

My Network, early 2012



**019 Speaker: Yeah so, here you can see the cluster maybe starting to form. That's maybe suspicious.

My Network, late 2012



**020 And now, whoa, there's something different. And so, that indicates that maybe a person's been radicalized or something. In my case, I got a new job. I graduated. I was done with my PhD.

Speaker: Sure, that's a likely story.

Speaker: And I got a job. Yes, clearly.

Speaker: Yeah, no that makes sense. And then the fact that people in that group are not connected to anyone else in your network is also relevant, right?

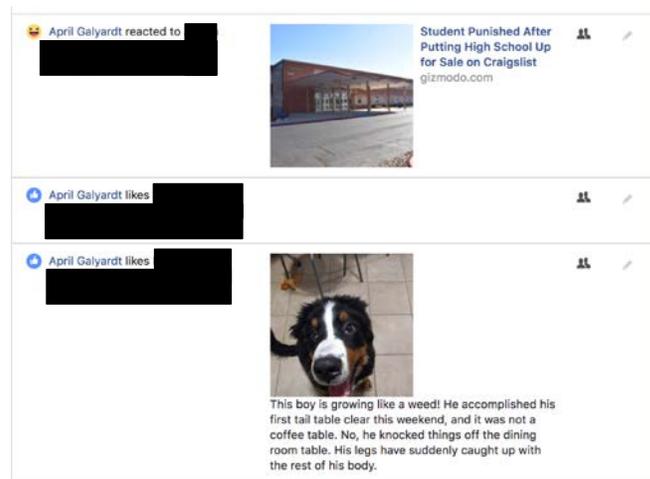
Speaker: Right, that could be somewhat alarming because if I've got a brand-new group of friends that aren't connected with any of my old

friends, that could be a big red flag.
Speaker: Interesting to note.

Speaker: Yeah, at a bare minimum
it's interesting.

What if we only see you, and your likes?

What if we only see you, and your likes?



**021 So, that's like if you see just
some basic connections. I know you.
You know me. You can still see a lot
of information there. But what if we
see only your likes?

October 2012

October 2012

PNAS PNAS

Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski^{a,1}, David Stillwell^a, and Thore Graepel^b

^aFree School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and ^bMicrosoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachtler, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and Caucasian Americans in 95% of cases, and between Democrats and

browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically and accurately estimate a wide range of personal attributes that people would typically assume to be private. The study is based on Facebook Likes, a mechanism used by Facebook users to express their positive association with (or “Like”) online content, such as photos, friends’ status updates, Facebook pages of interest,

Carnegie Mellon University
Software Engineering Institute

Digital Footprints
© 2013 Carnegie Mellon University

Distribution Statement A: Approved for public release and unlimited distribution.

22

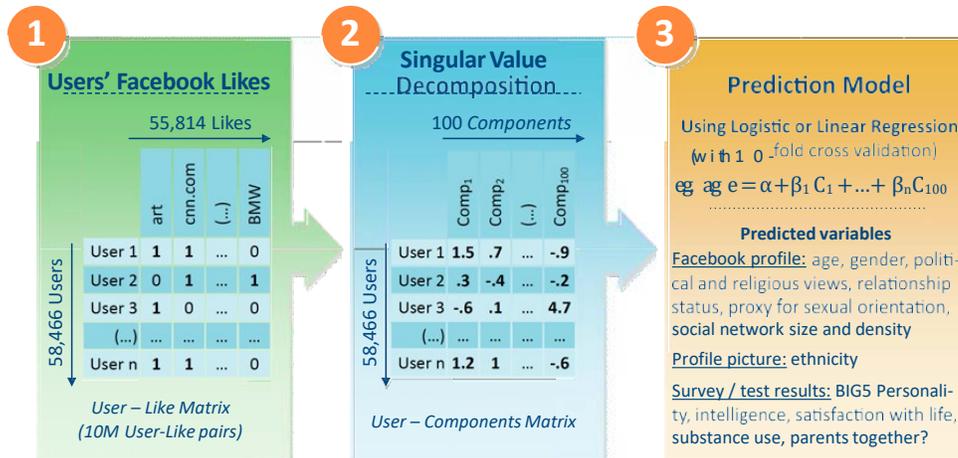
**022 So, this is a paper from 2012.
And this is a legitimate research paper. And I want to emphasize that the people who did this followed the rules.

Speaker: Okay.

Speaker: Because they actually set up a survey and had people fill out the personality quizzes and a lot of things because they wanted to see how predictive likes were and just kind of what information was there. And they did not share the data. They followed the research rules. But this study was kind of the inspiration for what went on with Cambridge Analytica because they did almost exactly what these guys did.

Kosinski, Stillwell, & Graepel (2012, PNAS)

Kosinski, Stillwell, & Graepel (2012, PNAS)



**023 So, here, the mathematics in this paper again are pretty simple. It's linear algebra. They took users and what they liked. And they-- linear algebra to find principle components, eigen values, and then they just put it in regression. Can we predict gender? Can we predict race?

Speaker: For people in the audience, I guess, for whom that's not obvious. This is a-- you can teach an undergraduate to do this in one week kind of level of difficulty.

Speaker: Right this is--

Speaker: This is not hard math.

Speaker: No, not at all.

Speaker: Right. We make sure everybody who is interviewing for our

group would be able to do this in their sleep, for instance.

Speaker: Right.

Speaker: And I also want to point out that the user/like relationship here is similar in an interesting way to the I've communicated with you relationship because it's just now the edge is we're in the same group again.

Speaker: Exactly. And now, we're-- this analysis-- and it's important to-- this is just like me and what I like. And so, the edge is between me and what I like. And the fact that I know you is not used in this data at all. What my friends like, anything like that, this analysis ignored that. So, this is super simple mathematics using very much a reduced form of the available data.

Speaker: Sure, so even if I don't know-- even if I'm in this group about dogs, and I don't know the other person in this group about dogs, we may still be similar enough because of our shared interest.

Speaker: Right.

Speaker: That's what's going on here.

Speaker: Yeah.

Speaker: So, we had an attendee question real quick. Just back from

the Paul Revere slide.

Speaker: Okay.

Speaker: He's asking-- from Joseph asking, "Why is the central figure in your network the one you care about?"

Speaker: Oh, sure.

Speaker: Yeah so, oh gosh, that's a lot of slides ago. I'm going to not go all the way back.

October 2012

October 2012



Private traits and attributes are predictable from digital records of human behavior

Michal Kosinski^{a,1}, David Stillwell^a, and Thore Graepel^b

^aFree School Lane, The Psychometrics Centre, University of Cambridge, Cambridge CB2 3RQ United Kingdom; and ^bMicrosoft Research, Cambridge CB1 2FB, United Kingdom

Edited by Kenneth Wachtler, University of California, Berkeley, CA, and approved February 12, 2013 (received for review October 29, 2012)

We show that easily accessible digital records of behavior, Facebook Likes, can be used to automatically and accurately predict a range of highly sensitive personal attributes including: sexual orientation, ethnicity, religious and political views, personality traits, intelligence, happiness, use of addictive substances, parental separation, age, and gender. The analysis presented is based on a dataset of over 58,000 volunteers who provided their Facebook Likes, detailed demographic profiles, and the results of several psychometric tests. The proposed model uses dimensionality reduction for preprocessing the Likes data, which are then entered into logistic/linear regression to predict individual psychodemographic profiles from Likes. The model correctly discriminates between homosexual and heterosexual men in 88% of cases, African Americans and European Americans in 85% of cases, and between Democrat and

browsing logs (11–15). Similarly, it has been shown that personality can be predicted based on the contents of personal Web sites (16), music collections (17), properties of Facebook or Twitter profiles such as the number of friends or the density of friendship networks (18–21), or language used by their users (22). Furthermore, location within a friendship network at Facebook was shown to be predictive of sexual orientation (23).

This study demonstrates the degree to which relatively basic digital records of human behavior can be used to automatically and accurately estimate a wide range of personal attributes that people would typically assume to be private. The study is based on Facebook Likes, a mechanism used by Facebook users to express their positive association with (or “Like”) online content, such as other people’s status updates, Facebook pages of interest,

**022 Yeah so, the central figure in the network, that gets used in a couple of different ways. So, one of the earliest things that advertisers did with network analysis and kind of looking at this was looking at well, the central figure in the network. They're the one that's connected to everybody. So, if I can give them my

product, and then they tell their friends, I will sell more product. And you see a lot of that kind of thing still with the highly-rated people on YouTube. If you've got a makeup show, they're giving you all kinds of lip gloss and things because you're going to sell more of it.

But also, from a national security perspective, if you're looking at the central figure, that's the one that everybody has to communicate with. So, if he's not the leader, then he's the one in charge of communications between groups. And so, that's still a high-value target for whichever set of purposes you have.

Speaker: Sure, that person knows a lot no matter what their actual role is within the organization.

Speaker: Right. And if we think about Paul Revere's role in the revolution, he wasn't the guy that everybody looked to. He didn't wind up president. But he was the one that connected everybody. He's the one that got everybody going.

Speaker: Yeah, I think also to bring in a concept that I've heard from design is like there's no average person. People are too very much diverse in order for you to say-- to design toward the average person. But there are groups of people who have a representation. And so, if you know this person is central in some network, a lot of people who are like-- a lot of people who are similar to that person, you can I guess make some

assumptions about them based on the fact that you know that they're pretty similar. And that degree of centrality or of being in the middle is a good proxy for a lot of people are like this person. And what works on that person is likely to work for other people who are similar to them.

Speaker: Right, it can help you find somebody who's representative.

Speaker: Yeah.

Speaker: That's a great point.

Speaker: Yeah, thank you for that question. And please, if you have any other questions, please keep them coming to us.

Speaker: Yeah, give them to us.

Speaker: We have one more in the queue. And this may be jumping the gun, so feel free to push on the end. But Ezra wants to know, "What would be the ultimate solution for data leakage prevention?" So, that's something we can push off to the end, or if it's relevant feel free to-- yeah.

Speaker: I'll go ahead and give a plug now that we're going to talk a lot about that in the next webinar. This is the first of a two part series. So, that's-- the next one is going to talk a lot more about that. But we'll try and--

Speaker: Even at the end of this webinar, actually I have some

material on good practices. Spoiler alert, there's no magic hammer, but there are some things you can do to make your life better at least.

Speaker: And another one just came in from Ellie asking, "In regard to important people in the online social networks, are they always the same people, the important people in real-life networks?"

Speaker: No.

Speaker: No.

Speaker: Okay.

Speaker: Online life is real life. No, reject the premise.

Speaker: Well, there's important in what way? There's-- because the people who are central in a network, they're-- the idea is that they're often the influencers, maybe the people that you listen to, or the people that can reach a lot of people. They're not necessarily the leaders that we think of as important in real life, the stand up and follow me. It's almost two different meanings of important.

Speaker: Sure, and they have the power to influence your ideas as well.

Speaker: Right because they can put ideas in front of you.

Speaker: Yeah, if I'm very similar to a lot of people, I'm very relatable to a lot of people. And something that matters to me, even if I don't have to

push very hard, it's like marketing but more organic.

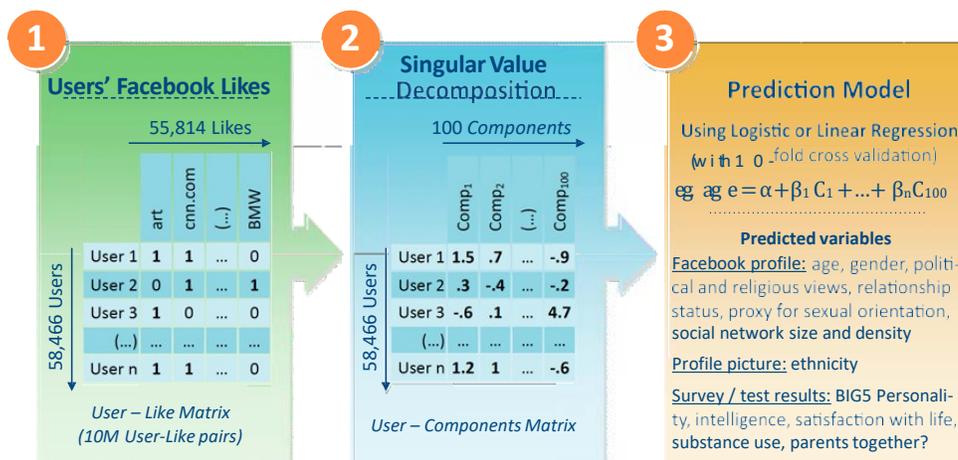
Speaker: And we can get into the social science, but one of the between people and the loose connections and being the person that connects one group to another group, those are, in real life, very valuable people to know. And we-- but I want to come back to that because that's maybe getting a hair off topic.

Speaker: Yeah, sorry. You were--

Speaker: Right so, here--

Kosinski, Stillwell, & Graepel (2012, PNAS)

Kosinski, Stillwell, & Graepel (2012, PNAS)



**023 They're just using the likes to try and see if we can predict a few things, age, gender, religious views, different sorts of things like that.

Kosinski, Stillwell, & Graepel (2012, PNAS)

Kosinski, Stillwell, & Graepel (2012, PNAS)

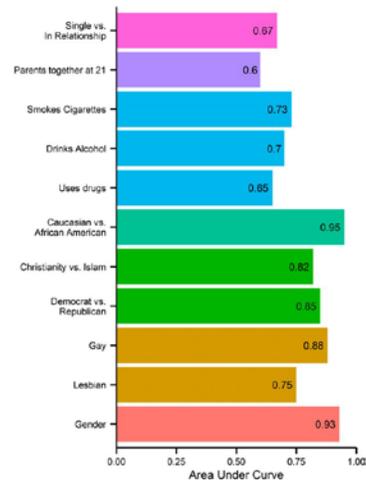


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

**024 And these are their results.

And some of them-- so, uses drugs was a harder one to predict. And that had an AUC score of sixty-five.

Speaker: The number here, higher means more accurate?

Speaker: Higher means more. Higher means that is incredibly easy to predict.

Speaker: Got it.

Speaker: So, predicting race, that was ninety-five percent accurate. And that's just from likes. But if you think about it, you think about music and all the things that kind of separate a lot of people, that is maybe not surprising.

Speaker: This is culture, yeah.

Speaker: Right. Gender, ninety-three, very easy to separate there. Sexuality, it's not quite as high as race and gender, but it is at that point eight. And it is very easy to predict just from likes as is political affiliation. Drug usage is a little harder. And predicting whether or not parents were together, that's again a little harder.

Speaker: Sure so, people will probably at least attempt to conceal some of these things.

Speaker: Right.

Speaker: Is it-- for instance, if they-- can the algorithm tell my sexuality just because I like sexuality-related pages or--

Speaker: Well, and--

Kosinski, Stillwell, & Graepel (2012, PNAS)

Kosinski, Stillwell, & Graepel (2012, PNAS)

Strong Predictors of male homosexuality included

- “No H8 Campaign,”
- “Mac Cosmetics,” and
- “Wicked The Musical,”

Strong Predictors of male heterosexuality included

- “Wu-Tang Clan,”
- “Shaq,” and
- “Being Confused After Waking Up From Naps.”

**025 That's a great question.

Speaker: Oh, okay.

Speaker: So, these are the ones that they published. And so, for the predictors of male homosexuality, liking a California House Bill 8, which was related to that--

Speaker: Yeah.

Speaker: That was very predictive. But the other two are not. Cosmetics, okay yeah, it's very easy to see that most straight guys would not like a cosmetics page. But "Wicked," the musical, that's just "Wicked."

Speaker: Sure.

Speaker: And why is being confused after waking up from naps predictive of heterosexuality?

Speaker: Right.

Speaker: And so, this is-- a lot of these things are not necessarily things we would think of as indicative of-- okay-- oh, I'm trying to-- Mavis Staples is coming. And she's going to go and sing here on Friday night.

Speaker: Sure.

Speaker: And she's awesome. What am I telling people about me when I say I like Mavis Staples?

Speaker: It's a lot more than just saying you like that one performer. It can say you like a genre. It can say that you subscribe to a political world

view, if they're related. It can say that you have--

Speaker: She does have some of those songs, yes.

Speaker: Sure. I apologize for not catching this reference, so I am just speaking. But it can tell you something about social economic status. There's lots of things you can tell about a person by just the fact that they like one performer.

Speaker: Right.

Speaker: And that's one of the really amazing things about these predictive models is that just a human might not have guessed that being confused after waking up from naps was something that was indicative of sexuality. But it turned out that that was the case just after having this amazingly rich field of data.

Speaker: Right, it was useful.

Speaker: Yeah.

Speaker: One of many things is the top three, going through the principle components analysis, they would have had a very long list. They just gave us the top three.

Speaker: Yeah, got it.

Kosinski, Stillwell, & Graepel (2012, PNAS)

Kosinski, Stillwell, & Graepel (2012, PNAS)

How hard do you think it would be to predict “parent”?
“gun owner?”

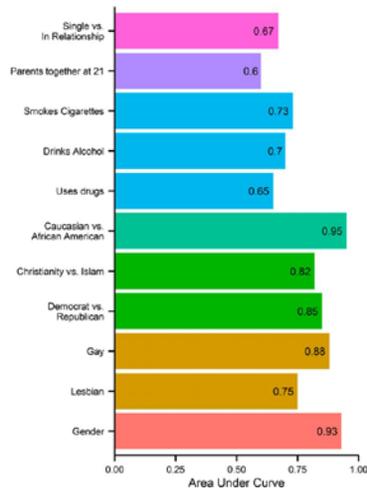


Fig. 2. Prediction accuracy of classification for dichotomous/dichotomized attributes expressed by the AUC.

**026 So, if you kind of think about this, and you think about okay, it's that easy to predict whether or not somebody smokes, it's that easy to predict what gender somebody is, how hard is it to predict parent? I tend to think that would be pretty easy. How hard is it to predict gun owner?

Speaker: Sure.

Speaker: These are things that maybe we don't make an effort to conceal. But at the same time, we don't-- we aren't always aware that we're giving this information away.

Speaker: Sure, or I guess to get a little bit more "Black Mirror," how hard would it be to predict likely to commit a certain kind of crime within the next year?

Speaker: Well, there's actually a lot of algorithms out there that people are working on that. I mean people are trying to predict, for criminal sentencing, recidivism. So, if we have somebody in front of a judge, and the judge is trying to decide what kind of sentence they should get, whether or not we think they're likely to recommit a crime is very pertinent to what kind of sentence the judge wants to give. But those algorithms are colored by so many different things.

Speaker: Sure.

Speaker: And that's maybe another talk entirely.

Anonymization is Easy to Break

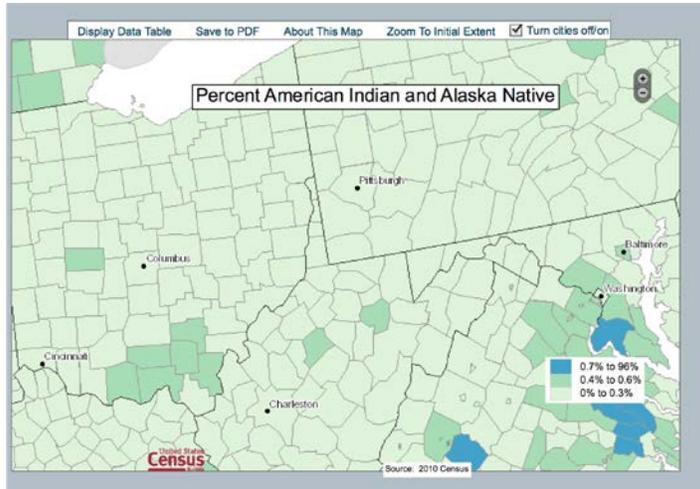
Anonymization is Easy to Break

**027 So, the next point here, anonymization is easy to break. Not just like it can be done, no, no, it's

easy. So, when you think data is anonymous, there's almost no such thing.

Anonymization is easily broken

Anonymization is easily broken



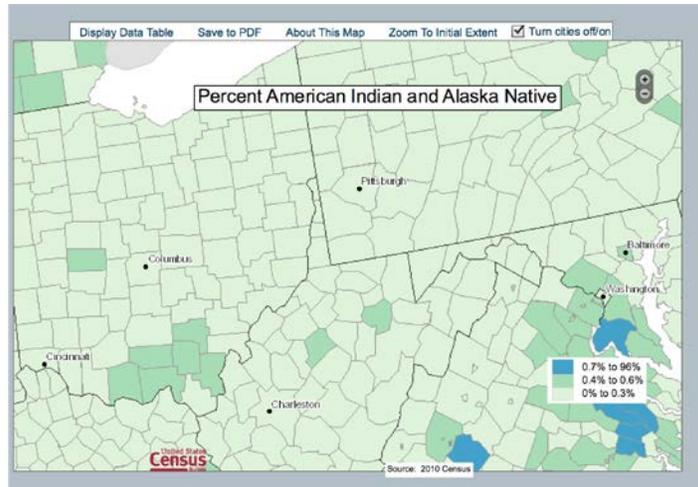
Earlier versions of the Census Data Mapper allowed you to map demographic information on a neighborhood level.

Now, only county-level maps are available.

**028 And the census has actually known this for a very long time. They've got a data mapper that will allow you to find some things out about your region. And in earlier versions of this, maybe ten years ago, that allowed you to get down and map demographic information on a neighborhood level, and you could look at-- you could go into the Pittsburgh maps, and you could see what are the rich neighborhoods, where are the poor neighborhoods, and get into a really fine grain level of detail. They don't let you do that anymore.

The US Census

The US Census



53% of the U.S. population can be uniquely identified if you have

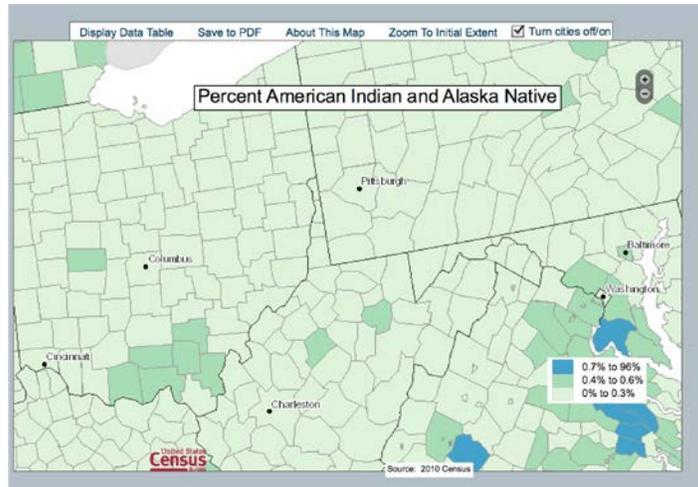
- place,
- gender,
- date of birth

Sweeny (2000)

**029 Because, as they were worried about privacy, Stephen Fienberg was one of the guys who was really helping them pay attention to this, that they figured fifty-three percent of U.S. population-- if you just have a place like I live in Pittsburgh, I am female, and if you have my birthdate, that's enough to identify me uniquely for fifty three percent of the population. But place, Pittsburgh has quite a few zip codes in it.

The US Census

The US Census



87% of the U.S. population can be uniquely identified if you have

- place => Zip Code,
- gender,
- date of birth

Sweeny (2000)

**030 So, if you have instead of just Pittsburgh, you have an actual zip code, you can identify eighty-seven percent of the U.S. population.

Speaker: Okay so, this is without knowing, for instance, your name. This is only with zip code, gender, and date of birth.

Speaker: Right.

Speaker: You can uniquely point at a single person.

Speaker: At a single person for eighty-seven percent of the population. And so, that's-- the fact that it's not just our fingerprints that are so unique, it's-- there's that much that makes us us.

Speaker: Right, and teaser for the end of the talk, there's a direct

analogy with your online habits that actually don't have anything to do with, a priori, your zip code, your gender, or date of birth, you can do a very similar kind of thing with just what websites you visit.

Speaker: And so, they, since 2000, have known that you can break anonymization, that things that are supposed to be anonymous are not.

The Netflix Prize

The Netflix Prize



**031 And Netflix found this out when they did the Netflix prize.

The Netflix Prize

The Netflix Prize

arXiv.org > cs > arXiv:cs/0610105

Computer Science > Cryptography and Security

How To Break Anonymity of the Netflix Prize Dataset

Arvind Narayanan, Vitaly Shmatikov

(Submitted on 18 Oct 2006 (v1), last revised 22 Nov 2007 (this version, v2))

We present a new class of statistical de-anonymization attacks against high-dimensional micro-data, such as transaction records and so on. Our techniques are robust to perturbation in the data and tolerate some mistakes. We apply our de-anonymization methodology to the Netflix Prize dataset, which contains anonymous movie rental records from the world's largest online movie rental service. We demonstrate that an adversary who knows only a little bit about this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we uncover records of known users, uncovering their apparent political preferences and other potentially sensitive information.

Subjects: **Cryptography and Security (cs.CR)**; Databases (cs.DB)

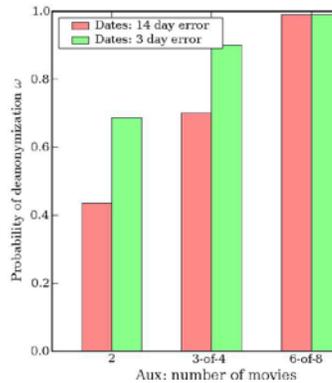
Cite as: [arXiv:cs/0610105 \[cs.CR\]](https://arxiv.org/abs/cs/0610105)
(or [arXiv:cs/0610105v2 \[cs.CR\]](https://arxiv.org/abs/cs/0610105v2) for this version)

**032 They had to discontinue this because these guys look at the-- Netflix was releasing anonymous data and having people put out-- try and come up with the best algorithm to predict what other movies people are going to like. They wanted to upgrade their recommendation engine. And these guys took anonymous Netflix data and the IMDb data, and they matched it up.

And they were able to match a huge number of records just from movie habits.

The record of movies you watch is a fingerprint

The record of movies you watch is a fingerprint



68% of records can be re-identified with 2 movie ratings and dates accurate to within 3 days

99% of records can be re-identified with 8 movie ratings (2 of which might be wrong) and dates accurate to within 2 weeks.

Figure 1: De-anonymization: adversary knows exact ratings and approximate dates.

**033 So, if you rated two movies on IMDb, and those dates are reasonably lined up with the Netflix dates, they got sixty-eight percent of the records matched up that way.

Speaker: Okay.

Speaker: They got ninety-nine percent of the records that they had rated eight movies. So, eight data points-- it's really only six because two of the eight might be completely wrong.

Speaker: Okay.

Speaker: And dates accurate to within two weeks. So, that's, by and large, I watched a movie on Netflix, and then like half a month later, I'm on IMDb, "Oh yeah, I liked that." So, ninety-nine percent of people, so behavior of the movies, those

timestamps of the movie, again, it's almost like a fingerprint.

Speaker: Okay so, the data that was involved in this was IMDb ratings and then Netflix watching history?

Speaker: They had Netflix ratings and dates on-- so, they had--

Speaker: Oh, it was also reviews for Netflix?

Speaker: Not any text.

Speaker: Okay.

Speaker: Just this was when Netflix was still using stars.

Speaker: Oh, okay.

Speaker: So, it was one star, five stars.

Speaker: Got it. And so, Netflix scrubbed the people information from it.

Speaker: Right.

Speaker: And just kept the--

Speaker: This person liked this movie, this movie, and this movie.

Speaker: Got it.

Speaker: And for each movie they watched, they had the rating. And so, you can think of-- if you were thinking of the dataset which you

have, you would have person, movie, rating, and date.

Speaker: Okay, got it. And so, this group was able to take a different dataset that was published by a different group.

Speaker: Right.

Speaker: Also, I guess anonymized in certain ways?

Speaker: Well IMDb, the ratings on it are public so that people can build profiles and the ratings of personas. You didn't have to use your real name on IMDb.

Speaker: But right so, they were able to take these two different datasets but then figure out who in this dataset corresponded to who in that dataset.

Speaker: Right, and they actually-- they had different ways to check this, but they actually talked to some of their friends who were-- they knew in the IMDb database and were like, "Hey, is this you?" And were like, "Yeah," so--

Speaker: Scary.

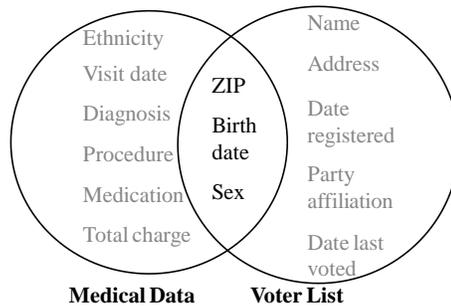
Speaker: Yeah so, that's the anonymization is-- and we've known for a while now that anonymization isn't this easy to break.

Speaker: Sure.

Health Records

Health Records

1997, Group insurance commission released *anonymized* health records from Massachusetts.



****034 Speaker:** This one is just kind of to drive the point home. So, in 1997, the group insurance commission, they released anonymized health records from Massachusetts. They had visit, diagnosis, procedure, all of this very personal information. And they had removed the names and the labels.

Speaker: But they did keep the zip, birthdate, and sex?

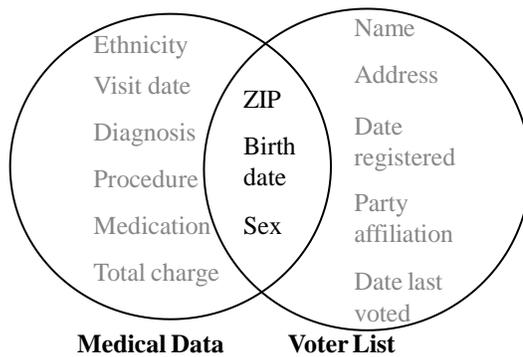
Speaker: Right. Well, I mean if you're looking at medical data, the age and gender of the patient are very important for the medical history. But that zip code, that's that third identifying piece of information. And so, they took the voter list, which has actually who you are, your name, your address, but it's also got your birthdate and your gender and

your zip code. And they matched them up.

Speaker: It'd be so easy, yeah.

Health Records

Health Records



Sweeny (2002), identified Governor Weld's medical records:

- 6 people shared his birth date
- Only 3 were men
- He was the only one in his zip code.

****035 Speaker:** And they actually identified Governor Weld, who was the governor of Massachusetts at that point in time because only six people share his birthdate. Only three were men. And he was the only one of those six people in his zip code. And so, now they had the governor's medical history. So, this is, if you go back, eighty-seven percent of Americans could be identified this way with anonymized data, with just that little information.

There are Real Consequences

There are Real Consequences

**036 And so, the final point here,
there are real consequences--

There are real consequences

There are real consequences

Marketing for the wrong products

- Annoyance for most of us
- Reveals a medical condition at work

Someone gets categorized as “Diabetes Interest”

- Ads for sugar- free products?
- High risk insurance category?

**037 When this sort of stuff

happens because we might think marketing okay, I like martial arts. So, Facebook can't figure out whether I'm female or male. And they keep showing me ads for the wrong underwear. Oh, well. But, at the same time, maybe I've got a medical condition. And the advertiser has figured out that I have this medical condition. And now, I'm seeing ads for chemo or dialysis, and seeing these ads at work. And now, everybody at work knows I have this medical condition. And that, for some places that people work, that is not a safe thing.

And you've also got the other issue. So, somebody gets categorized as diabetes interest, which that's a category data brokers use, okay so they see ads for sugar free products. Maybe that's benign. But now, they're suddenly in a high-risk insurance category depending on who has that data and how they use it.

Speaker: So, relevant question here from Joseph asking, "There's enormous value in this data to improve our lives and create economic value. How can we get transparency and control?"

Speaker: That is exactly the issue. That is, it. And there is-- these examples are-- I'm trying to emphasize the risks right now, but there are tremendous benefits. And I like the fact that Google can tell me oh, there's an accident on the bridge that you usually take home. Go another way. I love that. But that

means that Google knows how I go home. And so, that is the real question. And a lot of what we're doing, that's one of the reasons we set this up as a two part webinar was here, we're talking about what's possible and what can be done so we have the language to address these problems. And that question is really what we're going to focus on in the next one. What should regulation look like?

Speaker: Yeah, it's true. Yeah, please tune in for that one, as well. Our goal in this presentation is not to tell you that the world is bad or that data is bad. It's to show you that data is extremely useful.

Speaker: And powerful.

Speaker: And yeah, powerful. So, yeah in the next presentation, there's going to be extensive talk about what can be done, what are regulatory best practices, what should people be talking about.

Speaker: Well, and there was an editorial out recently. And I thought it really kind of hit the nail on the head because they were talking about how a lot of sciences have had a reckoning with dynamite. Dynamite is powerful, but it's also-- it can be destructive. And chemistry has had this reckoning. And medicine has had this reckoning. And is this the reckoning for computer science, that data is extremely powerful, but let's

not see things destroyed? Let's control it before we get there.

Speaker: Can we work in one more from Frank here asking, "How to prevent a document from losing its metadata after being shared in encrypted media like WhatsApp?"

Speaker: So, if you--

Speaker: I'll read it again. How to prevent a document from losing its metadata after being shard in encrypted media like WhatsApp.

Speaker: Okay, so there's two ways that he might mean losing there because losing could mean like the metadata's gone, or losing could be the metadata is now out and available. And that's kind of the thing about metadata is you can't keep it hidden because it's the outside of the envelope. If I send you an email, that record is there. And maybe see what's in the email. Maybe people don't. But the fact that that email was sent is-- that doesn't go away.

Speaker: Somebody owns these servers. I think one way that we didn't show in the video clip but that was in this same larger talks was, again, Zuckerberg was saying, "We're Facebook. We give you the opportunity to share all this data. Yeah, like we own some of this data."

Speaker: "It's what we do."

Speaker: "It's what we do, listen."
So, at some point, there's always

going to be a-- somebody's going to know where this data was going. I guess I have not heard of a communications company that allows you to say we don't know a single thing about how you're using our service. That would be an actually really cool like business model.

Speaker: Right. Signal does a lot of encryption, but I don't know-- I am not that up on exactly how Signal works to know at what level the encryption works.

Speaker: Yeah, so if there's-- if we maybe did not the address the root question that you were trying to ask, please ask it again to Shane in a more detailed way. And we'll try to get to that. Thank you.

There are real consequences

There are real consequences

Facilitates

- harassment
- stalking,

Exposes

- domestic violence victims,
- law enforcement officers,
- prosecutors,
- public officials
- ...

**038 Speaker: Okay, kind of

going to the real consequences, some of the incidental sharing of this data can lead to harassment, stalking. There's a reason that school principals never put their phone number in the phonebook because teenagers sometimes make bad choices. And so--

Speaker: Yeah. You, Shane?

Speaker: It's a thing. And so, having people who have known for a long time my phone number should not be in the phonebook, now suddenly everything is out there and exposed. That's-- there are risks there. One of the other risks, when Google created Google+, and they just kind of lumped everybody into your friends, and they made it very flat. Like Facebook is very flat. Everybody that is your friend is your friend. There's no hierarchies of friends. A lot of the people who had been emailing, they emailed their friends. But they also have emails that they send to their ex-husband. And at one point, he was not their ex-husband. And he had friends. And they might have been cc'd. And when Google+ opened that up and made it flat, suddenly they were re-exposed to that abusive relationship. And there were a lot of stor-- Google fixed that pretty quickly. But they didn't anticipate it.

Speaker: We have another question in the chat here from BJ Johnson. Thank you for the question. It says, "Scott McNealy said in 1999, 'There's no such thing as privacy any longer."

Get over it.' We've known about the situation for a long time. So, since you were showing us how gaining understanding of our data in this manner is so easy, it seems that by using social engineering, the bad guys can figure out our authentication credentials easily. How can anyone feel their data is safe at all ever?"

So, I want to start-- I think that there's a lot in this. So, thank you for that question. I want to start by saying the techniques we've talked about today are actually-- I don't think that knowing your username and password is the main thing that people are getting out of these techniques. If you are-- if you have bad username and password practices, then yes, you're at risk. But you can do simple things like don't use the same password for every single thing. Use a complicated password that's-- or use actually a password manager.

Speaker: Right. Well, this is-- a lot of what we're talking about here isn't even like authentication, how we log into the systems. This is the stuff that we think we're okay with sharing, but it leads to inferences that we might not be okay with sharing.

Speaker: Yeah, this is not authentication. It's not how can a person pretend that they're me. It's what can a person know about me and how does that influence their actions that affect me. So, there are definitely user rights best practices

that are not really related to this discussion. Your data is safe/we work at a cybersecurity group. There's always a way for somebody to gain access to the thing because there's always somebody smarter than whoever engineered your security. Usually, if you're not a super high-profile target, I feel uncomfortable saying you can probably rest easy because nobody's going to--

Speaker: Well, so I, after the Experian and Equifax and all the data breaches, my credit is locked down. So, I assume that after Target and Home Depot and Experian, somebody has my data. And so, I have my credit locked down where nobody can open an account in my name. It's just I assume that, at some point, my data's been stolen. But that's-- the hackers coming into a system and getting things that they're not supposed to have, that's almost a separate problem. It's a very real problem. It's one that we deal with in this building every day. But what we're talking about are the things that--

Speaker: I'm not afraid to share, right?

Speaker: Right, the-- it's public, right? It's just between you and me and what we can learn from the things that we think we're okay with sharing.

Speaker: Thank you for the question, BJ.

Speaker: Okay, I think we're done with that.

Real Consequences

Real Consequences

Real personal information, including

- Names
- Addresses
- Social security numbers,
- Bank account numbers

Attached to fake debts.

Packaged and sold to debt collectors.

FTC Privacy & Data Security Update: 2017

**039 So, actually, this example is-- kind of bridges that gap between the criminal activity and the what we thought we were okay with sharing this because this has happened. There were several big cases last year that were prosecuted. But people had taken identifying information, so names, addresses, social security numbers, things that they could get that were real people and real information, real bank account numbers, and then they attached them to fake debts. They made up-- so, I have your information. I know that you exist. I have your existence information. I'm not going to put you in a database and say you owe somebody ten thousand dollars. And then I'm going

to take that database of debts that I have made up and sell that because once that goes to the debt collectors, those debts have been sold. And the debt collectors can just collect on them. And so, these things-- this happened multiple times last year that the identity information was attached to fake information. And so, the debt collectors went out and were harassing people and collecting these debts, and some people paid debts that they did not owe, never mind the harassment cost of go away. And so, this is one of those that it's on that border of this is clearly criminal, but they have access to it through things that were not criminal.

Speaker: Right.

Personal Precautions

Personal Precautions

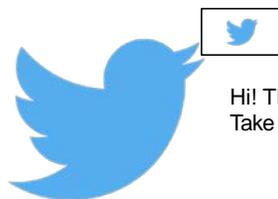
**040 Yeah so, for the rest, we

have about thirteen minutes left. And I want to talk about some good practices, some personal precautions. I'm even hesitant to say best practices because again, there's always someone wiler than you. But these are some things that are kind of a baseline here's how you can make yourself aware of what you give away even when you're not actually interacting with like a social network. These are just things that you leave around when you're using the Internet.

Cookies

Cookies

Small text files stored on your machine by websites you visit, to remember information specific to you



Hi! Thanks for logging in!
Take this card and show it to me next time you visit.

**041 So, for one, I want to talk about cookies. So, a cookie is not code. A cookie is a small text file. It's just data, just data, that's stored on your machine by websites that you visit in order to remember information that's specific to you. So, these were designed in order to

make the Internet more convenient to use. The use case that they were designed under was if you're using a shopping like website, and you want to close that window, do some other stuff in your day, come back and open the window and have your shopping cart remain, that's what a cookie is for because they want to be able to save on your machine some of the history of what you've been doing on that website. It's designed to make your life less annoying.

Another thing that it can be used for is authentication so that you don't have to type your password every time that you go to another page on the same website. The website wants to remember that you've already proven you are who you are. So, in this analogy, a cookie is like a card. And you're using a service, let's say Twitter. And Twitter says, "Hi, thanks for logging in. Take this card. Show it to me next time you visit."

So, why might they want to do that?

First-Party Cookies

First-Party Cookies

Cookies set or requested about the domain you're visiting



Hi! Can you show me your  ?
I want to see if you're logged in.

****042** Well, so there are actually two distinguishments to be made about cookies. And that's in their origin and in who's asking for them. So, a first party cookie is when a cookie is set or requested about the domain that you're visiting. So, you're visiting Twitter. And Twitter says, "Hi, can I see your Twitter cookie? I want to make sure that you're logged in. That way I don't have to bother you with typing your password again." This is in order to make your life easier.

Third-Party Cookies

Third-Party Cookies

Cookies set or requested about a domain other than the one you're visiting



****043** There's also a third-party cookie which is a cookie set or requested about a domain other than the one that you're visiting. So, if you're reading the news, let's say you're on CNN, and CNN wants to populate a box that allows you to post about this article to Twitter, CNN might say, "Can you show me your Twitter cookie? If you're logged in, I'll add a tweet box for you. And if not, then I will not." It's a way to make these websites more interactive. So, a third-party cookie is the website that you're on asks about information from a website that you're not on right now. So, that can be used for a fair-- I think that's a pretty benign use case, but the same kind of concept can be used for advertisements or for tracking in general.

Third-Party Cookies

Third-Party Cookies

Cookies set or requested about a domain other than the one you're visiting



Hi! I see that you're on a website about Pittsburgh public transit. I'll record that on and show you relevant ads.

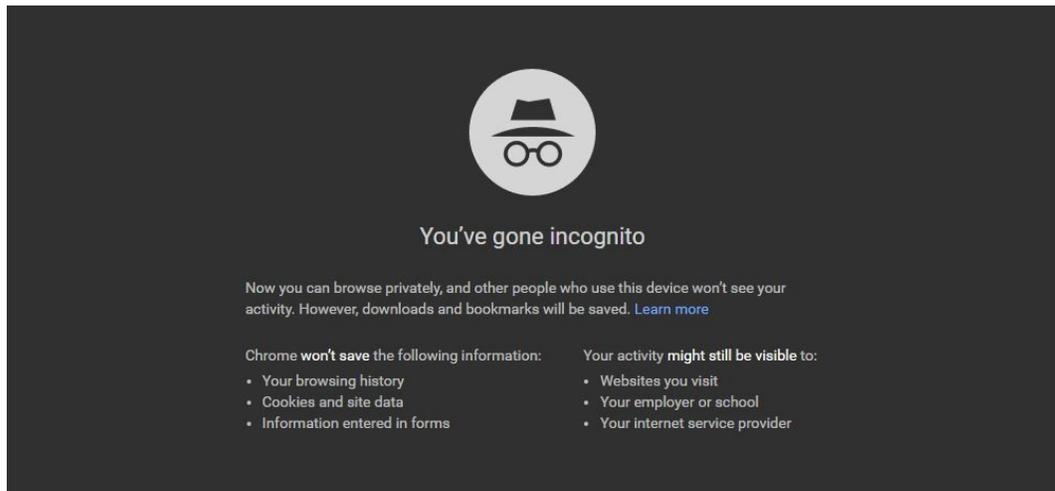


**044 So, Google AdSense is a very widely used ad platform. And so, AdSense says, "Hi, I see that you're on a website about Pittsburgh public transit. I would like to record that information. And can you please hold that information about yourself on your machine? And then next time you visit a different website, let's say it's about dogs, I want to record that you're interested in both Pittsburgh public transit and dogs and maybe one about babies." And so, you're probably never going to the AdSense website when you are surfing the Internet. But if you do not disable these third-party cookies, just a totally, a separate third party can acquire this information about you and build a profile. And it's specifically to see what you're interested in and give you ads that it'll think you care about more.

That's becoming closer to malice. And of course, if that's possible for advertisers to do, it's possible for people who like actually want to hurt you to do. And so, the great thing is--

Incognito/Private Browsing

Incognito/Private Browsing



**045 You can do something about it. So, for one is incognito or private browsing. You're probably already aware of this. But one of the things that this is really useful for is that it doesn't save anything on your machine from your website browsing. So, it will not save your browsing history. And it will not save your cookies and site data. So, this does not mean that everything you're doing is obscured from the world entirely. In fact, I mean as it says on the incognito site or the splash screen, your employer can see what websites you're still visiting. Your Internet Service Provider can still see

what websites you go to. This is saying I don't want to save that data on my own machine. And so, this is really useful for lots of reasons. One of them would be blocking cookies to you don't want to have information-- you don't want to facilitate the gathering of information about you.

Disabling Third-Party Cookies

Disabling Third-Party Cookies

First-party cookies are generally not harmful, don't need to be turned off

Third-party cookies can be disabled in each browser you use

Details vary based on which browser you use. E.g. some browsers only disable *setting* third-party cookies but don't disable *reading*

**046 Another thing you can do is actually just disable third party cookies. And so, first party cookies, the ones that are requested or set by the domain that you are trying to visit, are generally okay. They're usually there to make your life easier and don't need to be turned off. But third-party cookies can actually be disabled in each browser that you use. And so, you can just kind of search for I use, let's say, Safari or Chrome, and different browsers are going to differ in the settings that

they allow. But all of them enable some form of disabling third party cookies. Some of them will disable, for instance, setting, but not disable reading, which can still be problematic, but less so than just freely doing cookie things.

Speaker: Cookie things.

Browser fingerprinting

Browser fingerprinting

Websites ask your browser for information about itself and your computer so that content can be displayed most effectively, e.g.

What browser and operating system are in use?

What plugins are installed?

What is the screen resolution and color depth?

What fonts are installed?

This information can be used to **uniquely identify you**.

**047 Another totally different thing is so what if I actually just totally did not allow cookies to be transmitted or received at all? Can somebody still tell about me? And the answer is yes. So, browser fingerprinting, there's a source on this which is listed on the next slide. But you can go to amiunique.org for more information. But a website in this modern era will ask your browser for information about itself and about your computer

so that it can display content most effectively and most beautiful to you.

So, some things it's going to want to know, what browser and operating system are you using in order to know that I'm compatible at all, what plugins are installed, so how should I modify the screen based on that, what fonts are available, what screen resolution is there. So, that's just stuff about your computer. There's nothing that you would think is about you about any of that information. However, that information taken together, can be used to uniquely identify you.

Speaker: So, it's the equivalent of the birthdate and zip code and place.

Speaker: It is exactly like that. So, the point of this is they don't have to ask-- yeah.

Browser fingerprinting

Browser fingerprinting

Are you unique?
Yes! (You can be tracked!)

3.44 % of observed browsers are Safari, as yours.

0.12 % of observed browsers are Safari 11.0.2, as yours.

13.23 % of observed browsers run Mac, as yours.

3.52 % of observed browsers run Mac 10.12, as yours.

62.65 % of observed browsers have set "en" as their primary language, as yours.

4.73 % of observed browsers have UTC-4 as their timezone, as yours.

However, your full fingerprint is unique among the 714241 collected so far. Want to know why? [Click here](#)

source: amiunique.org

Carnegie Mellon University
Software Engineering Institute

Digital Footprints
© 2018 Carnegie Mellon University

Distribution Statement A: Approved for public release and unlimited distribution.

48

**048 They don't have to ask what's your name. This is not like somebody at the store knowing what your name is. This is like somebody at the store knowing oh, I've seen you before. You're the one with the brown hair and the thick eyebrows. And you bought diapers last week. I recognize you. And I can also talk to my friend who works at a different store and say yeah, they came in and bought diapers again. So, this cannot be disabled.

You can check out amiunique.org in order to figure out whether you are unique. In this case, this is a screenshot from my machine. And you can see only point one two percent of observed browsers were using the browser version that I did. So, that's already a super small number of people just without me knowing anything. And it turns out

that there's actually one of the data attributes that's hidden here is-- makes me completely unique. They're able to tell-- assign a unique ID number to me and say-- and watch me across the web. So, just like be aware that this can happen to you and that people can recognize you even if they don't know your name.

Speaker: I would go so far as not it can. It has.

Speaker: Yeah, they do.

Fingerprinting and cookies example

Fingerprinting and cookies example

Wall Street Journal 2012:

“Orbitz Worldwide Inc. has found that people who use Apple Inc.'s Mac computers spend as much as 30% more a night on hotels, so the online travel agency is starting to show them different, and sometimes costlier, travel options than Windows visitors see.”

Orbitz also used this data to influence ranking:

referring site: the site a user follows a link from to get to Orbitz
return visits: booking history and previous activity on the site
location

**049 And so, to give you an example of this in usage, in 2012, the Wall Street Journal talked to, or in some way interacted with, Orbitz. And Orbitz found that people who use Apple Mac computers are likely to spend more on hotels. And so, there was some kerfuffle about the

reporting. But eventually, it came out they don't give you a higher price for each hotel room, but they do, in their search results, sort higher priced hotels more to the-- closer to the top if they detect that you're using a Mac computer. And that's interesting. Is this malicious or not?

Speaker: It feels kind of malicious.

Speaker: So, I don't know, right? So, in an analogy, the point of Google is, as Google's searching, is to give the most relevant stuff close to the top.

Speaker: Right.

Speaker: There's going to be a million search results for every Google search that you do, but you want to see the stuff that you care about on the first page.

Speaker: Right, I want-- Google knows for me that when I type EDM, I mean educational data mining, not electronic dance music.

Speaker: Yeah, sure. And that's not annoying to you in order to get the right result. So, listen, they're keeping a profile on you. This is how the Internet works right now. It is not, on its own, malicious. But it is just how it works. So, there's other attributes that they used in order to influence the ranking. And these are just things that you-- that are not about like-- they're just about how you use the Internet.

Revoking access

Revoking access

Apps and Websites Logged in With Facebook

The screenshot shows the Facebook 'Apps and Websites' interface. At the top, there are three tabs: 'Active', 'Expired' (which is selected and highlighted in blue), and 'Removed'. The 'Expired' tab has a small circle with the number '17' next to it. To the right of the tabs is a search box labeled 'Search Apps and Websites'. Below the tabs, there is a message that reads: 'Data Access Not Allowed. These are apps and websites you've logged into with Facebook and may not have used in a while. They may still have access to info you previously shared, but their ability to make additional requests for private info has expired. [Learn More](#)'.

**050 And I just want to spend a couple more minutes talking about some best pra-- like some-- again, more good practices. So, this is specific to Facebook, but this kind of thing can be done with your other--

Speaker: I want to make sure we have time to get another question.

Speaker: Oh, sure. Okay.

Speaker: We'll--

Speaker: We're good? Okay.

Speaker: Yeah. Get through your slide, and then we can--

Speaker: Awesome.

Speaker: So, you can revoke access to apps that ask for access to your online credentials. What that means

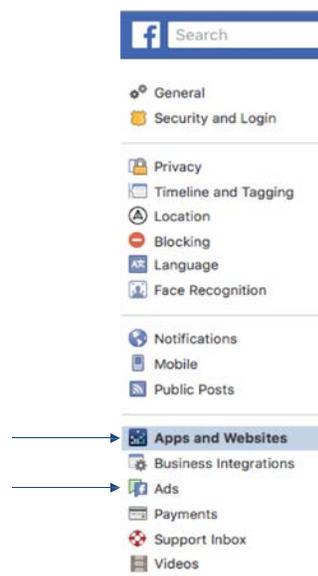
is that you say you're no longer allowed to gather information on me. It does not mean you must delete all the information you have on me. And in fact, that is just not available. That option is not there.

Speaker: Not in America.

Speaker: Yeah.

Seeing what's known about you, and controlling what's shared

Seeing what's known about you, and controlling what's shared



**051 And finally, if you do use Facebook, I highly recommend checking out these two tabs that are highlighted here, the apps and websites, and ads. I did this yesterday and was shocked, although I shouldn't have been, at how much they knew about me and how they were tailoring their content to me. Just-- I'm not going to say stop using Facebook. But I will say it would be

really a good idea to know how much they know about you.

Speaker: Right, well and you're going to be surprised because one of the things I bought something from another online retailer. And they had given my email address to Facebook. And so, Facebook knew that I had made a purchase from them even though I never told Facebook I made that purchase. And that's, again, some of the data that came-- that was available through that Cambridge Analytica API that-- did I agree to tell everybody I did business with this retailer?

Speaker: Sure. So, yeah, that is all the-- I guess, no, we've got a--

Takeaways

Takeaways

1. Metadata is Data.
2. There's no such thing as anonymous.
3. There are real consequences.

**052 Wrap up slide here. And then we can take some questions.

Speaker: So, the three big takeaways, I mean if you remember nothing else that we've talked about, metadata is data. There's no such thing as anonymous. And there are real consequences.

Speaker: Great, so we do have some backlog of questions here. So, we've got about a minute. So, we may go a minute over. So, we understand if people have to leave at two, but we'll get through as many questions as we can here. You guys mentioned earlier, this was not to cause fear. It was just to show you where data is-- how it's being used. So, there was a question that came in, "Do you have specific examples or case studies of these types of data being used for good things?"

Speaker: Sure, I mean neutral would be Google ranking relevant things closer to you. Good things--

Speaker: I have a pretty good one that there's a couple of-- kind of like there's Doctors without Borders, there's a couple of statisticians and data miners groups along those lines. And one of the-- there was a hackathon a little while ago, it was a couple years ago, but it stuck out because they had done-- they took all the records of who owned which properties and what complaints had been filed and things that are normally kind of all over everywhere. And they pulled it together. And they did some of the same social network analysis. And they were able to

identify the central landlords in this network graph and encourage the DAs-- provided evidence so that those delinquent landlords could be prosecuted. And also, let them know which targets they needed to go after first. So, it enabled the prosecution of some criminal activity in that case.

Speaker: A related-- or actually, a different example would be healthcare analytics is actively trying to help healthcare teams figure out whether you're at risk for certain things. So, this can be used to leak your diabetes information. But it can also be used for your doctor to say, "Oh, you're like these other one hundred people who had this particular illness. We know how to treat you. And we know how to contact you and say you're at risk."

Speaker: Right, and I think that that's a great example because it shows it's a two-edged sword. There's power for good. And there's power for not.

Speaker: So, a question from Michael, "With both your backgrounds in machine learning, what one book or resource would you recommend to a serious student of machine learning?"

Speaker: If you want to get into the math of it, I highly recommend Hastie and Tibshirani's "Elements of Statistical Learning." It is, by far, it's got everything in there. It's fantastic.

Speaker: To-- if you want to get

into the computer-like oriented side of it, I would highly recommend actually checking out the scikit-learn sort of just like tutorial. Scikit-learn is a Python package that is just dripping with low hanging fruit as far as how to do very basic machine learning.

Speaker: Right, between the two of those, you should be set for a while.

Speaker: Great. Great presentation today. Thank you very much. As Ellie, Carson-- or, Ellie-- April and Carson mentioned throughout the webcast, we are going to have a part two on June 20th. It's going to be Digital Footprints. Privacy and security will be the second part. So, we'll be able to touch on some of those many questions that came in about those aspects. Upon exiting today's even, we ask that you do fill out our survey. The survey tab is available in the chat window now. So, thanks again for everybody for attending today. Thanks again for a wonderful presentation. Have a great day, everyone.

Speaker: Thank you.

Digital Footprints: What Can be Learned from the Traces We Leave on Social Networks

Digital Footprints: What Can be Learned from the Traces We Leave on Social Networks

April Galyardt

Carson Sestili

Software Engineering Institute
Carnegie Mellon University
Pittsburgh, PA 15213

Carnegie Mellon University
Software Engineering Institute



[Distribution statement A] Approved for public release and unlimited distribution.