EPISODE 689

[INTRODUCTION]

[00:00:00] JM: When Aran Khanna was a college student, he accepted an internship to work at Facebook. Even before his internship started, he started playing around with Facebook's APIs and applications. Aran built a Chrome extension called Marauder's Map which used Facebook messenger's web APIs to track where people live, what their schedule was and other highly sensitive information. These were not public features of Messenger, but Aran was able to reverse engineer the APIs.

As a result of making Marauder's Map, Aran invitation to work at Facebook was retracted. Aran remained curious about the norms of publicly available social network data and the second order datasets that could be produced on top of those publicly available datasets. Out of his curiosity, – created a tool called Money Trail. Money Trail was another experiment in using publicly available social network data. Money Trail was built on the public Venmo data. If you've ever used Venmo, it's a payments app where much of the payment information is default publicly available, and he was able to use that publicly available data to build a graph of how users were paying each other.

Aran showed for a second time that data that seems innocent to share can be repurposed to identify, classify, and incriminate users. Developers of these online applications face tradeoffs between privacy, convenience and security. By interacting with these applications, we generate data that suggests how we think, what we like to do and who we are affiliating with. Google and Facebook probably understand you better than you'll understand yourself.

Aran Khanna was previously on the show to talk about machine learning at the edge. That was when he worked at Amazon Web Services. and Aran now works as a digital privacy researcher. He's working at a startup and his background in machine learning makes him well-equipped to talk through the subtleties of modern digital privacy.

In this show, Aran returns to talk through these finer points of privacy, data and artificial intelligence, and it was a pleasure to talk to him about the subject, because many of the conversations you hear in news outlets or in politics, they just are not realistic, because they're not rooted in the realities of how data is shared, of how data makes its way around, and of the advantages and disadvantages of building machine learning models, which is not the fault of politicians or media outlets because they're not trained in computer science, but I would like to have more of these kinds of conversations where we can truly dig into the realities of the issues, because we have an audience and a guest lineup that has an understanding of the science and the realities of these things.

Before we get started, I want to mention that we are hiring for a number of roles. You can find these softwareengineeringdaily.com/jobs. We're looking for writers. We're looking for a podcaster, and the podcaster job is a fulltime job. It pays well. It pays an engineer salary, but we have extremely high standards for the podcaster that we are looking for. The podcaster would have to have savvy in both engineering and in broadcasting/charisma/public speaking. But with that said, we would love to see applications for those kinds of roles, those editorial kinds of roles as well as we have some engineering roles that are posted. So we'd love to get your applications.

[SPONSOR MESSAGE]

[00:03:37] JM: DigitalOcean is a reliable, easy to use cloud provider. I've used DigitalOcean for years whenever I want to get an application off the ground quickly, and I've always loved the focus on user experience, the great documentation and the simple user interface. More and more people are finding out about DigitalOcean and realizing that DigitalOcean is perfect for their application workloads.

This year, DigitalOcean is making that even easier with new node types. A $15 flexible droplet that can mix and match different configurations of CPU and RAM to get the perfect amount of resources for your application. There are also CPU optimized droplets, perfect for highly active frontend servers or CICD workloads, and running on the cloud can get expensive, which is why DigitalOcean makes it easy to choose the right size instance. The prices on standard instances have gone down too. You can check out all their new deals by going to do.co/sedaily, and as a bonus to our listeners, you will get $100 in credit to use over 60 days. That's a lot of money to experiment with. You can make a hundred dollars go pretty far on DigitalOcean. You can use the credit for hosting, or infrastructure, and that includes load balancers, object storage. DigitalOcean Spaces is a great new product that provides object storage, of course, computation.

Get your free $100 credit at do.co/sedaily, and thanks to DigitalOcean for being a sponsor. The cofounder of DigitalOcean, Moisey Uretsky, was one of the first people I interviewed, and his interview was really inspirational for me. So I've always thought of DigitalOcean as a pretty inspirational company. So thank you, DigitalOcean.

[INTERVIEW]

[00:05:44] JM: Aran Khanna, you are a digital privacy researcher and you're working on a company and perhaps several other projects. Welcome back to Software Engineering Daily.

[00:05:54] AK: Thanks for having me, Jeffrey.

[00:05:56] JM: The last time we talked about edge machine learning, this was when you were at AWS, and that was a popular episode. This time we're going to talk about privacy and we'll eventually get into some overlap of machine learning and privacy. Starting with some background, three years ago you were going to be an intern at Facebook, and before that internship, you actually created a Chrome extension, and this Chrome extension that you happen to create got your internship offer retracted. What did your Chrome extension do that got your Facebook internship offer retracted from you?

[00:06:36] AK: Well, essentially, there was a feature in messenger, the Facebook messenger application, where by default every time someone would send message from their cellphone, the location of where that message was sent was automatically attached to that method. This was a problem, because people use messenger like they would text. So every second of every day they'd be in group chats and private chats sending messages, and every one of these messages would have a location attacked.

Now this was something that you could obviously turn off, but not many people did, and an avid messenger user myself, and I have an intern who really believed in the ethos of doing what was right for the customer and trying to get a baseline level of transparency about what the product was actually doing on the backend and what information is being collected. I thought it was curious that not many people turned off this feature firstly, because I could just go through my messenger list and see all location of my friends were sending. But I also sort of understood that because it was a default, people might not have thought to look into it or turn it off or even knew it was there.

Really, all the extension did was you would download it, and when we would open up your messenger, it would just take all of those locations that you'd sent out to your friends or your friend sent in to you and just plotted them on a map. So it was just adding a baseline level of transparency as to what the information the APIs on the backend were exposing.

[00:08:08] JM: This was a preface to some of the work and some of the questions that you are exploring right now. Why were you compelled to make this application that was kind of exposing some of the hidden aspects of, in this case, Facebook messenger? But more broadly, and application that people kind of misunderstood the privacy implications of using that application?

[00:08:37] AK: Yeah. Just to lead off, I think these technology platform, and I worked in many of the companies that developed them, have the ability to do and provide incredible value for the world. That said, the way they were being developed and the way a lot of them still are being developed is really like the Wild West. There is no sort of second thoughts about how when features and decisions about what information is collected, how that information is exposed and how that process is communicated to users. When that comes together, there's not a lot of deep thought about what the impacted scale is going to be, and that's very problematic because of the fact that these platforms reach incredible scale. A .1% scenario that seems incredibly unlikely in the mind of a programmer, such as a stalker using the stream of locations that one of their victims is unwittingly exposing over messenger to find them and track them down, those sorts of things will actually happen reasonably frequently at any large scale.

I think that this project was one in a number of projects including – I looked at Venmo as well, which had these default settings that were kind of poorly set and the information being exposed publicly was poorly communicated. I think this was a growing problem that really has the potential to hurt users and was actually actively hurting users and potentially endangering them and definitely exposing things about themselves that they didn't want to expose.

It really wasn't in the best interest of the users, and I think that that’s something that is really critical about all these projects. It was trying to add a basic level of transparency, at what data was being collected and exposed that should have been in the project – In the product rather in the first place.

[00:10:28] JM: In the Facebook messenger case – So your Chrome extension essentially made a map of the locations of your friends. Is that what it was?

[00:10:39] AK: Correct. So all of those locations that were being attached by default to the messenger messages were actually essentially scraped off the page by my Chrome extension and using Mapbox just plotted out there right the browser for you to see. So you could really just understand the pattern of movement of all of your friends, and it's actually – Most people have very regular schedules. It’s actually really easy with a month of messenger data even with a frequency of once or twice a day to understand, “Hey, this is where all of my friends are at any given moment with really high probability.”

[00:11:15] JM: Back then, you didn't have to be messaging them all the time. You just had to be connected to them?

[00:11:20] AK: Well, no. You had to message them, but it was all the time. The thing is that it was every message that got sent. At any time in history, if this default was enabled, would have the location attacked. The key thing here is that most people have really regular schedule. Most of my friends at school took a set of classes that repeated every week. So with one month, or even a little bit more of infrequent messaging history, I can actually build up that pattern, and if you just look at the map, you can see how much regularity there is and how predictive that set messages of that person’s kind of trajectory throughout any given day.

[00:12:00] JM: So for a given person, you had to message them throughout the day over the course of a month perhaps to get a reasonable sample size of where they were at different times throughout the day?

[00:12:10] AK: Yeah, maybe about throughout any one day, right? If you message me one day at 8 AM, one day at 9 PM, one day at 7 PM, I'm slowly building up this picture. What's really interesting is that a lot of these platforms have things like this. Even like Venlo, every single payment, you're like, "Oh, even if I am attaching my location, even if this payment is public, what's the worst that can happen?" I don't care about this one payment. But it really is about kind of a straw by straw, right? It builds up this rich picture of your activity over time every little bit of information you put out there, and people don't realize how much can be inferred from that.

[00:12:51] JM: Did you feel like there was some risk to making this tool, because you were literally making a tool that exposed a privacy risk of a Facebook application when you were about to be an intern at Facebook. I mean, there must've been some like kind of trolling the company you're about to work at. Did you feel that it all?

[00:13:14] AK: Well, to be honest, I think the ethos that Facebook was trying to convey to its interns was do what's right for the customer and move fast and break thing. Code wins arguments. These are all the things that I heard.

So in a culture like that, wouldn't it be good to go forward and say, "Hey, this is a feature that potentially is hurting users, and most definitely users don't have a good understanding of. If I can write some code to actually fix this problem, wouldn't that be a net positive in the culture that they have, basically, from day one, put forward for their employees and programmers to follow? If code wins arguments, wasn't this the right way to say, "Hey, this feature maybe should be changed."

[00:13:58] JM: Well, it would've probably been more diplomatic to email the product manager of messenger, it'd be like, "Hey, I built this Chrome extension and here's a private GiHub repo to it," and this is something that people could build. Doesn't that seem like a bad idea?

[00:14:13] AK: Yeah. I mean, it really was something that I think was known in the public as well. It was written about for years and years and years, right? So it wasn't like this was a bug that I was exposing that, “I can give you articles from 2012 where this feature first launched, and there are a number of publications that this is creepy and nothing happened.”

So if all I had to do was write something that showed instead of told to cause people to actually stand up and take notice, I don't think that that's something that requires special disclosure. In fact, all that Chrome extension was doing was something that you can do with a piece of pencil and a paper. You could go through your messenger feed, click on everything, the location, and draw a little dot for everything single location that was actually sent to you and exposed via the browser.

[00:14:58] JM: I mean, I don’t disagree that – I don't blame you for making this tool, but looking back now where you’re like, “I was really naïve to think that there was no risk of kind of losing my internship doing this.”

[00:15:12] AK: Yeah. I mean, I think it was I was still a student. There wasn’t a bit of naiveté. I did assume that companies, especially a company like Facebook, would it act in the best interest of their users and they would find a tool like this more useful than harmful? Clearly, that wasn't the case.

I think you know, especially with Facebook, the track record since then has not been great either in terms of being proactive about user privacy. Yes, there was some naïveté there. I didn't fully realized that companies are not there – In mean, in many cases are there sort of the user of the companies like Facebook were really not. The fact that they would rather obfuscate this feature and try to slide it under the rug rather than have a frank and conversation about, until it's fixed, how users can protect themselves, or at least understand the consequences of using it are? That was something that was eye-opening to me and I didn't realize before I had gotten into this whole mess.

[00:16:14] JM: I’m obviously a shill for the technology industry, but I tend to be sympathetic to Facebook when people take critical eye to Facebook, because if I look at just the problems they have today, like fake information spreading on WhatsApp and leading to the murders of people. If you're in Facebook’s shoes, that's just not – That’s not an easy problem to solve. There's no ethical handbook or guide of commandments or standards that you can look at and say, “We know what to do.” It's kind of damned if you do, damned if you don't. You're damned if you suppress the information and prevent rumors from spreading really quickly, and you're damned if you just let the information spread and people get killed.

I think this is – That's an extreme example that really illustrates that Facebook has been in this morass of really difficult problems for a really long time, partly because it's such a novel application and they’re exploring the cutting-edge. Do you think that – Is there some institutional problem with the way that Facebook has approached feature development or the way that the company ethos has been structured if you were in charge of Facebook, you would've made changes to earlier?

[00:17:30] AK: Well, I think that Facebook developed from a place where hyper growth was top of mind all the time, and move fast and break thing was, and still in lot of ways, is the ethos. I mean, now their motto from friends that I know who work there is move fast with stable infra. But still, the moving fast on feature development, while it’s good and leads a lot of the crazy growth, especially in the developing world that we’ve seen with Facebook, it also – It’s something that can potentially go awry in the cases where norms and kind of an understanding of local culture is tossed aside and it leads to things like people using Facebook to coordinate and hunt down people. It leads to a lot of negative things that if there were some local partner, if there was some amount of consultation with folks on the ground and obviously more staffing of the platform, instead of just pure automation and scaling, I think there would be at least some degree of kind of plausible deniability for Facebook. But in a world where they're happy, they just sit back on the technology they’ve built and push it out to the world and move as fast as they can with feature development without really going in and taking a hard look at how the features they're pushing out are actually impacting people. How people are interacting with them at scale in every locality they’re in. If they don't do that, then it's really a culture that prioritizes growth above everything else, which is not the way that I think of social platforms in the year 2018 with the scale that Facebook how it should be run.

[SPONSOR MESSAGE]

[00:19:24] JM: Accenture is hiring software engineers and architects skilled in modern cloud native tech. If you're looking for a job, check out open opportunities at accenture.com/cloud-native-careers. That's accenture.com/cloud-native-careers.

Working with over 90% of the Fortune 100 companies, Accenture is creating innovative, cutting-edge applications for the cloud, and they are the number one integrator for Amazon Web Services, Microsoft Azure, Google Cloud Platform and more. Accenture innovators come from diverse backgrounds and cultures and they work together to solve client's most challenging problems.

Accenture is committed to developing talent, empowering you with leading edge technology and providing exceptional support to shape your future and work with a global collective that's shaping the future of technology.

Accenture's technology academy, established with MIT, is just one example of how they will equip you with the latest tech skills. That's why they've been recognized on Fortune 100's best companies to work for list for 10 consecutive years.

Grow your career while continuously learning, creating and applying new cloud solutions now. Apply for a job today by going to softwareengineeringdaily.com/accenture. That's softwareengineeringdaily.com/accenture.

[INTERVIEW CONTINUED]

[00:20:56] JM: In 2018, I think user preferences are a little more straightforward. They're still not completely straightforward, but they're a little more well-defined. We have a little bit better understanding of what people want out of these platforms. Just to push back a little bit, Facebook –Contrast Facebook with Amazon. So you've worked at Amazon. I worked at Amazon very briefly. Well, 8 months. So not super briefly, but it's long enough to know that Amazon is more about like deliberation and long-term thinking, but there is also a speed and an agility and a rapid expansion component to Amazon. But Amazon seems to have a few more checks culturally on, "Are we doing something risky here?" perhaps, but even Amazon, in the pursuit of growth, they have challenges with their warehousing systems and like outsourcing delivery to third parties that may not have the most scrupulous practices in terms of managing the workers there. That's a challenge that's a little bit more well-defined. If you're Amazon, you know you want good working conditions for your workers. You know that there is a bottom line or there is a floor on the working conditions of the workers that you don't want to cross, but it's just – The nature of the expansion that the company undergoes makes it really hard to really enforce those guidelines.

At Facebook, you just have more challenges that are of a more distinctive nature, of the new nature. You look at the stuff that they are dealing with today and the stuff that YouTube is dealing with today and this human content labeling problem, like scaling up the data labeling in order to police the fake information on YouTube or Facebook. That's uncharted waters, and it feels like they were in uncharted waters before. Yes, there is this narrative of like growth over everything in the past. To some degree, that was true. But doesn't the novelty of the challenges that these companies are encountering, doesn’t that to some degree absolve them of blame?

[00:22:55] AK: I don't think it entirely absolved them of blame. I think it reasonable that these challenges do not get fixed overnight. I do want to go back and say that I think Amazon, and I think we both have worked there, it does have an incredibly fast growth mentality, but the key difference between them and Facebook, I think, is largely the fact that they view customer trust as one of the ultimate things that the product that you put out is trying to gain. The idea of customer centricity and alignment between who your customer is and who your platform is serving is much clear there.

In the case of Facebook, the clarity is not really there, and there is – And I have a lot of friends at Facebook especially back in 2015, maybe it’s different now, but there was not a lot of focus on customer trust. I think those sorts of cultural institutions within a company guide a lot of the downstream product development. While the challenges might be unique and there might be a lot of unknown unknowns, and unknown unknowns that are really difficult to tackle because of the novelty of what these social platforms are doing in the scale of content that they have to believe. It’s still is something that I think at a cultural level wasn't there when the company started.

This idea of customer centricity and earning customer trust of the users of the platform in this case, not of the advertisers that are pushing content to the users, that's not something that is really top of mind in any room when the decisions for the product are being made in these cases. Now, maybe it's a little bit different in Google case. I'm not sure, even though the model is very similar, but the sense I get is that there is a cultural gap as well. While the problems are hard, I don't think it absolved them of making a fair effort to try and solve these problems and noticing when things are spiraling out of control and being upfront about sort of fixing it. If it means turning off the platform for people in a certain part of the world, maybe that's a drastic response, but maybe it's worthwhile, right? Starbucks closes all its stores after there was an incident in one of them in Philadelphia. These sorts of things actually do, in the end, win over the trust of consumers these drastic measures. And I really haven't seen Facebook take any of them. So it does also kind of bely this sense that they don't have customer centricity and they're not striving every day to earn customer trust.

[00:25:29] JM: Let's focus in on the issue of privacy. Many consumers say that they care about privacy. Some consumers take actions that do reflect caring about privacy a lot, but other consumers may say that they care about privacy, but their actions reflect something different. What kinds of privacy do people actually care about?

[00:25:52] AK: I mean, I think that it's quite interesting, because a lot of my thinking about this has come from behavioral psychology literature. I think people really deep down do care about privacy, just like people deep down care about saving for retirement, or eating healthy. But because of the fact that the entire interaction pattern with the privacy settings in the platform is essentially controlled by the platform and historically, maybe not recently, but deftly historically, has been something that they have taken liberties to make, incredibly varied and inscrutable, difficult to work with in change. That causes people cannot act on their sort of core beliefs and instinct that they do want to go in as soon as they start using a platform and really set up those privacy knobs and levers. That isn't really top of mind, because the platform that offers those controls is very much incentivized to make sure that people actually share as much as possible and are really giving as much information over to the platform for the platform to redistribute as possible.

I think there is there is very much a misalignment, and I think the most extreme example of this is if you see design dark patterns which not even companies like Facebook, and Venlo, and Twitter, but more like the small indie mobile game and these startup software shops that have even a couple tens of thousands of users, they'll push a lot of design dark patterns, like buttons that are colored differently, or one that's much smaller, settings that are hidden behind five or six different menus that live in their product and cause an incredible divergence in user behavior from what the norm would be if these sorts of settings and options were put out in a sort of plain and simple way of the user to understand.

[00:27:52] JM: Right, and I think the dark patterns that manifest that impact our privacy – I don't know enough about ad tech to say this with complete confidence, but my sense is that the dark patterns, at least today, are not done – The privacy dark patterns, they're not done as much by the larger companies that people often complain about, like Google or Facebook. They're done more by the smaller ad tech companies, the advertising surveillance world out there, and those companies may or may not be pulling from data sources, like Google or Facebook. But it's unfortunate, because I think people sometimes blame the internet giants for leaking private information, because, "Oh, these ads are tracking me around the internet, and it must be Google." It's not necessarily Google. It can be other third-party data sources. You can pull together a scary amount of "surveillance data from you from a bunch of random ad tech companies that you've never heard of without going through Google at all."

[00:28:59] AK: Right, and I think that entire advertising ecosystem is something that has – Especially with the data resellers, right? The fact that your data was captured by one of these guys can be repackaged and sent off to 10 others, right? There's no control over it. While I think maybe GDPR in a way is going to try and attempt to fix that, I think we've yet to see any sort of measurable impact in terms of people actually being able to actually have a handle on their data.

Yes, mostly these small players that are really the worst offenders. While they don't have the scale of the Googles and Facebooks, they are incredibly complicit in this essentially dark data trade that's going on behind the scenes where private information from you and me and from millions of other people is exchanging hands every single day and spreading the more and more sort of digital institutions across the web.

[00:29:59] JM: So the GDPR, to me it seems like it's useful in the sense that like not an actuality, or an implementation, but in spirit. It's kind of useful, because it represents a big monolithic pushback against the surveillance of the Internet. Would you agree with that?

[00:30:17] AK: I agree with that. I think it's a very useful norm, and I think the California privacy protection legislation that's going through is also a really useful norm.

[00:30:27] JM: What is that one? I don’t know anything about that one.

[00:30:29] AK: There's a law that’s coming through California that is meant essentially be the American equivalent of GDPR and potentially could actually have some real privacy wins for Americans as a whole, because anything that goes through in California kind of has to ripple through the entire United States internet ecosystem.

But I think it’s useful than norm, it’s useful regulate the big folks that can be regulated, but in a sense, GDPR and these California privacy protection laws are largely only useful in so far as they’re enforceable. Most of the bad behavior is still going to happen with smaller players. I’m not saying it’s not because it’s not enforceable. In that case, we shouldn't do it. I think it's still incredibly useful and incredibly empowering to the government to allow them to, in a sense, that norms for the privacy practices of the larger companies they can enforce.

That said, I think that it’s still a first step of many, and the final solution in many ways I think will have to be technical, because the problem is deeply technical. But legislation does help and does I think the right kind of starting norms that we can jump off of and start making more and more improvement in this space.

[00:31:44] JM: On the other extreme, what do you think of China's privacy policy? Looking at humanity as a whole, is it useful to have an experiment in the opposite direction where people get normalized to not having privacy?

[00:31:57] AK: That's one of the more terrifying places that I’ve seen creating technology implemented towards the end of that. I think most folks in the west would say are rather dystopian. The fact that companies like Face++ exists in China and they have one of the most accurate facial recognition technology with incredibly high precision using deep learning target in pretty grainy stock CCTV photos most citizens of China. That's terrifying and that technology platform coupled with the fact that most of the citizens are already digitally connected to their banking digitally, etc., give centralized institutions an incredible amount of power over them, and there's a culture there where those centralized institutions are completely green-lit in wielding that power as long as it's serves the ends of the party in whatever way they see fit.

I think as an experiment, it's interesting from a purely academic perspective. I'm curious to see where it goes. From a humanitarian perspective, I think it's dystopian and rather appalling and a direction that I hope in the west, these technologies do not move in that direction at all. I think that there's an incredible amount of benefit that we can get from these technologies across the globe. We have to figure out how we're going to use our tools. I think it's interesting to see how the Chinese government and the Chinese tech companies are using these tools. But it's not necessarily the direction I would ever want to see anywhere else go in.

[00:33:27] JM: My sense is that in America, there are ways in which we might care too much about privacy interest in certain axes. So the example of medical records comes to mind. The fact that we have not been able to orchestrate a mass study of data across large swaths of medical records it. There's so much value there, and yet we have not been able to capture it partly because of this problem of anonymization of that data. Are there ways in which you think we care too much about privacy?

[00:34:04] AK: I think there's a little more nuance to it. In fact, the electronic health records, yes, if you want to distribute them as a company, there is this step of anonymization that you need to do, and I can actually talk a whole bunch more about that. I think that's one of the more interesting problems from a technical perspective, anonymizing and releasing data. But even if you assume that the data scientists or analysts at the end is incredibly trusted and doesn't need any sort of heavyweight anonymization procedure on the data, it's the fact more so that data is kept in a number of different silos, such as different hospital systems, etc., and data sharing between those silos is really the bottleneck to any large-scale studies in a lot of ways, not the anonymization.

You can even see this in fighting fraud online. It’s actually quite interesting that most financial companies, most credit card companies, don't actually share the data of the fraudulent transactions they have with each other, which would actually make everyone's models better and reduce fraud across all of these platforms. I think there're more sort of bureaucratic obstacles more so than privacy norms that are causing a lot of these sort of shortcomings in the studies that we can do on the data that we’re collecting.

[00:35:22] JM: Really? So you can effectively anonymize a set of hospital records? It's just challenging to share them?

[00:35:29] AK: You can effectively anonymize a lot of datasets. In fact, there’s an entire field of study that I'm involved in around it. I don’t know if you’ve ever heard of the term k-anonymization. That's kind of the first of many techniques in actually anonymizing datasets.

[00:35:47] JM: Explain what k-anonymization is.

[00:35:49] AK: K-anonymization is kind of this simple concept that if you have a set of data, say, an Excel table, you have one or two characteristics with an Excel table that you want to protect, say, the disease, the person that that table has, or the salary of that person, and then you have a whole bunch of information about that person that can be gathered from other public records; their date of birth, their ZIP code, their sex, etc.

Now, if you just strip the name away from the table and released it wholesale, what could happen if someone could go to, say, Facebook and say, “Hey, there is a female in this database that's from Arkansas that's born on this day that is – Whatever, another – Something kind of simple that you can get off that public data source and they could actually match it with the Excel table and actually find out even though it's anonymized, they could find out the disease standing of that person.

So what k-anonymity is doing is essentially saying, “Hey, every single record here has to be identical in terms of the quasi-identifier in terms of that publicly accessible subset of information. It has to be identical to k other records.” It’s this idea of hiding in a crowd in that release dataset.

[00:37:09] JM: Why is k-anonymization hard to implement, or are you saying that it's not hard to implement? Is it a solved problem?

[00:37:14] AK: I'm saying it's hard to do it effectively, and it's hard in many ways, because to actually implement it, one of the most common ways is through micro-aggregations. So instead of saying, "Female age 30 from Arkansas," you would say, "Female age 20 to 40 from Arkansas," and it actually messes with a lot of the statistics of the dataset. So there is a tradeoff when you k-anonymize. Actually, k-anonymity isn't sufficient in a lot of cases. You might want to, let's say, the variable that sends a data to something like salary, right? Or let's say it even is disease category. If the K people in your k-anonymity group all have cancer, then you know that if you can find any one of those people in a matching dataset, even though you're hiding in a crowd, that the sensitive variable is still being exposed, right?

So k-anonymity is not sufficient to guarantee privacy and it has some pretty steep tradeoffs. There are these things like l-diversity that help you sort of refine this notion of k-anonymity for the case where there is a common shared sensitive attribute in your k-anonymous group. If it's, let's say, salary and it's continuous, there is another definition, t-closeness to help you sort of further anonymize and protect against kind of statistical exposure of sensitive factors through analyzing the anonymize data.

There're a number of techniques and a pretty rich field of study around taking datasets, and you're your threat model is the data scientist that I'm going to release this to is potentially going to try and de-anonymize these people. With that threat model, these sort of things have found sort of use in the real world. With carefully [inaudible 00:38:57] around with the anonymization procedure, you can protect the statistics that you really want to expose from that dataset, to expose the value of that dataset.

Now, there's actually other kinds of privacy metrics, like differential privacy that work under a different threat model where the output of your modeling process is the thing that someone wants to attack, and that's I think where machine learning and these sorts of privacy procedures really meet head-on.

[00:39:23] JM: Explore that in more detail. What are you talking about with the overlap between machine learning and this field of anonymization?

[00:39:31] AK: It’s really coming from sort of a different threat model, right? The threat model in the k-anonymity case is we are a government, our hospital system, and we want to be transparent, but we want to perfect the anonymity of our patients, or our citizens. So we’re going anonymize this dataset and put it out there for anyone to see.

In the case of machine learning models, it’s usually the case that we have all these data as a hospital system or a government and there they trust the data scientists that we’re going to work with. That is going to build a model from the data, and then we’re going to expose that model out in the wild. What has been shown especially recently with things like model and version attack is the fact that when you take a machine learning model, put a whole bunch of data through it and push it out into the wild, within the weight of that model, a lot of that data is actually stored, and some of that data can actually private data.

Especially with these inscrutable deep learning model, which with their millions and millions and millions of parameters, can actually store a large amount of data, it's possible to do things like model extraction attacks. From an image network that was trained on patient data and scans, in some cases, you can actually invert that mode, and if you have full access to the model that’s released, you can actually do kind of a very similar procedure that was used to train the model to actually reconstruct one of the input images that was used or one of the input data points that was used to frame the model.

So these sort of things are now becoming more reasonable threats, and under that threat model, how do you ensure privacy? How do you ensure those weights don't hold private information? I think that the best way that we’ve really found to do that has been through this idea of differential privacy, which at a very high level, is saying, “Hey, the model is a function of the data.” right? So I have my data. I put it through some function and the model comes out the other end.

Now, if I change any one data point in my input dataset, the probability of my output model changing is less than some epsilon E, which is kind of the formal definition of differential privacy.

All that means is I'm able to not discern whether or not a given individual was in the training set for a model by just looking at the model. If you think about it for a bit, what that really means is that model isn't really exposing any private data on any individual. It's really just at its core, exposing aggregated data, or statistics that are aggregated from multiple people.

[SPONSOR MESSAGE]

[00:42:21] JM: Citus Data can scale your PostgreS database horizontally. For many of you, your PostgreS database is the heart of your application. You chose PostgreS because you trust it. After all, PostgreS is battle tested, trustworthy database software, but are you spending more and more time dealing with scalability issues? Citus distributes your data and your queries across multiple nodes. Are your queries getting slow? Citus can parallelize your SQL queries across multiple nodes dramatically speeding them up and giving you much lower latency.

Are you worried about hitting the limits of single node PostgreS and not being able to grow your app or having to spend your time on database infrastructure instead of creating new features for you application? Available as open source as a database as a service and as enterprise software, Citus makes it simple to shard PostgreS. Go to citusdata.com/sedaily to learn more about how Citus transforms PostgreS into a distributed database. That's citusdata.com/sedaily, citusdata.com/sedaily.

Get back the time that you're spending on database operations. Companies like Algolia, Prosperworks and Cisco are all using Citus so they no longer have to worry about scaling their database. Try it yourself at citusdata.com/sedaily. That's citusdata.com/sedaily. Thank you to Citus Data for being a sponsor of Software Engineering Daily.

[INTERVIEW CONTINUED]

[00:44:07] JM: The last time we spoke, you were working on edge machine learning at Amazon, and since then you've left and now you're working on these topics related to privacy and anonymization. Take me through that transition. Why did you end up studying these areas that you're working on right now?

[00:44:24] AK: I think it’s actually quite interesting, because the sort of models that are deployed with the edge people learning techniques often can hold private data, and this idea of figuring out how to anonymously train models and anonymously aggregate data from devices that models are deployed on. There’s a recent paper by Google on federated learning that each phone in a network can actually kind of privatize whatever data they create and send it as a gradient to a machine learning model the update. These sorts of techniques I think kind of meet head on in the edge deep learning world, especially because if you're pushing your model to someone else's hardware, they have full ability to inspect every single the weight, and if it's not privately trained, this threat vector actually becomes real, right? They can actually go in and steal all of the data or de-anonymize some of the data at least that was used to create that model they're pushing down to your device.

So I think there's a lot of interesting problems that sort of arise in this space, and the academic journey was very much driven by the fact that as you understand sort of the abilities of these sort of model compression techniques and how ubiquitous these sorts of deep learning models are going to become, the next thing, if you're a privacy-minded person that will come to the top of your mind is, “All right. Well, what does this mean for all the data that went into these models? Especially models that are going to be used in the healthcare space for health tech, etc.”

[00:45:54] JM: So what are you working on today more specifically?

[00:45:56] AK: More specifically, today, other than the startup, I'm working on largely private training procedures. So figuring out how to take deep learning models and train them in a private way by largely adding noise departs in the training procedure in ways that give you these sorts of mathematical guarantees, like differential privacy.

[00:46:16] JM: Adding noise. Explain what you mean.

[00:46:18] AK: Essentially, what that means is you are taking some random numbers and sprinkling them in throughout the training procedure. If you do that in a very directed way with the right sorts of distributions that you're taking these random values from when you're adding them into the training procedure, you can actually yield some pretty sort of provable and stable privacy guarantees about the training procedure and, hence, about that end model that comes out the other side. These guarantees and training procedures actually can be merged with a lot of existing training procedures, especially things that are currently used to train network that go down the edge, like half precision training and training while pruning away weights that reduce the size of the model. All these sorts of techniques are in a big toolbox and they can all be used together. I think that the ability to at least discover and put out what these tools are for folks who are going to launch in to trying to implement these networks into product and sort of actually touching real customer and human data with these sorts of networks, it's really important. It's really important to have that toolbox out there and to let people know what's possible and what sort of protections in terms of their training procedures and in terms of their modeling procedures are actually available to them.

[00:47:39] JM: So the problem you're defining is my model. I might train my model in a bunch of information, and I want that information to help inform the model, but I don't want to explicitly have that information be available to somebody who’s querying the model. I don't want somebody to be able to query the model and know what is Jeff's history of cancer, or some other thing that might impact my health insurance, for example. But I do want to be able to query the model and know a population's probability of cancer.

You’re saying that by potentially adding noise at certain phases in the model training, you can have stronger guarantees about the query ability or the anonymization of the datasets that have been used to train a model.

[00:48:34] AK: Correct. Mathematically speaking, what it means is you can't construct these sorts of attacks against that model once it’s put behind an API and served, and I think it really goes back to and dovetails well even with the work that I did with Facebook and Venlo where information was being exposed via these APIs and people didn't really take that simple step to go and check the API’s, see what the information coming over the wire was and what the consequences of aggregating that and kind of constructing a quasi-attack off of that aggregation would look like. I think in a very similar way, a lot of customer data then, in many ways is private, is now being put into machine learning models and put behind these APIs and people really aren’t taking a second look at, “Hey, if I can construct the number of queries against these APIs, can I actually extract some of that customer data?” I think getting ahead of that problem before it becomes a problem is something that in my mind is incredibly important and something that our community, the machine learning research community, has a real obligation to do sooner rather than later. I think they really have to start and we really have to start putting these tools in the toolbox making them really accessible and easy to use for – And the programmers are going to be using these technologies and exposing them on a real customer data behind real API endpoints to the whole world.

[00:49:52] JM: You are working day-today on startup also? I think that startup is related to analyzing the networks of crypto asset movement. Is that accurate?

[00:50:04] AK: Well, it's actually largely around model serving. That was one sort of modeling vertical that we got into right away, because there was a boom there and people wanted a lot of model developed on those systems. But it's largely around just model serving. So that component of taking that model that you've developed, putting it behind an API and serving it on set of input data. There's a lot of really interesting problems that arise just in that process largely around checking if the distribution of the data is the same or different coming in second by second, especially in spaces like digital asset markets. You have high degrees of non-stationary. So you have to kind of understand how things are shifting.

Retrainings are really big problems. So that sort of loop is part of the deployment in many cases and that's really not easy to get right, because you have the potential to really degrade the accuracy of your model in a lot of ways. Just generally, with deep learning models, optimizing for SLAs, like cost and throughput, is really difficult, because there are so many knobs and levers that come out of. When you're serving at scale, when you have hundreds and hundreds and hundreds and hundreds of people hitting this endpoint, this is something that could be a real problem.

Now, our hope is to eventually start to serve models across the spectrum and implement some of these sorts of privacy measures in retraining procedures and guaranteeing SLAs like differential privacy balance, etc., for the model being served.

[00:51:38] JM: I've talked to a few people who are working on problems in the machine learning model release process, all those challenges you're describing. I know somebody who worked on SageMaker, Algorithmia. Algorithmis is a sponsor of the show actually, and we did a show with them a while back. There's also there – There was Y Hat and there's also – Google has there's. There's Pipeline AI, I think. Why are there so many of these different people who are tackling this problem, and is there something distinctive that the company you're working with is doing that they saw as an opportunity in the market?

[00:52:19] AK: Yeah. I mean, largely, what our process is, is we go into existing companies with existing models and we essentially will use very recent sorts of techniques, like knowledge installation, to take those models and put them into like PyTorch and these sorts of more standardized systems that we have a whole bunch of infrastructure built out around. So it's a little more high touched, and I think a lot of the other folks out there, and we actually go in and instead of being a platform to serve the model, we actually touch the model and use our tools and our building blocks and construct the process specifically around that model and around that domain.

I think there's so much interest in the serving solutions, because each domain really has its own serving problems. Serving models for a medical application is very different for serving models for something that's trading assets, versus serving models for add a click through rate prediction. These are all very different challenges. While some of the infrastructure is common, most of the infrastructure has to be custom-built. So there's a huge kind of Greenfield in this space, and I think that's why so many folks are entering it with so many different sorts of offerings kind of across the spectrum.

[00:53:34] JM: Are you exploring this space looking for a product to build?

[00:53:38] AK: I think it's largely right now driven by the demand of folks that we're in contact with that many of the people in the company have relationships with and have done business with before. So there is some sort of latent demand from engineering orgs and data science orgs to say, "Hey, I have this model and no one's maintaining it. My data scientist doesn't want to work on maintaining this old model that this other guy schlep off and left here and is running something critical like our click through rate serving. Can you guys come in and maintain it?" That's really sort of the space that we're working no right now and what we're trying to solve for.

[00:54:18] JM: To wrap just to bring things full-circle, we've talked about a lot of different areas of privacy. I know you have been looking at cryptocurrencies and blockchain stuff for a while. I was looking at your website and you have – I think you have a post as far back as maybe 2013, or you talked about looking at crypto currencies in 2013. Where do you think we're at in terms of the evolution of applications being built around blockchains and what are the applications you think that will be built eventually that will impact our privacy?

[00:54:55] AK: I think that there is an incredible opportunity with the technology to build systems that kind of have a sort of provability of ownership for data built into them in a way that the World Wide Web protocols do not right now. So the way that our browser is built, the way that we access information does not have ownership of data built in as sort of a first priority kind of top of mind technical thing in the system. I think kind of at a high level, I see a big opportunity there.

More granularly, I think I'm really interested in the research side of this, especially the mechanism design side. There's still so much work to be done I think in terms of understanding – It's the intersection of two very, very deep and interesting fields of distributed systems and game theory. I think that the next generation of systems we see are going to be largely moving away from this sort of old-school proof of work that really limits scalability and toward more robust consensus mechanisms.

I think that that's really the evolution that's going to allow the platform to actually take on full amount of scale. So I'm still waiting for that to happen, but I think in the near term, there are a number of really interesting projects. I think I have a sense that we all do, in a sense, trust our government. I'm not a crypto anarchist, and I think that the breadth of applications that people are seeing out there might not actually be as broad, because we already have pretty good solutions for trust in most of these cases. But I think problems that you find on the internet, like DNS, and even data privacy, I think in a way, those are things that are unregulatable by government and a way that cryptosystems and protocols can come in and provide some sort of automated regulation.

Beyond that, I think just generally privacy preserving protocol that are, in a sense, very antigovernment, like the monetary use case, like Bitcoin, or like Orchid Coin, which is essentially taking [inaudible 00:56:51] and adding incentives to it to increase the bandwidth of the network or incentivizing torrent networks through tokenization. I think those are the sorts of systems where there is the most opportunity for crypto to really find a strong foothold.

[00:57:06] JM: Not to open a can of worms, but you mentioned that you think that the beyond proof of work systems are going to allow for the scalability, or I think that's what I heard.

[00:57:18] AK: Yes, very much so.

[00:57:19] JM: That's what you think. Okay. Well, there's another contingent of people who believe that proof of work is a really good backbone, and that second layer solutions are kind of the path to scalability.

[00:57:33] AK: Yes. Lightning and those sorts of things I think are stopgap solutions, to be honest. I think this might be my proclivity being largely from the academic side. I think from the ground up, we do need to design sort of the mechanism first for a lot of these blockchains. In the case of something like Casper, that sort of mechanism design is a very hairy probably. We don’t even know how to debug things like this, and at scale, what is this blockchain sort of implementation on a proof of state going to look like? What are the attack vectors?

I think there are very thorny problems there, but the upside is that in the sort of kind of terminal state, that sort of network is going to be more robust and infinitely more scalable than something that at its core is built on proof of work.

[00:58:19] JM: What makes you say that?

[00:58:20] AK: I think largely, the fact that you won’t susceptibility to attacks from individual institutions that can produce hardware that's burned in at the chip level, like ASICs to attack the chain, allows you to, in a way, provide more security guarantees and more sort of long live in these guarantees around the system. So we’ve seen that Bitmain has put out these – Not to open these can of worms, but like Bitmail has put out these ASICs that have, in a sense, all of these hardware in aggregate, even though it’s not all owned by Bitmain, has over 50% of the Bitcoin hash power and they just came out with Ethereum ASICs.

So already have seen these sorts of attacks that I think are limiting a lot of the scalability by virtue of a few large centralized institutions with these powerful technology tools getting control of a large amount of the currency and being able to manipulate the market and manipulate the price so gas prices go up, or things that were previously cheap become expensive, because these token owners who have artificially inflated the price are essentially extracting rent from the network by virtue of the fact that it's built on this proof of stake model, or proof of work model rather than a proof of stake where a priority you can actually set the tolerances on how much power any one person can have in the network and set the incentives on holding currency, which actually have a little bit of a cost and liquidity, but is able to provide an end market mechanism for managing the actual supply of currency in the market.

[00:59:52] JM: Aran, we've touched on a lot of different things. It's been really good talking to you. I'm sure will do it again in the future.

[00:59:58] AK: Yeah. Always great chatting. Thanks for having me on.

[END OF INTERVIEW]

[01:00:03] JM: Software workflows are different at every company. Product development design and engineering teams each see things differently. These different teams need to collaborate with each other, but they also need to be able to be creative and productive on their own terms. Airtable allows software teams to design their own unique workflows. Airtable enables the creativity and engineering at companies like Tesla, Slack, Airbnb and Medium. Airtable is hiring creative engineers who believe in the importance of open-ended platforms that empower human creativity.

The mission of Airtable is to give everyone the power to create their own software workflows, from magazine editors building up their own content planning systems, to product managers building feature roadmaps, to managers managing livestock and inventory. Teams at companies like Condé Nast, Airbnb and WeWork can build their own custom database applications with the ease of using a spreadsheet.

If you haven't used Airtable before, try it out. If you have used it, you will understand why it is so popular. I'm sure you have a workflow that would be easier to manage if it were on Airtable. It's easy to get started with Airtable, but as you get more experienced with it, you will see how flexible and powerful it is.

Check out jobs at Airtable by going to airtable.com/sedaily. Airtable is a uniquely challenging product to build, and they are looking for creative frontend and backend engineers to design systems on first principles, like a real-time sync layer, collaborative undo model, formulas engine, visual revision history and more.

On the outside, you'll build user interfaces that are elegant and highly customizable that encourage exploration and that earn the trust of users through intuitive thoughtful interactions. Learn more about Airtable opportunities at airtable.com/sedaily.

Thanks to Airtable for being a new sponsor of Software Engineering Daily and for building an innovative new product that enables all kinds of industries to be more creative.

[END]