ephemera
theory & politics in organization


ISSN 1473-2866 (Online)
ISSN 2052-1499 (Print)
www.ephemerajournal.org
volume 17(2): 421-430

# The terms of anonymity: An interview with Marit Hansen, German data protection expert

Götz Bachmann, Paula Bialski and Marit Hansen

## abstract

As data gathering technologies are permeating various corners of our lives, a number of stakeholders are attempting to map, track, analyse and define what is happening to our identity, our privacy, or our ways of being social. As notions like privacy, anonymity, data, unlinkability, or pseudonymity are being defined, many of these definitions, while sounding almost the same, shift meaning from discipline to discipline, from context to context, and from one political agenda to the other. In this interview with Marit Hansen, one of the most influential activists for data protection regulation in Germany, and the head of the Independent Centre for Data Protection (ULD) and the Data Protection Commissioner of Schleswig-Holstein, Hansen highlights the way in which her computer science discipline defines its terms and working categories, in a rapidly changing landscape of data gathering technologies. The interview draws heavily from her (co-authored with Andreas Pfitzmann) seminal paper in the computer science field around privacy, anonymity and 'identity management,' titled 'A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management'.

## Introduction

The profound changes in technologies of personal data collection have shifted our terms for understanding anonymity. As data gathering technologies are permeating various corners of our lives, a number of stakeholders are attempting to map, track, analyse and define what is happening to our identity, our privacy, or our ways of being social. These stakeholders include lawmakers and politicians, think tank members and lobbyists, entrepreneurs and marketeers, journalists and activists, legal scholars and lawyers, social scientists and computer scientists. Part

of these processes of defining the unfolding reality of our mass-scale data collection includes outlining the terms and definitions at stake. As notions like privacy, anonymity, data, unlinkability, or pseudonymity are being defined, many of these definitions, while sounding almost the same, shift meaning from discipline to discipline, from context to context, and from one political agenda to the other.

Computer science has very technical definitions for the terms of anonymity – terms that are used to build technical systems – simulating how anonymity in practice works, searching for failures and loopholes in various communication networks being built, and tweaking these networks in order to improve them. Perhaps the most seminal paper in the computer science field around the terminology used for a range of phenomena related to privacy, anonymity and 'identity management' online is 'A terminology for talking about privacy by data minimization: Anonymity, unlinkability, undetectability, unobservability, pseudonymity, and identity management'. The paper was written by the late-computer scientist Andreas Pfitzmann, and Marit Hansen. To put the authors' main argument quite simply: times are changing in the world of anonymity, data privacy and identity management, and nobody knows how to define what's happening. Let's attempt to do so anyhow.

We reached out to Marit Hansen in order to gain insight into the way in which a discipline defines its terms and working categories in a rapidly changing social landscape. Written in 1999, first published in 2000, and rewritten in multiple iterations until 2010, the paper was created during a period, when the way in which anonymity was 'done' online – meaning how communication happened and who could partake in intercepting such communication – was in rapid flux. Throughout our interview with Hansen it became apparent how data protection knowledge is shaped by a nexus of legal and technical knowledge alike, within historical, political and economic contexts, and with various decisions becoming politicised, often explicitly building on the history that came before it. All these elements become enmeshed. Pfitzmann and Hansen's paper tried to 'clean up this mess', while being fully aware that such a task is challenging, if not impossible.

At the moment, Hansen is one of the most influential activists for data protection regulation in Germany, and the head of the Independent Centre for Data Protection (ULD) and the Data Protection Commissioner of Schleswig-Holstein (one of the Federal States of Germany). She is a computer scientist by trade, and her work sits at the interface of law and technology. In her years as a researcher and data protection commissioner, she has pioneered the concept of data protection by means of technology and privacy-by-design, through which the ULD has gained its respected status. In 2007, she was furthermore appointed by the

European Commission as an expert in the 'Privacy & Technology' working group. The interview is based on an evening with Hansen in her office in Kiel. In two hours, we revisited the aforementioned seminal paper (for a short summary see also the appendix), and explored the paper's context as well as ways how the paper terms can be translated to scholars in the social sciences and humanities who are interested in working with these terms.

**Paula Bialski (PB):** I'd like to start by backing up about 20 years, to the moment before you wrote your paper with Andreas Pfitzmann. What inspired you both to write it?

**Marit Hansen (MH)**: I normally don't get asked this question! I have to really think back ... Well, in the year 2000, we had the first workshop on privacy enhancing technologies, called PETs. In fact this was the founding conference of the PET symposia [currently one of the most influential gatherings of researchers working on privacy technologies.]

**PB:** Can I stop you right here, and ask you to explain a little bit more about where the idea of privacy enhancing technologies comes from?

**MH**: You have to see that in the 1980s and 1990s it had become increasingly apparent that if you want to protect privacy, regulating information and communication technologies is not enough. You have to build such concerns and values directly into the technology, for example by developing technologies that minimise the collection of personal data.

**PB:** Interesting! This reminds me of Lawrence Lessig, who famously declared that 'code is law' – privacy becomes a job of technology, so to say!

**MH:** Yes, very much so! And at that time I had been working at the data protection authority for a few years. I started in 1995, and by the year 2000 we knew already how much misunderstanding there could be between lawyers and computer scientists. At that time, even the 'anonymity' or 'anonymisation' definition in the different data protection acts was different. Schleswig-Holstein [the state of Germany where Hansen is based] had a different data protection policy from the federal one, not to even mention the differences between the definitions of the different EU member states.

**PB:** Can you give me an example of what was at stake?

**MH:** At that time, pseudonyms and pseudonymity entered the legal debate and was turned into laws, but often with completely different definitions. That kicked off the identities management debate in computing and legal regulation. Several

people thought: 'isn't identity management – on the basis of different pseudonyms – the solution? Isn't this the future of data protection perhaps? If you can protect your identity yourself?' At least it was necessary to understand better how technical solutions could support the societal challenges of data protection.

**PB:** So what were the aims of the PET workshop in 2000?

**MH:** The workshop on all flavours of privacy enhancing technologies was organised by a colleague named Hannes Federrath, who, at that time, was a visiting scholar at Berkeley University. So I flew to California with several others from the team of Andreas Pfitzmann from Dresden University. There were Europeans and Americans – and both groups even had different ways of understanding how an infrastructure should work. We were talking about 'mixes' – and there were at that time different ways that mixes worked.

**Götz Bachmann (GB):** Before we talk about these differences between US-American and European approaches – and we surely should! – can you please first explain, what, 'mixes' are?

**MH:** Mix technologies were invented by David Chaum, who is often called the grandmaster of privacy technologies. If you want to achieve anonymity in a computer science context, you have several possibilities. In theory, you could have 'no identifiers' at all, which is, of course, not very realistic in a computer science world. Because, as we all know, there are always identifiers. But there are different types. One type is generated by using random data. This identifier doesn't contain information on the subject it is attached to. And then there are non-random identifiers, like a nickname based on my street name. An e-mail address or a phone number are of the latter kind, too, as they contain information on how to reach the user. To attempt to achieve anonymity, you try to use the former. But what's more important is an 'anonymity set', where people cannot single out specific individuals within this set, because the behaviour of this 'set' is the same. The 'mixes' I mentioned then work as a chain to achieve 'anonymity sets'. They take in messages from multiple senders, mix them, and send them back out in random order to the next destination.

**GB:** I think we might do well with another example ...

**MH:** Okay, let's say I wanted to send a letter to Goetz, and if I send it, you can intercept that. The idea of David Chaum was that we should assume within any communication network, that there is a big mighty observer. At that time, nobody could imagine that this observer was real. But Chaum said, that it doesn't matter if this observer is real or not – if you solve the problem (in computer science terms)

for a mighty and powerful observer, then all other observers are also solved as well. So you make the problem bigger. Even larger than you think is realistic. And if you solve that, the rest is solved as well. Now in the meantime we have found out that the mighty observer, or the powerful observer does exist! But this we did not know at that time.

**GB:** If I understand you right, in your paper the 'mighty big observer' is called the 'attacker', correct?

**MH:** Yes, that's correct. And it is true computer scientist always think about 'attackers', or 'adversaries'. But we shouldn't take the term 'attack' as negative, or aggressive. It can also be a passive observation. And therefore we have something called a 'passive attack' and the 'active attack'. So you always have to consider that this is strictly computer science terminology. It's often not well understood in other disciplines like the legal sciences for example. For legal experts, there usually seems to be motivation behind an action. The attacker is trying to destroy something. That's not the same understanding in computer science. Of course, the third party might be trying to gain access, but this could also be for a legitimate, lawful reason, such as when law enforcement needs to accesses something. Attack does not mean that it's forbidden or not, or morally good or bad. It's only about the power of the 'attacker' – and that's called in our language the 'attacker model'. We also discuss the level of power that this attacker has: are they very bright or intelligent? Or do they have very quick computers? Can they draw on their computer forces? Can they also input their own messages? Or can they only observe? These are all part of the 'power of the attacker'. So you can imagine that at that time, nobody thought that what we discussed was in fact a real attacker model. We thought it was too mighty, too powerful.

**PB:** I am starting to get the picture. And I think this is a good moment to return to your point about different US and European approaches to anonymity.

**MH:** Okay – although it's not really US versus Europe. With the TOR network, for example, there is a free possibility to find your route through a network. Let's call this the US American approach. Our approach, the one we were and are employing in the AN.ON network and its successors, was different. It was much more about knowing exactly the 'nodes', i.e. the 'mixes', the message will pass. We called this mix mechanism a 'fixed mix cascade'. In it, it is fully clear where the 'mixes' are, what the order of the 'mixes' are, and what 'the last mix' is. The Americans said: 'Forget about cascades. Cascades have to be coordinated, and this helps the attacker. If somebody decides how to put together the cascades, you become vulnerable'. But the German team at the workshop argued: 'Our method is reliable because there are guarantees of the service, and you also know' – and here is where

the first legal idea came in – 'where the mixes are situated. What is the local law at that specific "mix" location? Should the "mixes" be in the same country? Should they be in different countries? What are the pros and cons?'

You know from the routing protocol on the Internet that it's not necessarily by accident, where each item of information goes. Whoever says 'oh my route is very good, come to me!' may get most of the traffic. Which means that an attacker can also try to be one of those seemingly 'nice' mixes, and by that getting everything. So with TOR, for example, you send something, and it finds its way. 'Its magic! It's good! You don't think about it!' But almost every hop could be accessed or owned by the NSA. You don't know for sure, but it could. With the cascades of AN.ON this could, in theory, happen, too. But with the fixed mix cascade, you know, who is providing each mix. The mix provider has signed a contract – at least in our setting they did. So if you know beforehand, who owns the specific mix, you may visit them, you may think about their reputation, you follow up with them. You would think, that if one link in the chain of the mix is weak, it breaks. But within a mix cascade, the opposite is the case. If only one link is strong, that would be sufficient for anonymity. The attacker does not have the full information and thus cannot decrypt the data, and anonymity is not broken. So we think it is really something where we can do some lobbying for.

With the American perspective, on the other hand, the individual is responsible, and everybody who discloses information is responsible. And it's the 'once-it-is-out-we-can-not-help-you' approach. But this does not work well in a networked world. Who can really defend himself against so many data controllers? So the legal European model, the data protection model, means, we want to *trust* the data controller, but the data controller has to give guarantees and to prove its trustworthiness. And if they are doing something wrong, then we can sue them, or they can be fined, or something like that. So these are two different approaches.

**PB:** And all this comes to the foreground in this workshop in 2000! What happened next?

**MH:** At the workshop we found out that we need the right terms to find out what are the different pros and cons in this matter. At that time, Andreas Pfitzmann was lecturing on this topic, and he had some ideas of how to define these terms, because it was his need to have these categories organised in his head. But these terms were not really what we needed at that time. So we sat around and got into a lot of discussions. Andreas changed almost everything. That workshop in Berkeley wasn't about PowerPoint slides, but really about getting together with colourful paper and pens and transparent projector foils. And I remember that Andreas put the foils on top of each other to make different levels – to see 'now we

are discussing this level, now that level'. And this I thought was very impressive. It was such a nice atmosphere there. The weather was warm. We were sitting on the flat roof of this building and thinking about what he had presented. And through this process, we found out, 'yes, it makes sense, not only to have this debate, or one little facet of the debate, but it's great to really have some basics. To really have the same terminology'.

**PB:** In the following 10 years, the paper was constantly being updated. It almost became a public document, a sort of wiki written by some of the leading experts...

**MH:** Yes, and that is not typical, by the way! It is not typical to have an open paper, which is ready for discussion and amendments from the public. Andreas was a very open guy and said, 'This is so important, we need to get feedback from everybody who wants to give feedback'. These updates made sense for our times and for our discipline. You can look back to older versions and see the progress. We decided it doesn't matter where we publish it. We didn't want to publish it for the sake of publishing. We wanted to publish for the sake of the academic discourse. And at that time you could see several references from different fields, and different translations into all sorts of languages, but it was progressing slowly. If you count what is happening in the field of anonymity in different disciplines, it is very hard to, well, cover everything. Our paper worked, because we said it was not fixed. We wanted to get input from others in our field, because otherwise this term-building would not work. After my co-author Andreas Pfitzmann died, at first I didn't feel like continuing on my own – we always had so many discussions and argued about each word until we were satisfied. This process cannot be done by one person only. Several people asked me to continue and update our work. But one of the things I understood only recently is that I am now in a different position. I am the head of ULD, and this is the supervisory authority in charge of laid down data protection law. But the legal definitions are different from the computer science perspective. Even if we could achieve a connect, changes in the terminology paper could become political. That makes it too complicated.

**PB:** What would you change now? Where does the paper need updating?

**MH:** We only wrote this paper to define communication technology, but we never really addressed database terminology – which is, as it turned out, something different. At some point in time we noticed that and added a definition of the setting. But the discussions on big data and potential anonymity wouldn't fit well in the current structure of the paper.

**GB:** Could you elaborate on this difference?

**MH:** With a communication network, there is always a sender and a message and one or more recipients. Then there are always items of interest. This can be, for example, the message itself, or the relation between a message and sender. With database terminology, it is important to remember that a database contains many entries, many items of interest. This is important. When just analysing a communication system, we assumed that, for example, a third party doesn't look into the content of the message. So if the message contains 'I am Marit', I can encrypt it and do as much anonymisation to this message as I want, and nobody can read into what the content of the message is. So with communication systems, we assumed, that the message is not readable, and that it is encrypted in a way that it cannot be hacked.

But it does not make sense to discuss settings of databases with encrypted data. Why not? Because you cannot work well with encrypted data. So we always have to take into account the accessible information. And as you can imagine, a database often includes personal data. So then what do you do with this personal data? That becomes mostly a legal discussion: When do you anonymise, or throw the item of information away, etc. etc.? But this issue goes beyond singular databases: if there is a large amount of people in a medical database, and this database can be linked to other data sets in another database, then it may be very easy to get to the personal relation by linkage, by linking these two databases together. Databases contain much more information than merely the obvious. This was the case before in the 90s, but in our times of 'big data', this has reached a new dimension.

**PB:** This sounds like an even more pessimistic stance than the starting point of your paper, where you state that full anonymity is not achievable.

**MH:** A perfect world is not achievable, but still, we talk about it, right? Again, I think it is about the attacker-model. If some observers can observe so many things, or so much is digitised, or available in some way, then you can put in as much effort as you want to anonymise something, but it's still not achievable. I guess a person could be anonymous only by not being part of society. Since the last version of the paper in 2010, which came out still before the full impact of the hype of social networks, there are new things we have to consider. So many people have already left so many data traces and discussed so much online. From a computer science perspective, if information is out, it is out. But the legal world has introduced the right to forgetting, and the technical tools for protection are improving. I am optimistic that the level of data protection will increase if we design products and services with fundamental rights in mind.

## Appendix

Pfitzmann and Hansen's text starts with a 'setting', which contains 'senders' sending 'messages' to 'recipients' via a 'communication network', as well as an 'attacker', who aims to infer 'items of interests' (IOIs). Senders and receivers are both 'subjects', which can take the form of a 'human being (...), a legal person, or a computer'. Anonymity can be achieved, if 'the attacker cannot sufficiently identify the subject within a set of subjects'. The latter is called 'the anonymity set'. A system normally aims to provide more than 'individual anonymity' for one specific subject. But as 'global anonymity' for all its subjects is never achievable, the latter is a question of 'strength'. Pfitzmann and Hansen then introduce three further terms: 'unlinkability' refers to a state, where IOIs cannot be linked to each other, whereas 'undetectability' and 'unobservability' describe states where IOIs are hidden. Based on this groundwork, Pfitzmann and Hansen analyse sender-, receiver-, relationship-anonymity. 'Pseudonymity', on the other hand, is a state, where an 'identifier of a subject other than one of the subject's real names' is employed. It enables, for example, the accumulation of reputation. If one holder has different pseudonyms (for example for different contexts), establishing 'sameness' can be a goal, but also an open door to an attacker. Pseudonymity furthermore throws up questions of various forms of links between the pseudonym and its holder. 'Public keys' are one specific and particularly important form of pseudonyms, which enable its holder, and only the holder, to prove his or her holdership by the 'corresponding private key'. The last of the terms introduced is 'identity management'. It describes the 'administration of identity attributes', is thus more a practice than a state, and includes an invitation to increase user agency in a given setting.

Link from 'Pfitzmann and Hansen's text': https://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf.

## the authors

Götz Bachmann is Professor for Digital Cultures at Leuphana University Lüneburg, Germany, and currently a Research Fellow at Stanford University. He is an ethnographer by trade, and the author of a monography on "Kollegialität" (collegiality.) His ongoing research looks at radical engineers in the San Francisco Bay Area, who aim to develop a new digital medium.
Email: bachmann@leuphana.de

Paula Bialski is a postdoctoral researcher at Leuphana University's Digital Cultures Research Lab (DCRL) and currently a Visiting Research Fellow at the Institute for Advanced Study, University of Konstanz. Her past work ethnographically studied

"Couchsurfing.org" and online hitchhiking websites in order to map out digitally-mediated, mobile interaction. Her current topics include digital infrastructures, programmer worlds, anonymity, the sharing economy, and digitally-mediated sociality. Since August 2016, she has been conducting an organisational ethnography of corporate software developers in Berlin.
Email: bialski@leuphana.de

Marit Hansen is head of the Independent Centre for Data Protection (ULD) in Kiel, Germany, and the Data Protection Commissioner of the Schleswig-Holstein region. She is a computer scientist by trade, and her work sits at the interface of law and technology. In her years as a researcher and data protection commissioner, she has pioneered the concept of data protection by means of technology and privacy-by-design, through which the ULD has gained its respected status. In 2007, she was appointed by the European Commission as an expert in the 'Privacy & Technology' working group.
Email: mail@datenschutzzentrum.de