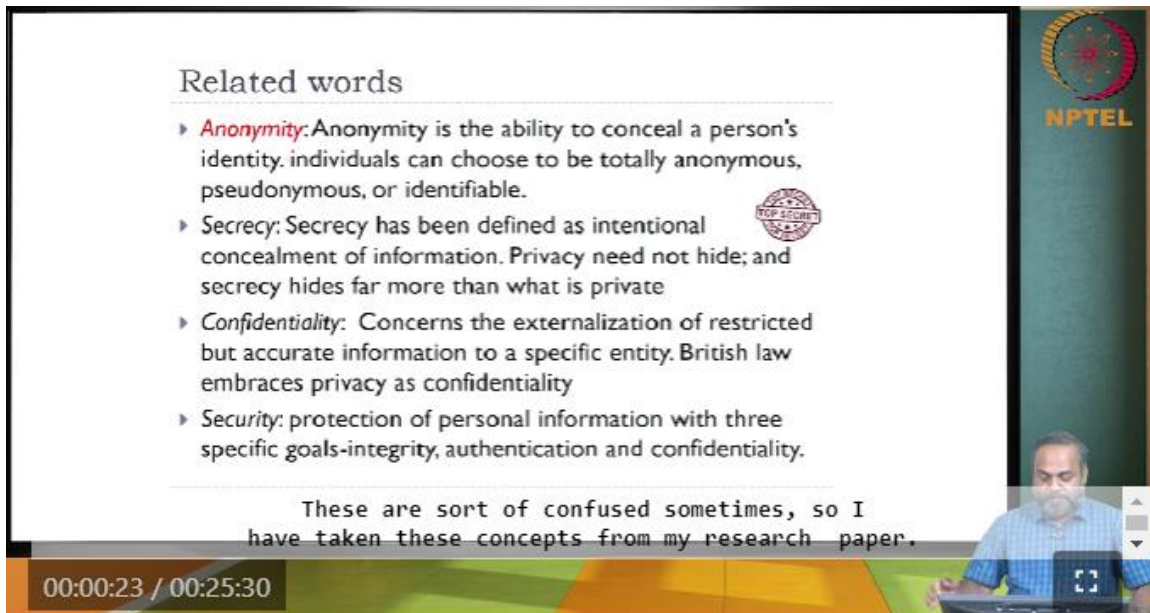


Course Name: Cyber Security and Privacy
Professor Name: Prof Saji K Mathew
Department Name: Department of Management Studies
Institute Name: Indian Institute Of Technology Madras, Chennai
Week: 10
Lecture: 29



Related words

- ▶ **Anonymity:** Anonymity is the ability to conceal a person's identity. individuals can choose to be totally anonymous, pseudonymous, or identifiable.
- ▶ **Secrecy:** Secrecy has been defined as intentional concealment of information. Privacy need not hide; and secrecy hides far more than what is private
- ▶ **Confidentiality:** Concerns the externalization of restricted but accurate information to a specific entity. British law embraces privacy as confidentiality
- ▶ **Security:** protection of personal information with three specific goals-integrity, authentication and confidentiality.

These are sort of confused sometimes, so I have taken these concepts from my research paper.

00:00:23 / 00:25:30

Now there are some related concepts as I said - Anonymity, Secrecy, Confidentiality, Security. These are sort of confused sometimes, so I have taken these concepts from my research paper. So, anonymity is the ability to conceal a person's identity, individuals can choose to be totally anonymous, pseudo anonymous or identifiable. Recently there was a discussion about sharing Aadhaar number with business entities and since if you share your Aadhaar number directly, business can actually misuse it or use it as your identification. For example, go for a telephone, a cell phone service, you give yourself Aadhaar number, they can actually use it for, you know profiling you. So, to prevent that Aadhaar came up with the idea that they can actually create a pseudo identifier or a number which is not same as the Aadhaar number, but it should be issued by Aadhaar.

Aadhaar can link that number to your number, but a business entity cannot or somebody else cannot. So that is pseudo anonymity, but pseudo anonymity is not pure anonymity because an agency can reconstruct your identity. But in any case when you hide your personally identifiable information and then that is called anonymization or anonymity. For example, you are a student and I want to analyse how well the class is doing, in terms

of your academic performance.

Now the class says no. Suppose I am not the faculty, somebody else wants to do that. So, you say no, I do not want to disclose my identity, you want to see how the course is doing or how academics is doing, you can just look at the grades, but do not link it to the individual. So, what I, somebody ask for with your permission I will say, I can give the marks of all the students to you, but I will number it as 1 2 3 4 5, I will not give your name, roll number, email id, anything that is personally identified, I just give the specific data or the attribute that is of interest, ok. Your grade is not a personally identifiable data.

It is a fact you know, it is a measure or a fact you know. We discuss fact table measures, right. Measure is something that you know organization wants to analyze and that is not identifiable data. So, I can give the measure. I can anonymize you.

And now you can imagine how data mining or analytics, this concept is important, you know a company needs to share individual data. It is important to anonymize individuals to protect individuals identity, ok. So, there is, there are actually anonymization techniques. I will discuss that in the next few slides. Secrecy is intentional concealment of information, where you not only want to be anonymous, you do not want to disclose anything to someone and you also can mislead people, ok.

Then there is confidentiality and security. In the next slide where I have a, I have a four quadrants a 2 by 2, the x axis being accuracy of personal information and the y axis is about amount of personal information externalized or shared, ok. So, based on that you can understand this four concepts of secrecy, transparency, anonymity and confidentiality. So, the fourth quadrant is the quadrant which is high on accuracy of personal information and amount of personal information, ok. That is your 100 percent transparent, ok.

So, so we say public institutions should be transparent, there should be nothing concealed. So, you share almost full information accurately, that is transparency. The opposite of that is secrecy. You do not want to share like countries. So, they may actually share information which is not correct or miss, it could be misleading also.

In order to protect your actual information and you do not share any extent or any volume. So that is secrecy, ok. So, secret is a, is a problematic method or a status. We have personal secrets. So, we may hide that somehow from others or we may try to mislead others as well.

Anonymity you can see is the first quadrant where accuracy of personal information is low, but you may share a lot of information. Like in data mining, it cannot be accurately

linked to you. You may be sharing all the information, all the facts about what is going on like your grades, but personal information is very limited, it is not personally linked to anyone. That is the anonymity part. The confidentiality is the other way.

You do not share. Certain information is confidential, ok. Therefore, you make the extent of sharing limited, but when you share, you share the complete information with, you know identity. That is one way of understanding these concepts. Confidentiality, secrecy, anonymity and transparency, but keep in mind that confidentiality is typically used in the organizational context.

Privacy is typically used at the individual level, generally in literature. So, let us understand anonymity. There are some browsers which offers anonymity, you know in browsing, incognito browsing. Each, each browser has its own terminology. Private browsing, incognito etc.


But what is that mean actually? Does it ensure anonymity? Service provider knows ok, but who does not know? Yeah, yeah, yeah, yeah, yeah the local machine that you use to browse. If you do anonymous browsing, it does not keep a history. It does not keep the passwords, it does not track the history. When you actually exit from that browsing, there is no record in the machine, but as he rightly said, the service provider knows the sites that you accessed, ok alright. So, it is sort of providing anonymity to some extent ok, when some users want it.

So, as we said there is also privacy preserving data mining, which is an important topic when it comes to analytics, data analytics. Because companies collect and store large amount of personal information and this personal data, if it is shared with an external entity or if when you outsource analytics ok, it is going to a data processor. And in order to protect individuals, it is important to hide identity of individuals from external service providers. When the data, the data collector or the data owner is conscious of the issue of privacy. So, therefore, one needs to anonymize the data, ok.

So, there could be total totally anonymous data sharing, there could be pseudo anonymous, we talked about it or identifiable, where you share the full information. So, in privacy preserving data mining, there are four, three methods broadly. One is randomization, second is anonymization and third is encryption. There are three main techniques for privacy preserving data mining. Randomization is rather easy to understand, ok.

So, the data provider or the one, the data collector, the one who is the, has collected and stored the data, when it is shared with a external agency or data analyst or a data processor,

you, before you send the data to the data processor, you randomize the data. So, as given in the last paragraph, if x_i is the value of a sensitive attribute, you add a random term e_i , ok. So, what the data processor receive is not the same data, ok. It adds a random value to that data, where e_i is a random noise drawn from some distribution, ok. So, maybe the data processor receives a distributional property of that data, but not the exact data.



Anonymization

- ▶ k-anonymity model widely used (Sweeney, 2002)

Definition 2.2 (k-anonymity requirement) Each release of data must be such that every combination of values of quasi-identifiers can be indistinctly matched to at least k individuals.

- ▶ Uses suppression and generalization

Suppressed

Table 1. Original table

Name	Race	Birth	Sex	Zip	Disease
Alice	Blank	1965-3-18	M	02141	Flu
Bob	Blank	1965-5-1	M	02142	Cancer
David	Blank	1966-6-10	M	02135	Obesity
Helen	Blank	1966-7-15	M	02137	Gastritis
Jane	White	1968-3-20	F	02139	HIV
Paul	White	1968-4-1	F	02138	Cancer

Generalized

Table 2. Anonymization of table 1

Race	Birth	Sex	Zip	Disease
Blank	1965	M	0214*	Flu
Blank	1965	M	0214*	Cancer
Blank	1966	M	0213*	Obesity
Blank	1966	M	0213*	Gastritis
White	1968	F	0213*	HIV
White	1968	F	0213*	Cancer

Look at these two tables. They actually illustrate this concept. In anonymization or in k-anonymization

00:12:38 / 00:25:30

That is randomization. So, randomization is one technique where data processor does not receive the same data, but a randomized data with some pattern, ok. But what is more interesting for us to see is anonymization, ok. Anonymization, I have taken this from a research paper by Sweeney published, Sweeney is from Harvard Business School and this paper discusses a concept called k-anonymity, ok. Just like you have k means clustering, in cluster analysis, there is a k-anonymity and look carefully at the definition of k-anonymity. K-anonymity is defined as each release of data must be such that every combination of values of quasi identifiers can be indistinctly matched to at least k individuals.

So, essentially saying that k has a value, ok. Suppose I ask you your name and your, your roll number, give me your roll number, you are uniquely identified. That is distinct. You are distinctly identified, but I do not ask your roll number. I ask you where is your place of birth? Where is your place of birth? Maharashtra.

Maharashtra, he just said Maharashtra, Maharashtra. Which year you are born? 99. 99 sorry, ok. So, are there any people from Maharashtra here, any other person from Maharashtra or no? You are uniquely identified and with 1991. Suppose there is one more,

one more person from Maharashtra here, who was also born in 99.

Since I am not distinctly identifying you, I have some quasi identifiers, but the problem is with these two quasi identifiers, there are two people, ok. The k becomes 2. These two, these two attributes together can be mapped to more than one person which is 2, k becomes 2. Suppose I make it even more blurred, then it may map to three people I just say where are you born? Maharashtra. There are, say it is a large class, there are eight people from Maharashtra.

There is some identification still, but it has anonymized you to k extent. So, that k is the extent of anonymization. Look at these two tables. They actually illustrate this concept. In anonymization or in k anonymization there are two techniques.

First technique is called suppression, the second technique is called generalization, suppression and generalization. So, compare table 1 and table 2. Can you figure out what is suppression? Table 2 is the anonymized table, table 1 is the original table. So, it has employed two techniques. Table 2 employs two techniques, suppression.

Suppression and generalization, can you figure out these two? Suppression, a column is completely removed, right. You do not find the name column in table 2 at all. So, suppression means completely removing a column.

It is suppressed. It is removed. Some, it has the potential to identify a person uniquely or you know closely. Therefore, that column is removed, but it also applies generalization. Can you compare the birth column in table 1 and the birth column in table 2? Yeah, yeah, yeah, yeah, it has generalized the date of birth. It has removed the date and month it only gives the year ok. So, it is an attempt towards anonymization, not complete anonymization, but anonymization.

You also have to run this idea in mind well analytics, for analytics to be effective, you need some characteristics of individual. So, it is trying to share, well this person was born in this year, but we do not say the date of birth. That may actually lead to identifying the person. So, it is trying to strike a balance between what value you can generate out of analytics, at the same time it does not become identifiable. So, generalization is an effort towards that, year is disclosed, but date and month are not disclosed.

That is generalization. You can see that in zip code also, right. Basically it is generalizing. It may give a sense of which area one belong to or which county, I do not know, but the last digit is removed, right. So, two techniques are used in separation and sorry two techniques, separation and generalization for anonymizing.

Now look at the first record in table 1. It pertains to an individual called Alice. Alice of course, race is not disclosed, date of birth is given, sex is given, zip code is given. These identifiers together makes a unique, there is a unique record for this person. And this person's fact or measure is that the person has a disease called flu, ok.

So, person is sick and the sickness is flu. That is a sensitive data, but it can be identified with the rest of the identifiable data, ok. Now when it comes to table 2, you see there are two people. Blank the first two records. You know blank 1965 M 0214, blank 1965 M 0214, the identifiable data is the same. So, therefore it is, you will say who has flu, it can be record 1 or record 2, ok.

The sickness flu can be attributed to two people. You look at the definition of k anonymity. It can be indistinctly mapped to two people. k equals 2. Actually this table anonymizes and it makes it possibly linked to two people, not one person, ok You can give the same interpretation to other records also.

So, the purpose of anonymization is to remove that unique identification of certain fact with an individual and it tries to anonymize using anonymization techniques, suppression and generalization. And in practice you see that, right. In anonymization is something that you come across in your Aadhaar number, if you have a wallet, the last 4 digit, ok and your credit card information 6709, ok. There are more data. Is it what type of anonymization is this- is it generalization or suppression? When you remove, when you show only 4 digits of an Aadhaar number or a credit card number, then you did not get it right.

In suppression you remove a column as it is. You completely remove it. You look at the zip code, ok. That zip code is generalization, you disclose it, but you remove part of it. Or the date of birth, only the year in a similar way.

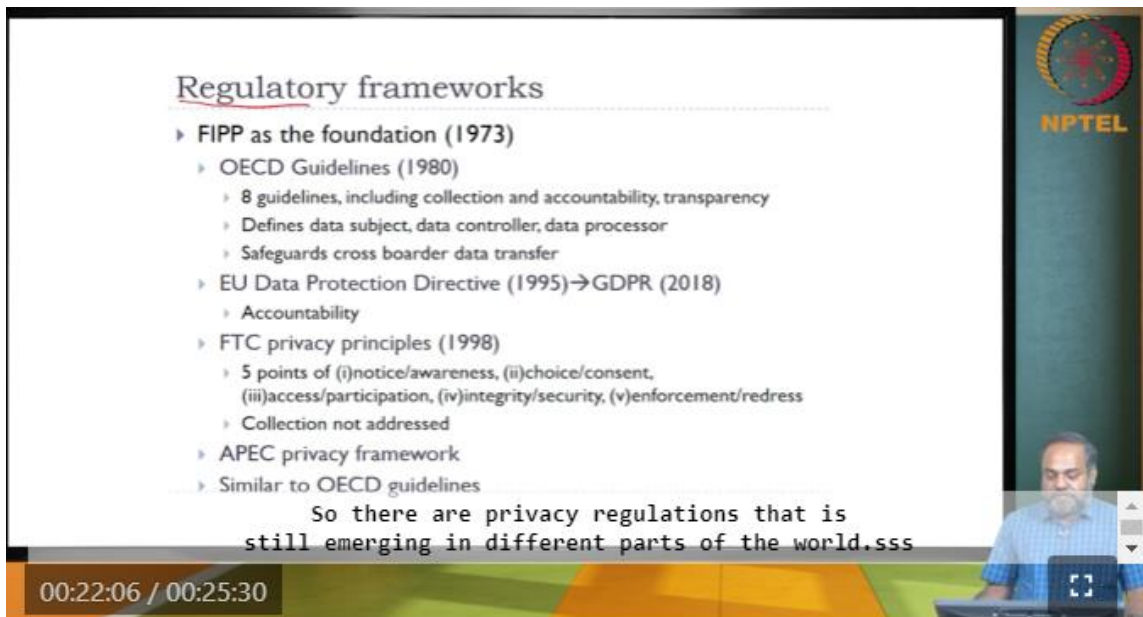
This is generalization This is generalization. You remove a part of it and disclose, only a small part. But these techniques have limitations. When you have certain quasi identifiers, you can also collect more data about those records and link multiple records and create make you uniquely identified. That is known as linking and that is what analytics does in many cases and there are algorithms for linking which can very effectively create identities for individuals, especially in several context like medical or for government. So linking is a, is a particular algorithm in data mining for reverse tracking the identity of individual, when identity is anonymized.

From anonymity you can actually link more data and create identity and that is known

as linking. The last slide, risk versus the last, but one slide, risk versus utility of data, known as R U maps . Risk and utility is a trade off. Privacy is a risk, there is a privacy risk. When you disclose complete information, there is a disclosure risk and that risk is very low when your disclosure is less, ok.

When you completely disclose, your risk goes up. But when your risk goes up, you can see when you move towards this part of the graph you look the you can see, you can see here the risk is going up here, ok or privacy is very low. But the utility from the data is very high or analytics insights are very high ok, but if you actually anonymize data, you move towards here , what suffers is the value of analytics. Especially you cannot profile customers and you cannot recommend products because you do not know who the customer is, at an individual personalization is not possible, ok. So there is a trade off between risk and data utility and that is where you saw the generalization where it generalizes to some extent, but it does not completely remove your identity. So you need to act somewhere, operate somewhere here, so that the value of analytics is not completely compromised or diluted.

At the same time, privacy is protected, ok. So this trade off is illustrated in this graph, risk versus utility of data. Yeah, so our aim today is to understand given individuals and organizations and concern for privacy and organizations need to deliver services effectively, how this sort of need to balance two aspect, need to be done at the country level at at the governance level and that is known as regulation, ok. So there are privacy regulations that is still emerging in different parts of the world.sss OECD regulations came, the FIPPS was the first. We discussed that yesterday in the United States, these were guiding principles.



Regulatory frameworks

- ▶ FIPP as the foundation (1973)
- ▶ OECD Guidelines (1980)
 - ▶ 8 guidelines, including collection and accountability, transparency
 - ▶ Defines data subject, data controller, data processor
 - ▶ Safeguards cross boarder data transfer
- ▶ EU Data Protection Directive (1995)→GDPR (2018)
 - ▶ Accountability
- ▶ FTC privacy principles (1998)
 - ▶ 5 points of (i)notice/awareness, (ii)choice/consent, (iii)access/participation, (iv)integrity/security, (v)enforcement/redress
 - ▶ Collection not addressed
- ▶ APEC privacy framework
- ▶ Similar to OECD guidelines

So there are privacy regulations that is still emerging in different parts of the world.sss

00:22:06 / 00:25:30

NPTTEL

This was not a regulation. These are principles, Fair Information Practice Principles, ok. In a similar way, even in European Union there was no regulation till 2018. It was only a directive, EU data protection directive, not a law, not a regulation. Regulation means it is enforceable by law, but in 2018 it, the directive D got replaced by R, it became a regulation, that is 2018. So you can note that the world is changing in terms of regulating data, regulating data because of the, you know the huge digital data storage and flow that is happening, it has become a necessity.

So in the United States as you will appreciate, there is no one regulation that is binding or that is overarching for all exchanges of personal data, but they have domain specific regulations, like the HIPAA for healthcare data, ok. So for several domains, you will find different guidelines and regulations and this also vary from state to state. US is having, the states have the right to have laws to their own state laws and that also brings an another dimension of complexity, in the US. So, so you will, you will see this landscape is very different but there are certain common attributes. Maybe GDPR would be a very broad, very broad based and this is something most countries are trying to emulate, including our country and some states in the United States also.

So we will actually focus on this in the next two sessions. Tomorrow, next discussion will be on GDPR and subsequently we will look at the Indian context of privacy and security and we will get more insights, along with certain case studies. So today we have a case study which is from the US because we are focussing on America in today's class, ok. So along with the case, you also see some of the limitations. because there is no country wide regulation, alright.

So I will close here. Any questions? Ok. So I will hand over the session to the next group to present the case and discuss the case. Of course, it is a fairly detailed case and it is very challenging, but I am sure you are prepared, alright. Thank you.