

Online Privacy
Professor Ponnuram Kumaraguru
International Institute of Information Technology, Hyderabad
Week 7
Voter and Browser Leaks, Profiling form PII – (Part II)

(Refer Slide Time: 0:18)

This slide shows a search engine results page for the query 'identity theft'. The search results are displayed in a dark theme. The top result is from 'Outlook India' with the title 'What is Identity Theft And How Can You Protect Yourself?'. Below it are results from 'The Financial Express' and 'Times of India'. A red circle is drawn around the word 'identity theft' in the search bar area. The NPTEL logo is in the top left, and the International Institute of Information Technology Hyderabad logo is in the top right. A small video feed of the professor is visible in the bottom right corner.

This slide shows a search engine results page for the query 'identity theft'. The search results are displayed in a dark theme. The top result is from 'Outlook India' with the title 'What is Identity Theft And How Can You Protect Yourself?'. Below it are results from 'The Financial Express' and 'Times of India'. Red circles are drawn around the words 'identity theft' in the search bar area and around the title 'What is identity theft?' in the first search result. The NPTEL logo is in the top left, and the International Institute of Information Technology Hyderabad logo is in the top right. A small video feed of the professor is visible in the bottom right corner.

Now that we have seen how personally identifiable information can be collected from public sources and how it can be actually crawled, collected, stored and analyzed all that. Now, let us look at how somebody can actually misuse this information. How this your information about let us take your voter ID that is being publicly available can actually be misused or your other information, Aadhar information or your other personally identifiable information that is publicly available.

So, if you really think about it one of the biggest problem that can happen because of this information that is publicly available is something called as identity theft. The identity theft happens because this information is publicly available. Now, that I can go actually create a user, create a credit card, create a Aadhar number from the information that is publicly available about you.

So, I can go actually apply for a credit card, I can go get the credit card delivered to my home and then use that for actually withdrawing money, creating a loan, all of that. So, interestingly the identity theft is actually a big problem across the world. If you look at, at least, the US there is a lot of statistics to show that how much money every citizen in the US actually loses because of this identity theft problem.

One of the biggest ways by which the identity theft problem gets created, the identities getting stolen is because of this information that is publicly available. I just did a Google search just to, while preparing the slides, just for the word identity theft and you would see lots of information that is around. And main focus I wanted to give you was this number of hours, lots of things happen in a short span of time on this topic.

Showing that this topic is very, very relevant to the society and very very important in that sense, news articles are being written people are talking about it all that.

(Refer Slide Time: 2:14)

Sourcing, I personally identifiable information, this is necessary because identity theft actually gets created because of getting all this information. What are the different ways that

you can actually think of getting this information? Shoulder surfing is, I am typing the password for getting into my email, somebody is actually behind my shoulders looking at it, sitting next to it and then just peeping the keyboard on what password you are typing.

A general practice that you want to have when people are typing the password is just look away from the keyboard, look away from what they are doing and if you are really curious to know what they are typing probably you will look at it but general practice could be is just to avoid seeing the keyboard. Dumpster diving is another way by which you can actually collect very personal information about users.

For example, the idea of dumpster diving is meaning you throw trash, you probably use credit card bills or you throw your Big Basket bills onto your trash and that gets thrown out of your home into a larger place and now if I was wanting to actually find out personal information about people I would actually go dig that dumb, trash. In that what all I can get, I can get a lot of information.

I actually have seen personally some people's income tax file return, filing returns, all that lying around the trash cans or around the place that I used to live earlier. Not very common, it is not that you will find the income tax filing returns every day, but it is something that you will actually see if you dig into some of these trash. Just think about what all we throw in our trash, can you actually use some of it to create a profile out of you.

A good exercise for you would be to actually look at your trash for a couple of days and see whether you can actually create profiles, if somebody else could actually create profiles about you using that information that is publicly available. So, that is how identity theft could be created and that is how dumpster diving could be done.

Social engineering – Social engineering is a very old technique by which, for example, if I were to get your information I could actually call up, let us take if you are working in a company or you are studying in a college, I could call up, let us take if I end up finding which college you are studying, I could call up the college and say that, “Look I am So, and so, So, and So, is relative, I actually wanted to talk to him.

I am not able to reach to him, it is an important message that I want to give can somebody actually give me cell number from your records, I lost my phone. I am not able to reach to him, all that.” You could actually create a lot of convincing articulation by which you can get

somebody to believe that this information is necessary and this information only you can actually give that is the college.

And this has been done, I mean, Kevin Mitnick, if you want feel free to go look at Kevin Mitnick, one of the first known cybercriminal, one of the techniques that he used versus social engineering, figuring or calling somebody, getting some details about the person A, using that information going to person B, getting some information about person C, putting all them together creating a profile and using it.

Social engineering is a way by which you can actually collect a lot of personal information, like for example; one of the sub-categories of social engineering could be is like phishing. Phishing attack is sending an email, getting you to click on a link and showing a website asking you for a username and password. So, that kind of attacks can actually be very fruitful in collecting personal information.

And when you think about this phishing attack itself I could get you to go to a page, I could get you to download an image and during the process I could actually get a few lines of code on your machine and actually get to see your browsing history. I think last time we talked about, earlier we talked about the browser extensions.

I could actually look at what browser extensions are you using, I mean your browsing history all of that can actually also be captured. Social media – We have been talking a little bit about social media, but there are loads of information that I could actually collect from social media about a user and it is been known that people have misused the connections between people in terms of collecting personal information.

For example, I could actually create a fake profile on some, using some profile information, connect with people, start asking for some personal information, meaning, start asking for secret details of the projects that they are working on, all of that. If you know that they are involved in a government project, if you know that they are involved in something that is not publicly available, can you actually use social engineering, can you actually use these social media to collect that information.

Then the part that, So, this is the part that I just described, this is the part that we already know that government services actually have a lot of information about us online, your voter ID, your Aadhar number, your driver's license, all of this information is publicly available.

So, one question that I have for you to think about is a difference between these sources. Before you go ahead with the video, if you can think about it for a second and scribble some ideas about what could be the difference between these two that will be nice.

The main difference between these two is that the first one is not that reliable sometimes, So, you get some information from social, you get some information by social engineering, it is not clear whether that information is very, very reliable. Dumpster diving, unless if you get handle on something really, really with my name, with my details, then probably you can associate with it, otherwise you know that from the society or trash that we got there were these three patterns that we were able to find.

It will be hard to associate it with PK. Whereas the one of the bottom government data source is like phenomenally reliable, if I get your access to your driver's license with your name on it, I mean, there is nothing that can be wrong in that information, whereas if I go to a profile which says that, let us take Karthikeyan. I do not know which Karthikeyan this is. I do not even know whether Karthikeyan exists in real world with that profile picture, with that face, with that image, whereas in driver's license that is that is available on government sources.

Information is perfectly reliable. So, reliable information from government sources makes it much more harder, much more, So, to say vulnerable for people to use and create identity theft and do malicious activities around it. So, I am going to show you some examples of how government, publicly available information on government websites can be put together and some profile can be created and can it be used for something. I think I will leave it to your creativity to think about that.

(Refer Slide Time: 9:33)

NPTEL

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

Open Govt. Data Sources

- Publicly available
- eGov initiatives
 - Improve services provided – aadhaar changes done online, ...
 - Easy accessible for citizens
- Improved data availability, easy checking / verification
- Public information can lead to privacy breach, disclosure

And why do government sources put this information? One way to think about is that the information is publicly available but you should also ask the question of why is even government putting this information. Why should anybody know what PK's Aadhar number is, what PK's driver's license number is, what PK's voter ID number is. The government is actually making it available publicly because of some specific reasons.

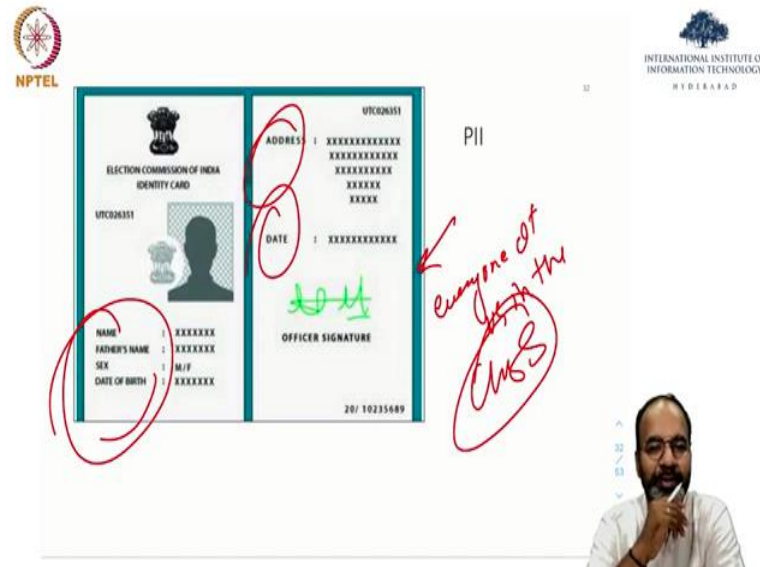
They have this e-government initiatives by which they want to reduce the time, the citizens are actually submitting their applications, citizens are waiting for the details, all of that. By putting the information public also the services that the government is providing us is also increasing. Now you are able to actually sit at home, pay your electricity bills, pay your water bills, pay your, So, to say, real estate bills, taxes, income tax has become very much online.

You could you could apply for Aadhar changes that you need in your Aadhar number, let us take home address or something; it is very easy to sit and do it at home now. That is the main reason why many of this information is also made publicly; information, services, all of that is made publicly available. Improved data quality and easy checking verification also So, data quality, information is available, So, you can actually check.

You could actually make an argument that, oh, something is not correct, you can go back and fix it, verification is also very, the verification becomes much more simpler when you are able to see what information the government has about you, but unfortunately if the

information is publicly available, then it could be actually used for breaches, it could be actually used for identity theft that I said before.

(Refer Slide Time: 11:35)



So, this is a, this slide shows you about Election Commission of India, potentially an example, just shows what all information about you is publicly available that can be actually crawled. Name, father's name, gender, date of birth, address, date of issue all of this could be actually got. Not just yours actually, imagine if this is gotten for every one of us in the class.

Then we can actually find out some group information that is available that may not be directly available also. So, it is not about just individual information, for example, if I could get this information and I could actually profile your family together, who are you living with, what could be potentially the relationship that you have with that person, those can become pretty damaging.

(Refer Slide Time: 12:39)

The slide is titled "PII" and features the NPTEL logo on the left and the International Institute of Information Technology Hyderabad logo on the right. A central list of personal identifiers is shown, each with a red checkmark to its right:

- Voter ID
- Name
- Father's name
- Age
- Gender
- DoB
- DL Number
- PAN
- Phone #

A small video inset of a man is visible in the bottom right corner of the slide.

A simple list of what PII could be voter ID details which is which can give you a lot more information, your name, father's name, age, gender, date of birth, driver's license number, PAN number, and phone number. We can generate a longer list also I am sure you can actually add some very specific attributes, information about us in this list, if you find something that we can add to this list and which are personally identifiable please send it in the email list, we can actually take a look at it and have a conversation there.

(Refer Slide Time: 13:16)

The slide shows a screenshot of the yasni search engine interface. The search results for "Brijesh Gupta" are displayed, with several items circled in red:

- The search bar and the search results header.
- The profile card for "Brijesh Gupta" (I'm Brijesh Gupta).
- The "Connections & Addresses" section, which lists two entries for "Brijesh Gupta" with addresses in "Hydrabad, India".

The yasni logo is visible on the right side of the slide. A small video inset of a man is visible in the bottom right corner of the slide.

Here are some services that was available at some point in time for looking at the personal information, looking at publicly available information, profiling them and seeing the details

together. This is Yasni.com. So, this is one of my students that used to work at that point in time, Srishti Gupta, searching for Srishti, you can get details about old, some home addresses, books, some profile pictures, some details about Srishti.

So, this kind of services have become very, very popular in the last decade or so. Before the services were useful for let us take job verification, you are applying for a job, the company wants to give you the job, but the company wants to make sure that all the information that you are provided is correct and want to do some background checking all of that, So, the middleman who has this aggregation of all the information about PK becomes very very valuable in terms of verification.

(Refer Slide Time: 14:21)

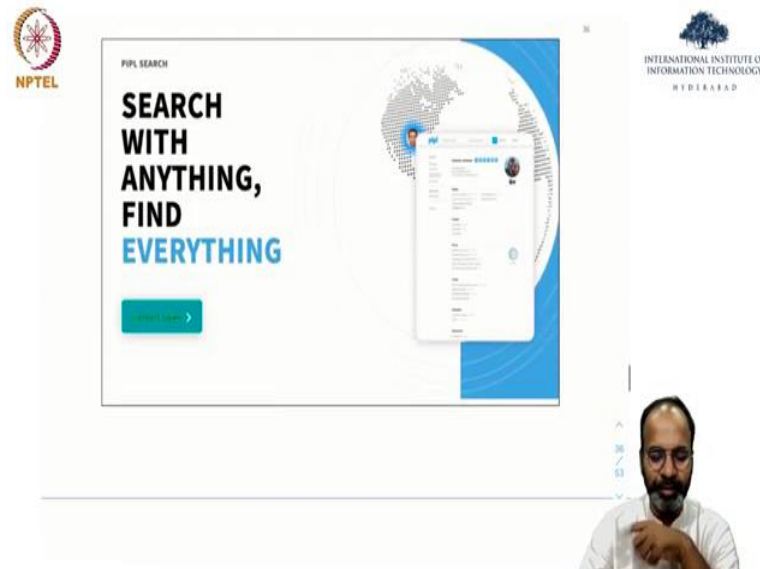


Another site which is actually popular when I was preparing the slides I actually went to the website to see what are the things that they are doing now, So, again the same search Srishti Gupta that is her actual picture, that is her actual LinkedIn page, and then Srishti Gupta, New Delhi, So, giving information, more and more information that is available, if you feed it to these kind of systems they become much more accurate.

I am sure you can go back and relate to the great hack video that we saw in the week one, where the argument was the Cambridge Analytica was able to actually find information about us, about US citizens particularly with like 3,000, 5,000 features of ours which is what I am talking about, if you are able to feed in more and more Ponnurangam Kumaraguru IIT Hyderabad, Hyderabad, professor, computer science.

Studied in the US, lived in a particular city and there is all these information if you actually give the system can be more and more effective, more and more accurate, the precision can be much, much higher.

(Refer Slide Time: 15:38)



So, this is, that is the website PIPL.com, feel free to go, take a look at it again, they seem to have forked into a lot of services which can be very, very useful for finding out frauds. I think, one of the on their website talks about why this profiling is extremely important is to find out frauds, find out fraudulent users, find out fraudsters in general.

Because knowing that PK information that is given in one place is not the same as in the other one can actually put a red flag, So, when we are actually being, PK is being looked for some information, this this red flag can show up and appropriate decisions can be made.

(Refer Slide Time: 16:24)

The slide features the NPTEL logo on the left and the International Institute of Information Technology Hyderabad logo on the right. The title is 'Country specific systems with Open Govt. data'. A table lists four services with red annotations: 'IndianKanoon' (India) with 'Legal search engine' circled and 'indexes judgement of supreme court and high courts' underlined; 'Opencivic.in' (India) with 'Data about state assembly elections, and profiles of MPs of states' underlined; 'ABQRide' (USA) with 'City buses, fares for other public transportation' underlined; and 'Illustreets' (UK) with 'Crime, education, transport, census data of a location' underlined. A speaker is visible in the bottom right corner.

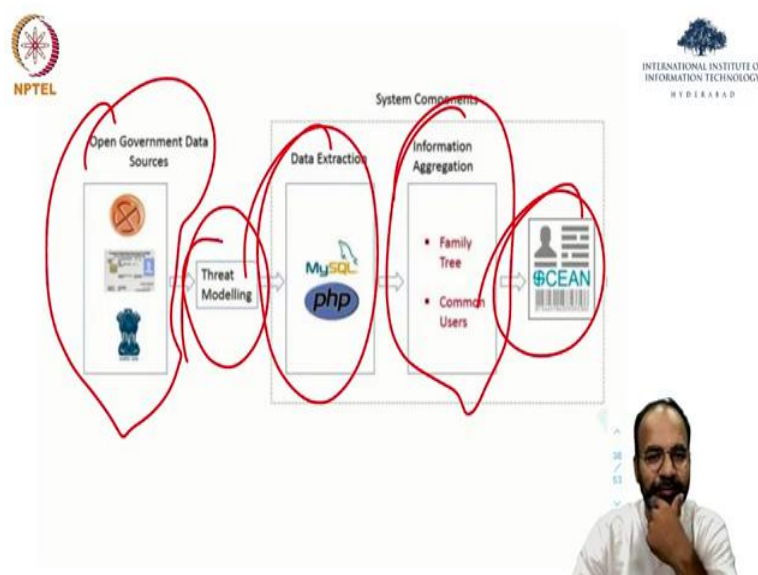
Name	Country	Description
IndianKanoon	India	Legal search engine indexes judgement of supreme court and high courts
Opencivic.in	India	Data about state assembly elections, and profiles of MPs of states
ABQRide	USA	City buses, fares for other public transportation
Illustreets	UK	Crime, education, transport, census data of a location

Just a quick list of services that are available for looking at government open data, what information is available on some of these services. The first one is IndianKanoon which actually gives you a nice search engine where you can actually search for judgments of Supreme Courts and High Courts. And Opencivic dot in is again based in India or based on data from India, data about state assembly elections, and profiles of MPs of states.

The next two are outside India; ABQRide gives you the details of city buses, fares for other public transportation. So, these kind of services have become much more prevalent now in using publicly available information, using these kind of transportation information that is publicly available, can you actually put them together to create some service with it.

Illustreets gives you details about crime, education, transport, census data of a location. So, these kind of publicly available services are also become very, very important to do some interesting analysis, interesting pattern predictions, interesting user behavior, So, there has been a lot of studies and a lot of need for having this open data also arguments to make for having this open data.

(Refer Slide Time: 17:59)



So, data collection, So, as I said I am going to show you one example where publicly available government information was actually crawled, analyzed, profile was created, profiles were created to see what we can do with it. The general architecture for many of these infrastructure that is built is first is some type of crawling, then pre-processing of information and storing of information and then using the stored information can be actually, can we actually create profiles with them.

That is exactly what is in this picture also government sources, driver's license, voter ID, all this information that is publicly available, can you do a threat modeling. We will actually see what a threat modeling is and how threat modeling becomes very handy. So, the idea of threat modeling is that you want to know where the loopholes are, where the weakest link is in the complete ecosystem that you have built, for example, ATM machine,

ATM machine, ICICI setting up an ATM machine scene or the idea of setting up an ATM machine in a particular location, you want to do the threat modeling to understand what are the threats that are available in that particular location. The threats in different locations can be different, So, threat something that is in Bangalore Mahatma Gandhi Road may be very different from an ATM machine that is kept on let us say Kukatpally in Hyderabad.

So, these threat modeling becomes very essential to actually build solutions around it. So, then the data extraction which is MySQL and PHP and then information aggregation, can you

put this information together to create a profile and then there was a system called OCEAN that was built, which actually became an interface for seeing this information.

(Refer Slide Time: 19:54)

Next couple of slides are the ways that you could actually see these driver's license, voter ID, what input is given and what output is got. These may not be the services, these may not be the way it is now when you go look at it and also remember all of this is from Delhi, So, if you are from cities that are away, other than Delhi, the entire interface, the entire information that is available may be very, very different.

So, this one, enter the driver's license number, input and the output is date, address, parent's name, blah blah, date of birth, all of this is showing up and these are all illustrative purposes,

but yeah. And if you see how do you get the driver's license, you again as we said before can we actually look at other sources to get this information. Social engineering that is what would happen, I get information one, let us take your name.

I get the information that your name uh is Srishti Gupta itself in this example. I got Srishti Gupta, I will use the service one, which takes input as a name and output as something else, input is name output is something else. Let us say output is age, gender and let us take driver's license number. Now, I will use this driver's license number as an input to another service and then get something else. So, this is how I will actually build a profile of Srishti Gupta.

(Refer Slide Time: 21:36)

The screenshot shows the 'Voter ID' portal of the Chief Electoral Officer, Delhi. The search criteria are: Assembly Constituency Number & Name: ROHINI (GEN) (circled in red), Voters' Name: SUSHI, and Father's Name: NOT REQUIRED. The search results table is as follows:

Sl. No.	Part No.	Serial No.	Section No.	Voters' Address	Voters' Name	Relation	Relation Name
13	1	1788	1	305/A-1, A BLOCK RAJA VIHAR BADLI	SRISHTI SINGH	Husband	ANAND SINGH
13	23	310	1	C-115, BLOCK C MILLENNIUM APPTT. ROHINI SEC-18	SRISHTI NIM	Father	PREM CHAND NIM
13	25	1078	3	E-3/206, PNT E3, LIG FLATS ROHINI SEC-18	SRISHTI SATWARI	Father	PURSHOTAM SATWA
13	28	130	2	E-5/75, E-1 BLOCK PNT-7 ROHINI SECTOR-15	SRISHTI VATS	Father	YOGINDER SINGH VATS

The screenshot shows the 'Voter ID' portal with search criteria: Assembly Constituency Number & Name: ROHINI (GEN) (circled in red), Voters' Name: SUSHI, and Father's Name: NOT REQUIRED. The search results table is as follows:

Sl. No.	Part No.	Serial No.	Section No.	Voters' Address	Voters' Name	Relation	Relation Name
13	1	1788	1	305/A-1, A BLOCK RAJA VIHAR BADLI	SRISHTI SINGH	Husband	ANAND SINGH
13	23	310	1	C-115, BLOCK C MILLENNIUM APPTT. ROHINI SEC-18	SRISHTI NIM	Father	PREM CHAND NIM
13	25	1078	3	E-3/206, PNT E3, LIG FLATS ROHINI SEC-18	SRISHTI SATWARI	Father	PURSHOTAM SATWA
13	28	130	2	E-5/75, E-1 BLOCK PNT-7 ROHINI SECTOR-15	SRISHTI VATS	Father	YOGINDER SINGH VATS

Voter ID, same thing, if you get to know where does Srithi live even approximately, write it some name Kamal Vihar or some Kukatpally that level of information is available, you could actually feed that information, which district is she part of and then you can actually get some relevant information. So, in this case it is saying father's, mother's, husband name, house name, voter photo ID card number. So, you could actually start using information from different sources and creating a profile.



(Refer Slide Time: 22:15)

The screenshot shows a web interface for tracking PAN/TAN application status. The main content area is a form with the following elements:

- Header: "Track your PAN/TAN Application Status" with the NSDL logo.
- Section: "Please select a type of application" with a dropdown menu for "Application Type" (options: PAN, New, Change Request).
- Section: "Please select Status and ID release field" with a dropdown menu for "ACCORDANCE DOCUMENT NUMBER" (value: INPUT).
- Form fields: "Last Name/Surname" (value: SRIHARI), "First Name", and "Middle Name".
- Date fields: "Date of Birth/Registration" and "Agreement/Partnership or Your Usual residence or Body of Individuals/Partnership of Persons".
- Footer: "Name should be as mentioned in the application form. Agreement other than Individuals should only fill name in the field for Last Name/Surname only." and a "SUBMIT" button.

Here is another one and then input being name and some date of birth and mother's maiden name, those are the kinds of information that is asked beyond just the name, which can actually be a little harder for people to get but once you have it I think you can get a lot more information from...


(Refer Slide Time: 22:36)



PAN Output



Your PAN Application Status

Acknowledgment Number	: BTM1100024451	OUTPUT
Name	: MOHIT MOHANI	
Category	: Individual	
Status	: Your PAN card has been delivered on 22-Feb-2019 by <i>Express Courier</i> . For L12 courier vide address of no. 225440076 at the address for communication* indicated by you in the application.	
Permanent Account Number (PAN)	: ASDP4652G	

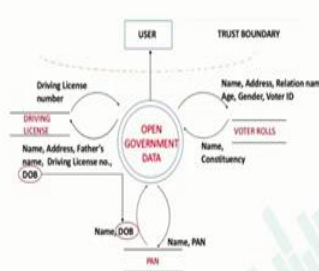


So, this one was actually pretty, very intrusive in that sense which is to get the pan card details about people.


(Refer Slide Time: 22:50)

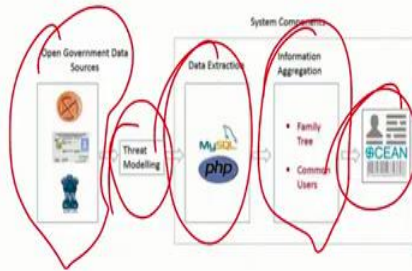


Threat modeling



43 / 53





And threat modeling - So, as I said what is threat modeling, threat modeling is a way by which you see how you can actually protect the details of the infrastructure that is being built. So, that is how you do data collection from publicly available sources particularly in this case government websites. Let us look at some specific examples of where, how information can be collected from online.

(Refer Slide Time: 23:24)

License Details			
DL Number	DL-0200000790	Last Activity	Opticon
DL Issue Date		DL Expiry Date	
DL Non-Transport valid from		DL Non-Transport valid To	





Driver License

Online License Details	
Enter Driving License No. <input type="text"/>	Search <input type="button" value="INPUT"/>
DL No.: <input type="text"/>	Old DL No.: <input type="text"/>
Name Of Applicant: Srishti Gupta	OUTPUT
Son Daughter/Wife of: <input type="text"/>	
DOB: <input type="text"/>	
Address: <input type="text"/>	
Pin Code: NEW DELHI	
License Details:	
DL Number: DL-62000000750	Last Activity: Database



So, here is an example of how to get driver's license. So, the input is named, So, again for practical purposes for the entire lecture, let us keep it Srishti Gupta as the input wherever it is, Srishti Gupta input is given and accordingly you could actually get some information from this driver's license website. So, here is how I would see how do actually social engineer getting more and more information.

So, we have Srishti Gupta, let us take we know the name. From the name look at services that takes only his name and gives outputs as something else, let us take driver's license. From here, So, now you know the name and you know the driver's license number here, can you find a service that takes these two as input and then produces something else as output.

So, doing this will help you to just add more information to Srishti and make the profile more and more, So, it is a lot more details can be added to these profiles. And it is also interesting to see how much of personal information you can actually draw from these government services only using some simple information to start with. And these are examples taken some years back.

(Refer Slide Time: 24:42)

Official Website of
Chief Electoral Officer, Delhi

Check Your Name in the Voters' List (Electoral Roll)

Assembly Constituency Number & Name: KOHAS GDS INPUT

Voter Name: [Redacted]

Voter No.: [Redacted] NOT REQUIRED

Voter Photo Identity Card No.: [Redacted]

S. No.	Poll	Ward	Section	Voter Name	Address	Voter Name	Poll No.	P. Card No.
12	1	108	1	2004 L 4 BLOCK RAJA VISHW SAGAL	Ward No. 108/108	25	108/108	[Redacted]
13	10	108	1	C 108 BLOCK C BILKISHAN APPT. KONGA SEC 10	Ward No. 108/108	25	108/108	[Redacted]
14	10	108	3	K 108/3 PVT. EL. LUG PLATS KONGA SEC 10	Ward No. 108/108	25	108/108	[Redacted]
15	10	108	3	K 108/3 PVT. EL. LUG PLATS KONGA SEC 10	Ward No. 108/108	25	108/108	[Redacted]

NSDL
Securities, Trust & Bank

Track your PAN/TAN Application Status

Application Type: New

PAN/TAN APPLICATION NUMBER: INPUT

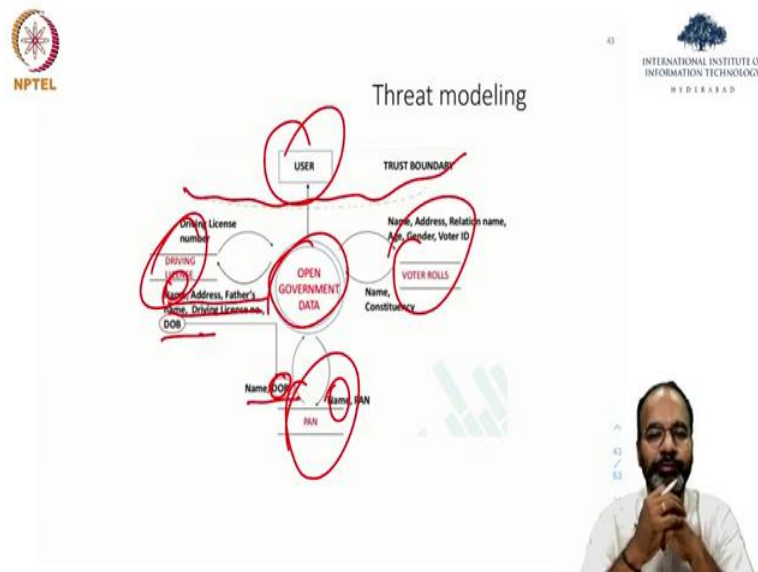
NAME: [Redacted]

S. No.	Name	Address
1	[Redacted]	[Redacted]
2	[Redacted]	[Redacted]
3	[Redacted]	[Redacted]

So, the exact page may not look like this, but some of these services now are available at some point in time. So, this is a voter ID, you again you find, So, from knowing Srishti Gupta you have got some voter ID details, So, from there that which district, which location she lives, can you use now Srishti Gupta.

And that location is an input for a service which is taking input as name and the location and giving you something else which could be the voter ID. So, that is the, from a driver's license number, from a driver's license detail you can come to voter ID. So, that is the input, can you actually give the input as Srishti Gupta, let us take Kukatpally or Delhi, it should be, it could be Hauz Khas or Okhla Phase 3, something like that, use that and then can you take some, can you get the voter ID details from the service. That is the pan output.

(Refer Slide Time: 25:38)



So, threat modeling, I mentioned earlier what a threat model is. Threat modeling is a important idea to understand when any systems are built you want to know what the threats are. Threats particularly, you want to understand threats from various points. So, in this example we were seeing the government services being here, user being here and then that is the threat boundary and if you see there is driver's license, voter ID, PAN card, these are different services that are available.

So, now the question to do threat modeling is that can you understand what threats are there if I get access to let us take name, address, father's name, date of birth from here, date of birth is here, name and date of birth from here. If I put to put these two data birth together can I actually find this name, which may be unique.

So, you have to enumerate the different threats that is there and of course, there is a formal mechanism of dread model, which I think a slide later or one later I have a table to show how to actually measure these values but these dread models are, these threat models are necessary to do to understand to get a feeling of what the systems are being built up.

So, dread models come in, I mean, the same dread model can be used for, let us take a security for the campus or the security environment compound wall and the security guards that you probably have in your society where you live or the security at the airport, all of this is places where threats happen, So, threat modeling becomes very, very relevant and you can actually build solutions depending on what you find from these threat models.

You can again classify it as low medium, high and depending on the risk that are there you can build solutions around it.

(Refer Slide Time: 27:28)

Attacks, creating fake voter ID

E-REGISTRATION (Form-6)

Application for inclusion of name in electoral roll

District: Please Select

To: The Electoral Registration Officer

Assembly Constituency: [Dropdown]

I request that my name be included in the electoral roll for the above Constituency. Particulars in support of my claim for inclusion in the electoral roll are given below:

1. Applicant's Details

Name: [Text] Surname (if any): [Text]

DOB: [Text] Sex: [Text]

Age as on 01 January 2018: [Text] [Dropdown]

I request that my name be included in the electoral roll for the above Constituency. Particulars in support of my claim for inclusion in the electoral roll are given below:

Name: [Text] Surname (if any): [Text]

DOB: [Text] Sex: [Text]

Age as on 01 January 2018: [Text] [Dropdown]

Attacks, creating fake voter ID

E-REGISTRATION (Form-6)

Application for inclusion of name in electoral roll

District: Please Select

To: The Electoral Registration Officer

Assembly Constituency: [Dropdown]

I request that my name be included in the electoral roll for the above Constituency. Particulars in support of my claim for inclusion in the electoral roll are given below:

1. Applicant's Details

Name: [Text] Surname (if any): [Text]

DOB: [Text] Sex: [Text]

Age as on 01 January 2018: [Text] [Dropdown]

I request that my name be included in the electoral roll for the above Constituency. Particulars in support of my claim for inclusion in the electoral roll are given below:

Name: [Text] Surname (if any): [Text]

DOB: [Text] Sex: [Text]

Age as on 01 January 2018: [Text] [Dropdown]

So, some of the attacks that you can think of like the point of talking about all this was the identity theft, some of the attacks that you can actually think of is creating fake voter ID, I know that Srishti Gupta lives somewhere, can I actually apply for a particular fake ID as Srishti Gupta and actually get it on Srishti Gupta's name.

So, if you see this is a registration form for getting voter ID, it is asking for name, date of birth, father's name and gender main. Some of these information you could, we saw that we could actually get it from other sources, even other information is also village, town, district,

So, if you have access to some services that we discussed before we should be able to generate some of this information.

(Refer Slide Time: 28:15)

Attacks, view tax statements

Sr. No.	Name of Deductor	TAN of Deductor	Total Amount Paid (₹)	Total Tax Deducted (₹)	Total TDS Deposited (₹)
1					

Attacks, view tax statements

Sr. No.	Name of Deductor	TAN of Deductor	Total Amount Paid (₹)	Total Tax Deducted (₹)	Total TDS Deposited (₹)
1					

So, this one was a little stretch, we were able to we were able to get to the state because we had some more sort of insider threat, some information that we know because we, some of the users, because we actually know these users also otherwise. Insider threats are problems that happen because of people inside the organization itself.

I could have, meaning, IIIT Hyderabad having some information being leaked is because somebody inside IIIT Hyderabad had to bother access to that information they decided to actually make it public. So, one of the other things in this case, we were able to get to this

level because we had access to some insider information. We were able to get to this level of income tax file returns details.

(Refer Slide Time: 29:01)

The slide displays the DREAD model table with handwritten annotations. The table is as follows:

Rating	High (3)	Medium (2)	Low (1)
D Damage potential	An attacker can subvert the entire system, get full authorization, run as administrator and content.	Leaking sensitive information.	Leaking trivial information.
R Reproducibility	The attack can be reproduced every time and does not require a timing window.	The attack can be reproduced, but only with a timing window and a particular race situation.	The attack is very difficult to reproduce, even with knowledge of the security hole.
E Exploitability	A novice programmer could make the attack in a short time.	A skilled programmer could make the attack, then repeat the steps.	The attack requires an extremely skilled person and in-depth knowledge every time to exploit.
A Affected users	All users, default configuration, key customers.	Some users, non-default configuration.	Very small percentage of users, obscure feature, affects anonymous users.
D Discoverability	Published information explains the attack. The vulnerability is found in the most commonly used feature and is very noticeable.	The vulnerability is in a seldom-used part of the product, and only a few users should come across it. It would take some thinking to see malicious use.	The bug is obscure, and it is unlikely that users will work out damage potential.

Vertical text on the right side of the slide reads: D R E A D M o d e l. A presenter's video feed is visible in the bottom right corner.

The slide shows a threat modeling diagram titled "Threat modeling". It features a central "OPEN GOVERNMENT DATA" node surrounded by various data sources and users, all enclosed within a "TRUST BOUNDARY".

- USER** (top): Connected to the central data node.
- DRIVING LICENCE** (left): Includes fields like "Driving License number", "Name", "Address, Father's Name", "DOB", and "Driving Licence no.".
- VOTER ROLLS** (right): Includes fields like "Name, Address, Relation name, Age, Gender, Voter ID" and "Name, Constituency".
- IRAN** (bottom): Includes fields like "Name, DOB" and "Name, PAN".

Handwritten red circles and lines highlight specific data points and connections within the diagram. A presenter's video feed is visible in the bottom right corner.

So, this is the dread model that I said which is dread stands for Damage Potential, Reproducibility, Exploits, So, dread affected users and discoverability. Low, medium, high, these are the ways by which as I said you can actually classify the rating and you can find these dreads can be done for every aspect of this, every aspect of the infrastructure that you have and use that to decide a high, low, medium for every single point.

So, there may be many gates in your campus, house, housing society, can you actually define threat models for each of them. We have online So, there are many exit points and entry

points for our internet services that we use on a campus, can you actually find out what the threat models are, can you measure them, can you quantify them. Quantifying helps because depending on that you can decide solutions again.

So, this one is dread model, dread model here is So, high, medium, low, the damage potential, the put, the chances of the potential of making the attack, leak sensitive information because it is, you can access from the government websites, you can get a lot of sensitive information from there. If it was low it will be leaked trivial information, if it is high, attacker can subvert the security system, get full trust authorization, runners administrators, upload content.

It is not necessarily through this or giving Srishti Gupta as the input and taking it as details from the interface, all that, even if I have, if it is a high risk environment somebody as an administrator can get access to the database and just change the details, upload a new set of files and then make all these details and to show up there.

Reproducibility, the attack can be reproduced every time and does not require a timing window, does not require a particular context. The attack can be reproduced but only with a timing window and a particular race situation, which is a particular context is necessary if the medium level of reproducibility is possible. And then the lowest it is just hard to reproduce the attack.

Exploitability, a novice programmer could make the attack in a short time, for if it is high of exploitability somebody easily, let us take an undergrad student or a 12th standard student should be able to get access to the database and do attacks. For a medium little bit of a skill, for a low the attack requires extremely skilled person and in-depth knowledge every time to exploit. Affected users - All affected users default configuration key customers.

Imagine affected users is high in a context of where let us take our LinkedIn user username, passwords is getting hacked, got hacked. In that the affected users are high because it is everybody, whoever was in the database, whoever was the users on LinkedIn. Some users non-default configuration, it is just medium because whoever, let us take if the attack happened on a campus or on a network whoever was part of that network got affected.

Obscure features affects anonymous users, the affected users is really, really small, again the context is very small, it is taken, in one room there was a breach of some information.

Discoverability - Discoverabilities how does these attacks gets discovered, discoverability is high if published information explains the attack. The vulnerability is found in the most commonly used features and is very noticeable.

It is easy to reproduce these kind of attacks, not necessarily, I mean, think of stuck nut, probably it is in the low side, the bug is obscure and it is unlikely that users will work out the damage potential, stuck nuts kind of thing would be at that level, discoverability is extremely low. In in the medium side the vulnerability is seldom-used part of the product, it is something potentially possibly you could actually find out if you spent some time.

They all look slightly subjective and they also look slightly not very quantifiable, but that is how this method is where you just mark it as a high, medium, low, for every entry or exit points. Then you kind of accumulate all of them, find a way to put them all together and then you can say that look IIIT Hyderabad has the, from the output of the dread model to IIIT Hyderabad has So, much of risk.

(Refer Slide Time: 33:42)

The image shows a voter information card for Srishti Baswat. The card includes the following details:

- NAME: Srishti Baswat
- VOTER ID: [Redacted]
- FATHER'S NAME: Prem Baswat
- Age: 23-25
- Sex: Female
- Address: [Redacted] Karnool Pura Baswat

Below the personal details is a family tree diagram titled "Family Tree of Srishti Baswat". The diagram shows a family structure with three individuals: a father (Prem Baswat, Age 57, Voter ID [Redacted]), a mother (Baswat, Age 54, Voter ID [Redacted]), and two children (Srishti Baswat, Age 24, Voter ID [Redacted] and another child, Age 35, Voter ID [Redacted]). A red circle highlights the family tree diagram.

The slide also features the NPTEL logo in the top left corner and the IIIT Hyderabad logo in the top right corner. A small inset image of a man is visible in the bottom right corner.



• Can be used / mis-used



Information in the chart:
Age, Gender, Source
station, Destination sta-
tion, Seat number, First
Name, Last Name, Pas-
senger Name Record
Number



You could do this for every single system and in this context it is done for the government services that are publicly available. So, I kept talking about Srishti Gupta as an example. Here is again an example, where you could actually think of going back to the train example, going back to the train picture that I showed earlier. Yeah, I showed earlier that what information is publicly available.

Can you actually use this publicly available information and have a conversation with the person who is actually next to you next in your train, in your compartment, that is the kind of motivation that we had when we started looking at this problem. So, in this case this output is, input is let us assume that somebody who is in your train, in the compartment you saw the name, you wanted to look at who this person is.

You gave the input into the system, Srishti Gupta again let us take, you get the voter ID, all that, the most interesting aspect here in this example is Srishti Rawar, the most interesting thing is the family structure that you can get. I said earlier also that building these kind of services to put information together and create profile not necessarily only you can be actually attacked or something, it can be the group that you are connected to, the family that you are connected with all of them can be actually attacked here.

So, here it is the family structure that you can actually build and family structure helps because, now when you are getting into a train you can actually arguably say that, oh, I know So, and So, from this group, is this your father, is this your brother, did your brother go to the school, all these kind of conversations can be built because information is publicly available about people.

(Refer Slide Time: 35:35)

The slide features the NPTEL logo on the left and the International Institute of Information Technology Hyderabad logo on the right. The title 'Challenges, aggregation' is centered at the top. Below the title, two bullet points are listed: '• Common names like Manish Gupta is hard 😊' and '• Not many users link their profile'. A red circle highlights the name 'Manish Gupta' in the first bullet point. Below the text is a screenshot of the Linktree website. The screenshot shows the headline 'The Only Link You'll Ever Need' and a list of social media links. A red circle highlights the URL 'https://linktr.ee' at the bottom of the screenshot. A small video inset of a man speaking is visible in the bottom right corner of the slide.

Of course, there are lots of challenges; it is not that trivial to create these kind of profiling services very easily, So, for example, common name like Manish Gupta, if you search for it in in India, in north India, in Delhi databases you are going to get like tons of Manish Guptas. So, it is very, it will be very hard to find out which Manish Gupta are we talking about, who is in the train.

And also that, for example, this this service Linktree is extremely popular because they argue the argument that the Linktree makes it as look the only link that you need to know is the Linktree link, because where users come and link Facebook, Twitter, LinkedIn, Instagram, all profiles to this one service and Linktree knows all the profiles that you have in different platforms. Why is that helpful?

That helps because it makes lives much more simpler to actually maintain my profile, I do not have to tell everybody that look my profile in Facebook is this, Twitter is this, Instagram is this, I just say that, okay here is my LinkedIn tree, Linktree profile and please go take a look at all my profiles, but on the contrary given that all these you as a user providing this information the easiness of doing an attack is also very, very high.

Because I get to know all the profiles of yours in one shot, So, that is where the Linktree becomes very useful or very not useful in that context and not many users are actually using Linktree. For example, I do not have a Linktree profile, So, it may not be that easy to get all users on Linktree. I am going to also ask, So, if you find yourself using Linktree, if you think that is how useful, how not useful it is please share it on the mailing list, we can talk about it.

(Refer Slide Time: 37:33)

The slide features the NPTEL logo on the top left and the International Institute of Information Technology Hyderabad logo on the top right. The title 'High risk users' is centered. Below the title, there are two bullet points: 'Users with Voter ID, DL number, PAN' and 'Around 2K people in the DB had all the above'. A video feed of a man in a white shirt is positioned at the bottom right of the slide.

High risk users

- Users with Voter ID, DL number, PAN
- Around 2K people in the DB had all the above ☺

So, this government was services, users with voter ID, driver's license number, PAN number all of them are, all of them we could actually crawl from these services and around 2K people had all the above, So, which is I could actually find out Manish Gupta whose driver's license is this, whose voter number is this, whose PAN card number is this.

(Refer Slide Time: 37:57)

The slide features the NPTEL logo on the top left and the International Institute of Information Technology Hyderabad logo on the top right. The title 'User feedback' is written vertically on the right side. There are five yellow speech bubbles containing user feedback. A central icon shows three stylized people. A video feed of a man in a white shirt is positioned at the bottom right of the slide.

User feedback


"It was an eye-opener to a common man."

"I am really shocked that the exact ID numbers are available online without much security against data mining at this scale."

"A great shortcoming and security flaw has been pointed out by OCEAN. Great work."

"Waiting for an upgraded version which will work for other states also."

"Good system. Great work I Didn't know such a system existed."



Srishti [Redacted]

NAME: Srishti [Redacted]
 VISIT ID: [Redacted]
 FACEBOOK NAME: Srishti [Redacted]
 Age: 23-25
 Sex: Female
 Address: [Redacted] [Connect For Detail]

Family Tree of Srishti [Redacted]



Parents: [Redacted] (Age 33) and [Redacted] (Age 34)
 Children: [Redacted] (Age 24) and [Redacted] (Age 25)



And at some point in time we decided that we will make this service, we will talk to people about what do they think about these kind of services, information put together, all that, meaning I am pretty sure let us take, for example, I will ask you also how intrusive do you think this is which is starting from your name going to actually a family structure. Again reply, post on the email, we can take a discussion around this also.

So, the responses that we got was it is an eye opener to a common man waiting for an upgraded version which will work for other states also because we are done only for Delhi. I am really shocked that the exact ID numbers are available online without much security against data mining of at this scale. A great shortcoming and a security flaw has been appointed out by OCEAN Great work. Good system, Great work, we did not know such a system existed.




(Refer Slide Time: 38:53)



51

Takeaways

- Large amount of personal information is available on government services
 - Will be great if some of you can explore this idea in your work city / state websites
- Information aggregation → PII
- Risk with these information publicly available



52

OCEAN: Open-source Collation of eGovernment data And Networks

Student Name: Srishti Gupta
IIT-D-MTech-CS-IS-13-MT11012
November 20, 2013


Indraprastha Institute of Information Technology
New Delhi

Thesis Committee:
Dr. Ponnurangam Kumaraguru (Chair)
Dr. Yasuyuki Nishii
Dr. Mottakaloshi Rejzunas

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science,
with specialization in Information Security

©2013 IIT-D-MTech-CS-IS-13-MT11012
All rights reserved.

<https://arxiv.org/pdf/1312.2784.pdf>



So, you feel free to take a look at the master says itself I have later in the slides, the pointer, So, what is the takeaways for all that we saw in this week. Large amount of personal data is publicly available, which can be used, misused, will be great if some of you can explore the idea for the city that you are part of. Delhi is something that we did, I am sure many of you are from different cities.

See, I mean, what I would like you to do is just go out look at government services in your city, in your state, do some sort of say searching around to see whether these information are publicly available, whether you can find any services for driver's license all of that, for the city, I mean, I think doing it for your own city is very exciting, because you will get to know something about your own city.

Pulling all this information from publicly, you can actually create personally identifiable information and a profile of users. Super high risks at some for some, because many information is available about some where you can actually triangulate this information and get the user profile as, I said large number about 2000 users, all details were publicly available, So, that is the thesis I mentioned.

Take a look at it and if you have any questions I am happy to actually answer later. So, that is the end of week 8. I hope you got a good sense of the voter leak that we saw, the browser extension problem that we saw and now we saw about publicly available information from government websites how it can be actually used and misused.

Again please feel free to ask any questions on the mailing list. I see some questions coming but sometimes they are all administrative questions, it will be nice to have technical questions also on the mailing list and see you next week.