

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

Welcome back for week six lecture. I hope you are enjoying the class. I hope the topics that we are seeing in the classes are relevant to you and making some sense in your real world interactions itself, both online and offline, I think privacy as a topic is very, very connected to both the online and offline world also.

So, and thanks for being some, thanks for being active on the mailing list for some of you and thanks for asking questions. So, what we will see this week is we will actually look at these topics voter privacy leaks, we will look at browser privacy leaks and we will also look at, we have defined what is personally identifiable information before if you remember that.

What we are going to see is that can we actually use those personally identifiable information that is publicly available and can you actually misuse it for something. I mean, what are the ways to use it and probably there are ways to misuse it also can we actually do that. And I will show you some examples of how it can be actually used or misused.

You may be able to recollect some of the lectures earlier that we saw about social security number being predicted, how Facebook, I mean, how users can be re-identified on Facebook all that, So, that is the kind of things, but here we look at specific examples of publicly available information, particularly within India and I will show you some examples where we can actually use those information from publicly available data.

So, voter privacy leaks, browser privacy leaks, and profiling people, profiling users from publicly available information; that is the kind of content that we are going to cover today, this week.

(Refer Slide Time: 2:08)

NPTEL

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY  
HYDERABAD

### What we have covered until now

- What is Privacy?
- Why study Privacy?
- Fair Information Practices
- Right-To-Privacy
- Contextual Integrity
- Privacy Policy
- Privacy Enhancing Technologies
- Privacy Invasive Technologies
- Social Media Privacy
- Identity resolution
- Privacy nudges
- Cookies
- Ethics / IRB

- Why anonymize – AOL, Netflix
- Methods for anonymization
  - K-anonymity
  - L-diversity
  - T-closeness
  - Differential privacy
- Cost of Reading Privacy Policies
- Conducting (User, Lab, and Online) Studies
- Reading research papers

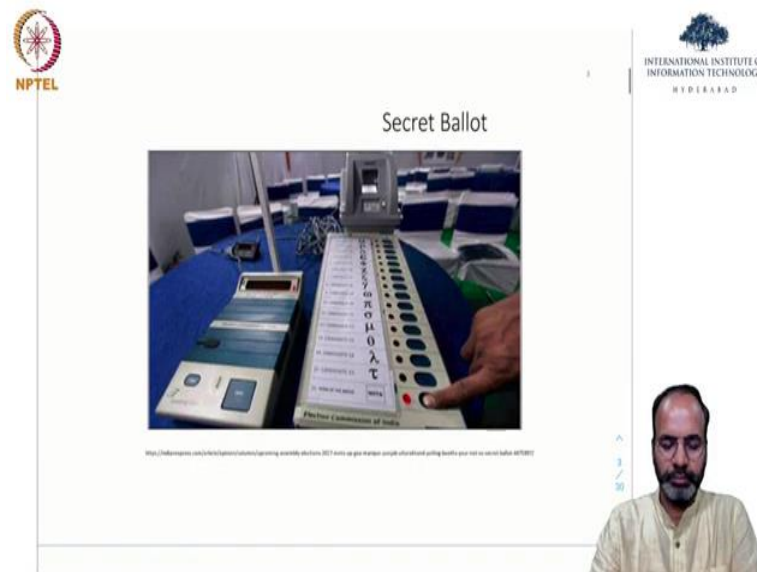
So, generally also to see what all we have covered until now. We have covered, in general, I think I am guessing that by now you should have a good handle on what privacy is, how privacy is different from So, to say security, what are the reasons why we should study privacy, fair information practices, right to privacy, contextual integrity, all of this I think I am guessing that you would be able to understand and sometimes even be able to talk about it in with others I think.

Privacy enhancing technologies, privacy invasive technology, social media privacy identity resolution, cookies, we will actually come back to cookies a little bit more also again later in the semester. Looking at some specific studies on how cookies have been used to study user behavior. We also talked about anonymity, I think we spent quite a bit of a time in anonymity, cost of reading privacy policies.

Last week we also saw conducting different types of studies, user studies, think aloud studies, survey, focus group discussions all those kind of methods by which you can actually collect data for studying privacy. I think it is also critical, I mean, I have been talking to you about projects I have not heard anybody send any email to me about projects, if you are thinking about it and if you have some cool ideas please feel free to drop a email and we can catch up later. Last week we also touched on how to read research papers.

I actually gave you an exercise for reading research papers, So, I hope you will be able to understand different roles that for reading privacy, for reading the papers.

(Refer Slide Time: 3:55)



What is secret ballot? I think most of you, I do not know whether most of you would have actually casted your vote in elections until now, um but if you are sitting in UP probably you may have casted your vote as we see this lecture. Secret ballot - Secret ballot is a concept where we, as in citizens, get control or protection from not sharing for whom we actually voted for that is I cannot force you to share who you voted for.

That is why if you see when we go to cast our vote there is always this box which nobody sees, if you are looking at the electronic voting machine as you are seeing in the slide, electronic voting machine if you are casting your vote there is always a box around it so, that nobody sees who you cast, who you are casting your vote for. And the secret ballot is a great protection for democracy also.

Now I am going to actually talk about how the secret ballot sometimes particularly in the context of citizens sharing their casting behavior, sharing their political affiliation on social media can actually be used for surveillance to understand how patterns of users casting out. Again secret ballot is an idea where we get protection for not being forced into revealing who we voted for. But I am going to show you some examples where we are sharing our political affiliations, our political, who we voted for users by themselves.

(Refer Slide Time: 5:40)

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 '#Elections2019' is centered at the top. Below the title, there is a bulleted list of four questions:

- Who all voted?
- Who all are 1<sup>st</sup> time voters?
- Who all posted on social media?
- Why did you post on social media?

A man with a beard and glasses, wearing a light-colored shirt, is visible in the bottom right corner of the slide frame, appearing to be the presenter.

It is a question for you, if you voted for, if you voted in the 2019 Elections, I mean, probably do a reaction on the mailing list, I am guessing that many of you will be a first time voter also first time voter now or first time motor in 2019 elections. I am sure some of you would have posted about your casting, that you casted your vote on social media. Even if you have not posted it I am pretty sure that you have seen others post on social media.

Why do we post? Why do we, I mean, if you just think about it, your friends may have posted it or your colleagues may have posted it, you may have seen your relatives sharing it on social saying that, “Oh! I inked my vote right.” People share for sharing that they are doing their duty for democracy. Many reasons why people, I mean, you would also see people taking selfies, pictures with people in the booth and with celebrities in the booth and sharing it saying that they are casting their vote and all of that.

(Refer Slide Time: 7:01)

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 'Voter Privacy Leaks' is centered at the top. Below the title, a bullet point asks, 'What all do you think can be inferred?'. A small navigation panel on the right side of the slide shows a list of slide numbers, with '5' highlighted. A video feed of a man in a light-colored shirt is visible in the bottom right corner of the slide area.

One question that I have for you is just think about it for a second before you move on to the slides later, think for whether what can be inferred, let us take, if you actually shared that you casted your vote and some information with that tweet or Insta post, what do you think can be inferred from that social media post?

(Refer Slide Time: 7:24)

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 'Voter Privacy Leaks' is centered at the top. Below the title, a bullet point states: 'Twitter users reveal their political inclinations, directly or indirectly, by tweeting the party or candidate they voted for'. Below this text, three small screenshots of tweets are shown, labeled (a), (b), and (c). Below the screenshots, a caption reads: 'Fig. 1. Users reveal their votes while posting tweets on Twitter. We have taken examples of the top three most notable political parties in India, viz., BJP, INC, and AAP.' A small navigation panel on the right side of the slide shows a list of slide numbers, with '6' highlighted. A video feed of a man in a light-colored shirt is visible in the bottom right corner of the slide area.

irectly, by tweeting the party or candidate they voted fo

NPTEL

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

(a) Example reveal of BJP (b) Example reveal of INC (c) Example reveal of AAP

Fig. 1. Users reveal their votes while posting tweets on Twitter. We have taken examples of the top three most notable political parties in India, viz., BJP, INC, and AAP.

NPTEL

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

(a) Example reveal of BJP (b) Example reveal of INC (c) Example reveal of AAP

Fig. 1. Users reveal their votes while posting tweets on Twitter. We have taken examples of the top three most notable political parties in India, viz., BJP, INC, and AAP.

Location  
Candidate Name  
PM  
Political Party

So, what we are going to look at is Twitter users, So, what we did was 2019 Elections we collected a lot of data in that, one question that we started asking is -- can we actually understand user preferences of who they voted for? Which is going back to my point of secret ballot, can you find out whether secret ballot is not, So, to say, users do not really think of secret ballot.

Even though they are casting their vote in a very closed environment they end up actually sharing who they casted to their vote for publicly. So, here are some examples, sometimes tweets talk directly about the voting preferences, sometimes indirectly. Here are some examples that you can see which is people revealing that which political party they casted their vote for.

Generally, their tweets would look like this, which is, ‘I voted for,’ there are some very common hashtags that we saw during the elections which people were using, it is later in the slides also you will see those hashtags. For example, hashtag I voted. So, I voted and then location, let us take Gachibowli, for example, your constituency that you will mention.

And then hashtag Hyderabad, hashtag Telangana and at the rate political party name, political party’s Twitter handle, at the rate the candidate that they voted for, and at the right Prime ministerial candidate. This is one extreme which is it is giving you all details, which is location, candidate’s name, prime ministerial candidate's name and political party. So, this if you have and that is the question I asked you to think about earlier.

Let us take you have access to this tweet what can you do with it? And then this is one extreme, but just think about it, even if you do not have all of this available, which is location, candidate name, prime ministerial candidate name, political party all of that, if I just know the candidate's name that he or she voted for, that they voted for it is easy to find out a lot of other information.

Because except for a few, there is a very unique connection between, relationship between candidate’s name, location, candidate to political party and candidate to the prime ministerial candidate. So, therefore, if you just even say that I voted for at the rate So, and so, that is good, you can actually infer a lot more information from just that tweet.

(Refer Slide Time: 10:17)

The slide is titled "Elections Data" and features the NPTEL logo on the left and the International Institute of Information Technology Hyderabad logo on the right. It contains two bullet points: "Politicians snapshot" and "Hashtags". Handwritten red annotations include a box around "#GE 2019", an arrow pointing to "#Elections 2019", another arrow pointing to "A/May 2019", a third arrow pointing to "Sep 2018", and three upward-pointing arrows at the bottom. A small video inset of a man is visible in the bottom right corner.

So, what all did we do in elections data? Elections data when we collected this 2019 Elections data, we actually started, the elections was in 20 2019 April-May, we started collecting data in September 2018. And what all did we collect? We collected politicians snapshot. Politician snapshot is basically taking verified handle because we took only verified annual because it is the confirmed way by saying that whoever is, whoever we are track, whoever we are collecting data is actually the person who he or she claims to be.

So, what did we do? We took, let us take politician A, we look at their followers, following tweets, likes, everything that you can actually get from a profile of twitter API, we take that. Let us take today at 12 in the midnight, tomorrow again 12 in the midnight we do exactly the same thing. What does this allow? This allows us to know that what change happened in these 24 hours.

This is very useful, because now I know in this last one day the users had, the politician had increase in number of followers, increase in number of posts, increase in the number of likes that they got, all of this can be actually analyzed if you have the snapshot. Whereas, now if you go collect data for politician A right now, you are going to get only the snapshot what is right now in twitter.

So, that is why we did the snapshot from September 2018, 20 September month till about April, May so, we get the snapshot of politicians for every day. I am sure you can already think of what all super exciting questions that you can ask. Wait for a little while the slides has pointers to all this data. You can you can probably take it and use it for doing some analysis also.

The other one which is where the large amount of data that we collected was the politician's snapshot is not really a large amount of data right in that sense. Hashtags is the one that we did, which is for example, General Elections 2019 was the official hashtag for the elections, So, we start from there. We then expand the hashtag as we look at the other tweets with the General Elections hashtags.

So, query expansion, General Elections, I mean, take a tweet of, let us take thousand tweets with General Elections, look at what are the other hashtags, if they are relevant to elections start collecting data for those hashtags. For example, from General Elections 2019 we can actually go to hashtag Elections 2019. It could be political party name and 2019.



You expand these hashtags and start collecting data for all the hashtags. We did this also there were like thousands of hashtags that we actually collected data for and again those data is also made public.

(Refer Slide Time: 13:35)

**Datasets**

- > Lok Sabha Elections: Ministers Social Media Report Card Data [README.md](#) | [Download](#)  
> MD5: 8c1afc0bcd393a5ce61e436ae72da746  
> SHA1: 956c9b722c12eaceaac2e160c56701793828f187
- > Lok Sabha Elections: Delhi Candidates' Social Media Details [README.md](#) | [Download](#)  
> MD5: 53c70b2f5305bb7080b12256362a0b5  
> SHA1: a2f574736b0413bca643bca4b4f510714402fc6b
- > Lok Sabha Elections: Phase One Candidates' Social Media Details [Download](#)  
> MD5: d3a9be45738eecc1d622c3d965796e51  
> SHA1: 89b74a4fc2e37ae65a62b39a57ae05836d4b637
- > Analysis of General Elections 2019 in India. [Download](#)  
> MD5: 9454e1b94765ff3b5efe29d9bbabdb13  
> SHA1: 5ca66aa3037f0f02891149303a7f644ae3438880

<https://precog.iit.ac.in/resources.html>

You could actually, So, yeah, here is a pointer for the data sets, you could actually take all those, So, I just put the screenshot for all the Lok Sabha Elections data. You could take all the data from this website and play around with it.

(Refer Slide Time: 13:50)

**NPTEL**

**Chowkidar Amit Shah**  
Member of Bharatiya Janata Party / President, Super Cluster, Maharashtra

**Chowkidar Arun Jaitley**  
Member of Finance and Corporate Affairs, Government of India

**Chowkidar Jyoti Z Irani**  
Chief Minister of Madhya Pradesh, Government of Madhya Pradesh / Minister, Panchayati Raj, Government of Madhya Pradesh

**Chowkidar Jayesh Mehta**  
Member of Bharatiya Janata Party / Minister, Panchayati Raj, Government of Madhya Pradesh

**Chowkidar ramakrishna varma**  
Member of Bharatiya Janata Party / Minister, Panchayati Raj, Government of Madhya Pradesh

**Chowkidar Haritha**  
Member of Bharatiya Janata Party / Minister, Panchayati Raj, Government of Madhya Pradesh

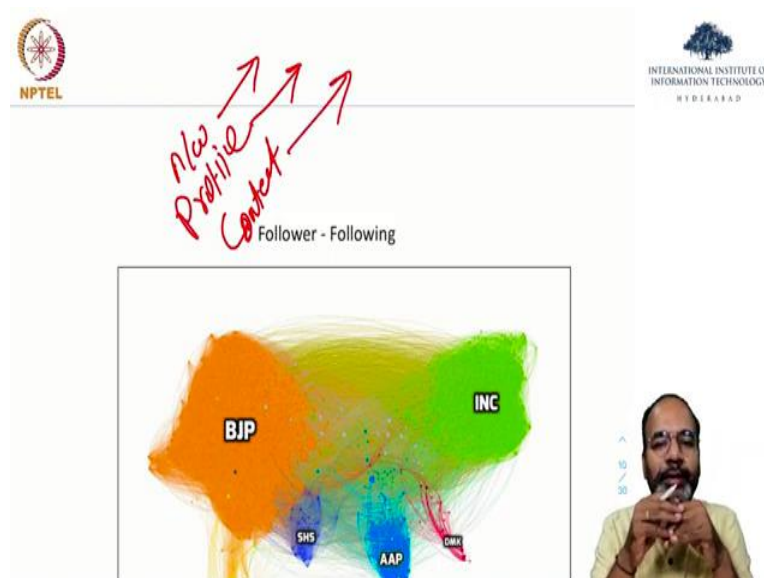
During elections many, many things happen, particularly during elections many, many online things started happening. This 2019 Elections was phenomenally influenced by, I mean,

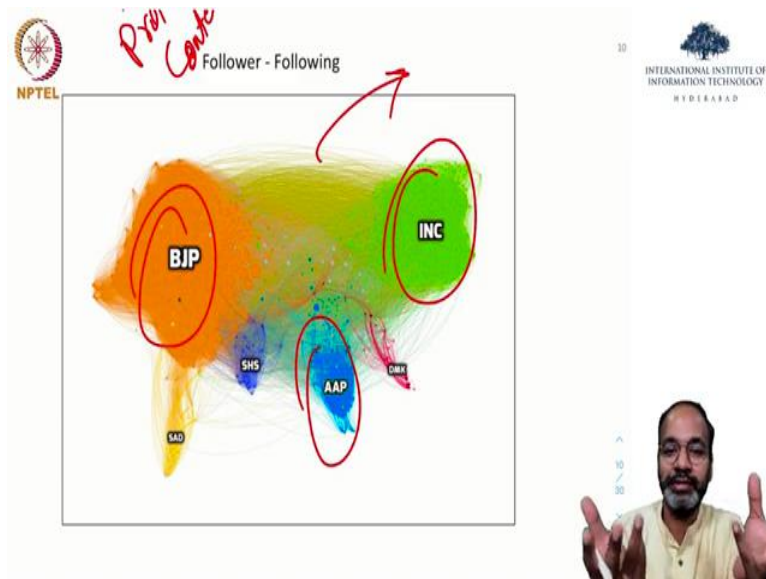
activity online was very high. And I am guessing that 2024 Elections is going to be even more online driven, let us see how that happens. So, this slide is actually showing you the details of how username changed during the elections.

So, if you see Chaukidar started, I am sure you will remember hashtag Mebhichaukidar was trending and was getting popular all of that. And this is the, the one on the top is politicians, the one on the top is general twitter users, who are also changing their names with the Chaukidar. And this is interesting, I mean, this is interesting from sort of say how campaigns work, how effective this campaign was, all that.

But let us stick to the privacy questions also. We should, we could actually use this data to identify users and also see how these people participated in campaigns, what campaigns did they, which side of the campaign were there, all of that we could actually infer from this. So, we collected this data.

(Refer Slide Time: 15:11)

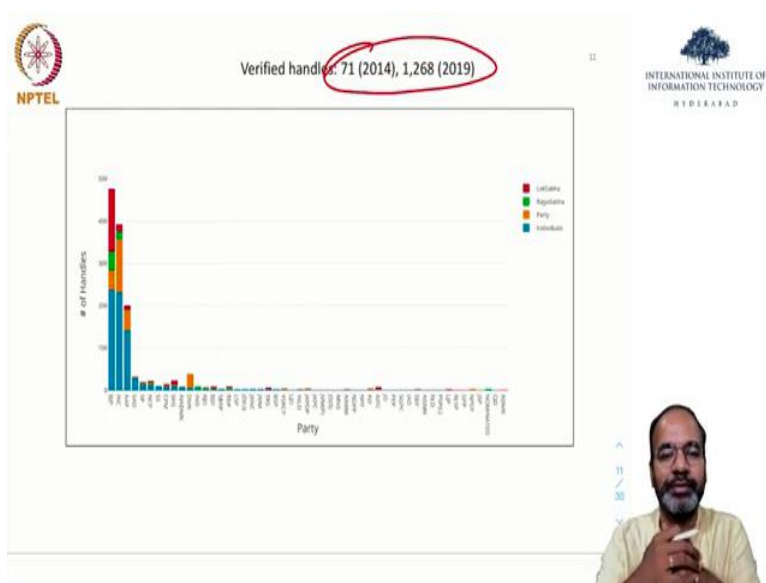




Another analysis, another common thing that you will see, So, generally in social media you are going to see network, profile, content. Network is what the slide is, profile is what the last slide that I showed you, handle everything, content is just the post itself. So, this is and the slide that I have now is actually showing you the interactions between follower-following relationship between the handles that are affiliated with the political party.

So, BJP, INC, So, different colors represent different political party, and the interactions; every edge is either a following or a follower. Again this kind of information can be very, very useful in inferring, what the interactions are, how dense the interactions are, all of that.

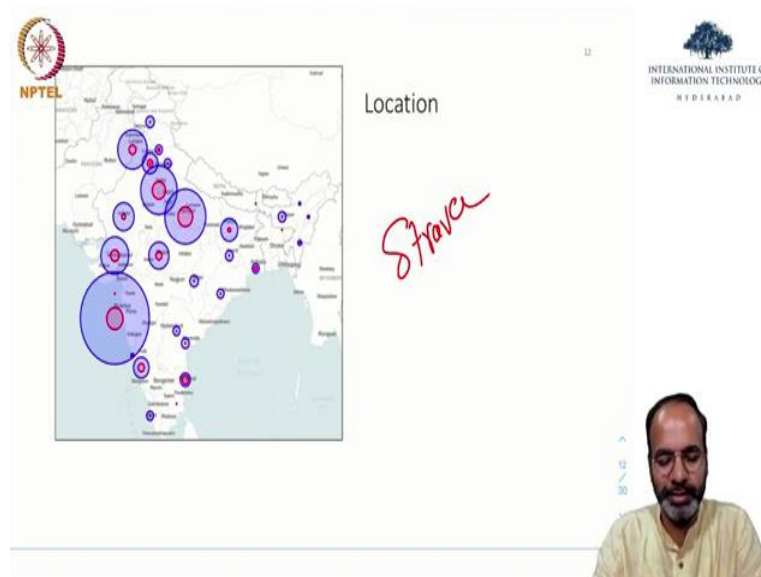
(Refer Slide Time: 16:08)



I said few minutes back the 2019 Elections was very much online, here is an example; here are some statistics to emphasize that point. X-axis is the political party, y-axis is number of handles. And particularly take a look at this, in 2014 only 71 handles that were involved in elections were verified, whereas in 2019 Elections it is 1,268 handles. Just just look at the magnitude. And again these, let us see what's happening in 2024.

But this can be very, very useful, because now I know PK's verified handle is actually the PK professor at IIIT, Hyderabad, whereas if you see other sort of say Ponguru or saying that professor at PK, professor at IIIT, Hyderabad you cannot believe, whether this is actually the PK who is teaching this online privacy class.

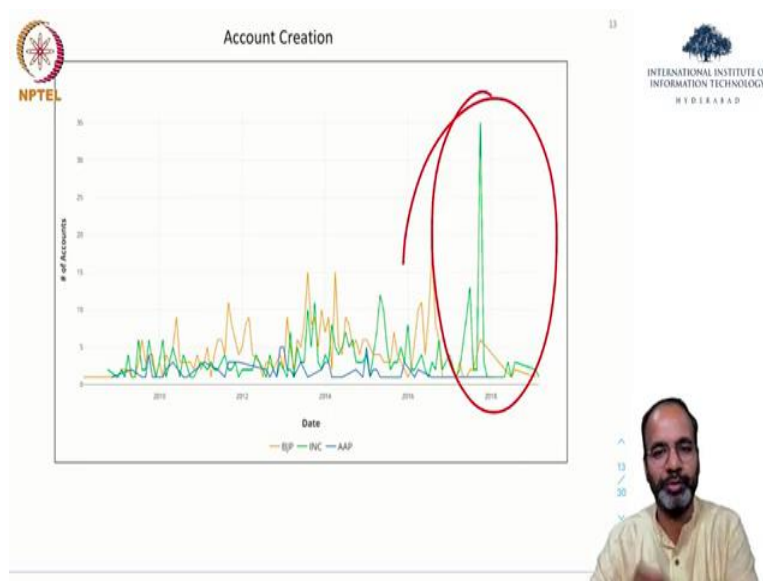
(Refer Slide Time: 17:08)



Location information – Location information again, we will back to this a little later in the semester, looking at how location information can be inferred from publicly available information. Remind me if I do not talk about it. There is a good example for Strava social media where the location. Strava is a health social media, it is a social media platform used by people who run bike all of that.

And we will talk about how you can actually use Strava to identify some information. This slide is actually giving you tweets with the location. Think about how this can be used and misused for the thing that you want about the users involved in elections, the larger the circle the large number of posts, all of that.

(Refer Slide Time: 17:58)



This one, So, I am going through some of this in slightly detailed because you need to understand the kind of data that we can actually collect from social and then use it for the privacy question that you are asking. So, this slide is actually showing you account creation. Why this slide is interesting because if you just see some at some days, some weeks, there is like spike in the accounts that are getting generated on these platforms. This is comparing different political parties.

(Refer Slide Time: 18:36)

**Is change the only constant? Profile change perspective on #LokSabhaElections2019**

**Authors:**  
Komari Neha (IIT Delhi), Shashank Srikanth (IIT Hyderabad), Sonali Singhal (IIT Delhi)  
Shwetanshu Singh (IIT Delhi), Arun Rajaji Baduru (IIT Delhi), Pommarangam Kumaraguru (IIT Delhi)

**ABSTRACT:**  
Users on Twitter are identified with the help of their profile attributes that consists of a username, display name, profile image, to name a few. The profile attributes that users adopt can reflect their interests, belief, or thematic inclinations. Literature has proposed the implications and significance of profile attribute change for a random population of users. However, the use of profile attributes for endorsements and to start a movement have been under-explored. In this work, we consider #LokSabhaElections2019 as a movement and perform a large scale study of the profile of users who actively made changes to profile attributes centered around #LokSabhaElections2019. We collect the profile metadata for 49,451 users for a period of 12 months from April 5, 2019 to June 5, 2019 and #LokSabhaElections2019. We investigate how the profile changes vary for the influential leaders and their followers over the social movement. We further differentiate the organic and inorganic ways to share the

**URL:** [http://precog.iitd.edu.in/pubs/Username\\_Change-Elections2019.pdf](http://precog.iitd.edu.in/pubs/Username_Change-Elections2019.pdf)

*Handwritten notes:* PK. Prateek, Pangur, P. - K. K.

I showed you earlier, So, actually we have even seen the profile change sometime back in the semester which is when we saw the identity resolution content, where you want to infer PK

dot Profgiri and Ponguru and Ponnurangam Kumaraguru are all same. This question if you remember we saw like I think like third week or fourth week.

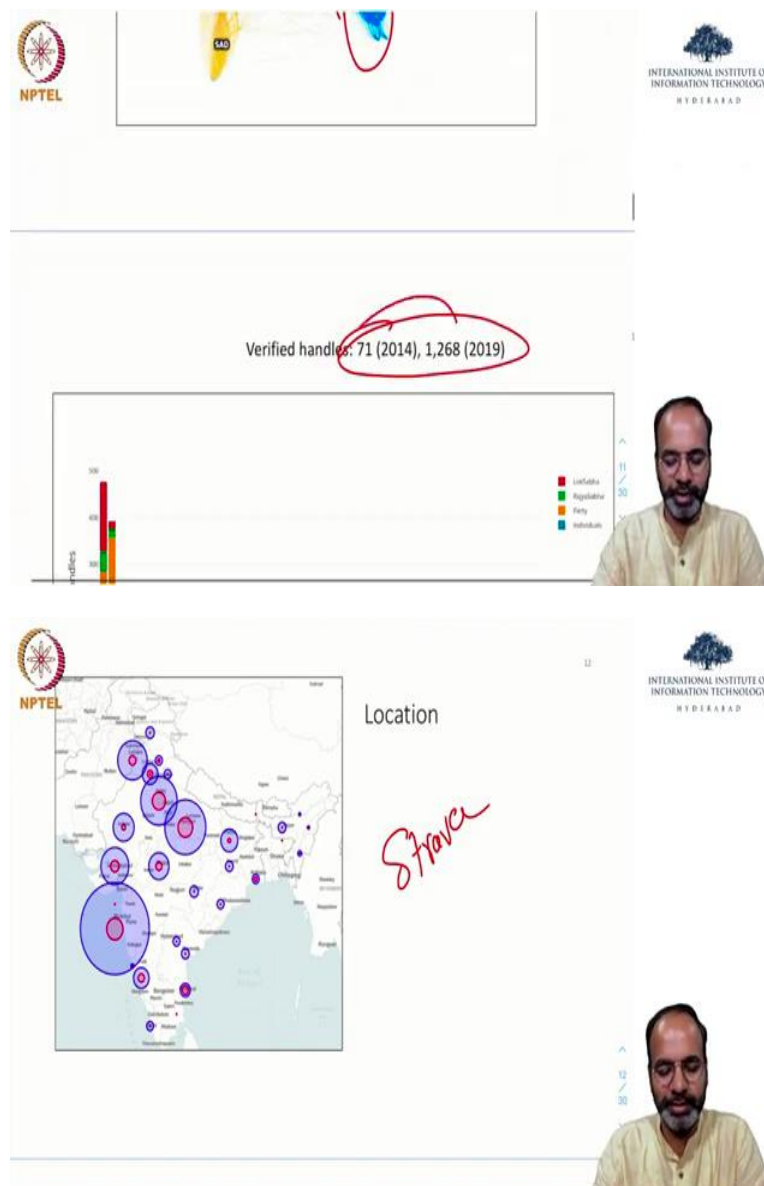
While seeing that question we saw important result which showed that if you, not necessarily that I should do the comparison only now which is what are my handles right now, if you know the handles that I had in the last 5 years or 10 years or some delta time, the comparison, the resolution of the users can be much more effective.

So, keeping that in mind we also looked at, is it the same elections, people change, the maybe Chaukidar, the thing that I showed you, how many people change, what changes it had, what was the influence about it all that is what was discussed in this paper.

(Refer Slide Time: 19:42)

One of the other thing that we ended up doing was that I had done some work, we had a master student who looked at 2014 Elections and we had built some dashboards for collecting social media data, all that. 2014 Elections we took and then we compared the data that we collected in 2019 and saw what was the differences.

(Refer Slide Time: 20:00)



One comparison that I just showed you was this number of verified handles. In the similar way you could actually do a lot more other comparisons, which political party was more active, how many, has the strategy over the years changed, has the manifesto over the years changed, has the online presence made any difference in the political party internally.

Can you care, any I think cool things would be, can you find out the organization structure using social media data itself, and or, which is the structure, is the real world structure, same has in the online structure. There is this hierarchy of people involved in elections or people involved in politics, can you actually recreate that structure using social data.

(Refer Slide Time: 20:47)

The slide is titled "Hashtag analysis" and features the NPTEL logo on the left and the International Institute of Information Technology Hyderabad logo on the right. The central content is a screenshot of a research paper titled "Hashtags are (not) judgemental: The untold story of Lok Sabha elections 2019". The authors listed are Sourabh Gupta, Arun Rajaji Bhaduri, Amit Kumar Singh, and Purnamanganam Kumaraguru. The abstract discusses the use of hashtags on Twitter to spread information and opinions during the 2019 Lok Sabha elections in India. The speaker, a man with a beard wearing a light-colored shirt, is visible in the bottom right corner of the slide.

And as I said before we collected hashtags also. So, we actually very, very closely analyzed hashtags from the elections also.

(Refer Slide Time: 21:00)

The slide is titled "Filtering process" and features the NPTEL logo on the left and the International Institute of Information Technology Hyderabad logo on the right. The central content is a list of data points related to the filtering process, with some items circled in red. The list includes: "Data that I described earlier", "Efforts like #KBPM-Selfie, by media house to share voted pics", "#gotinked #voted #firstvote", and "User handles mention", which is further broken down into "Candidate" and "Party". The speaker, a man with a beard wearing a light-colored shirt, is visible in the bottom right corner of the slide.

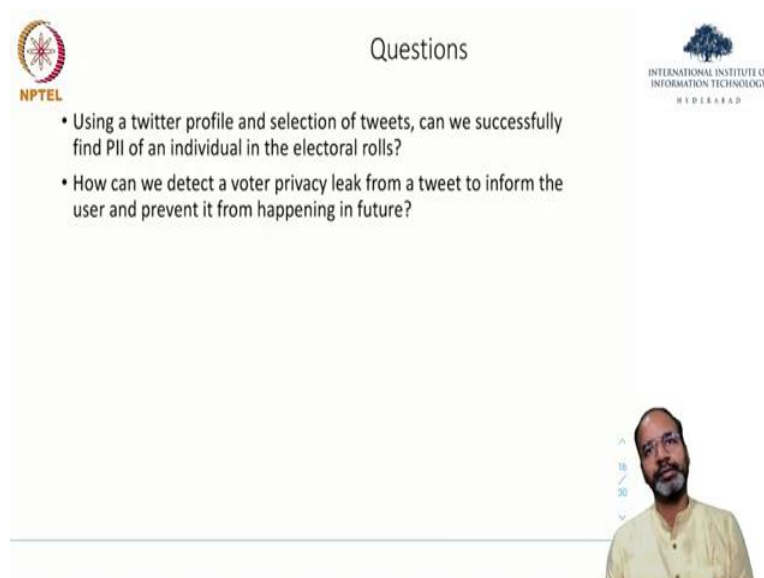
You can take a look at this paper again. So, what we did? Coming to the privacy question given that you understand this background of what all data is coming, what privacy, we are talking about secret ballot, we are talking about voter surveillance, all that. So, with this data they were, I mean, why there was more information on social about elections, because there was also the other side.



The media houses also gave incentive for people to post and put some hashtags with it, So, their tweets are getting posted by these media houses all of that, So, this is one example, but there were many other examples that were going on at that time. So, these were the hashtags that i mentioned earlier which is 'gotinked', 'ivoted', 'firstvote'. These are the different hashtags that were there for elections when people shared about them casting the vote.

And then I already talked about user handles mentions candidate and party, different ways by which you can actually collect what they are talking about, political party, location, constituency, all that. So, with all that if you filter the content with political party, with the location, with constituency, with the candidate with the prime ministerial candidate all that, you will get some content.

(Refer Slide Time: 22:13)



The slide is titled "Questions" and contains two bullet points:

- Using a twitter profile and selection of tweets, can we successfully find PII of an individual in the electoral rolls?
- How can we detect a voter privacy leak from a tweet to inform the user and prevent it from happening in future?

The slide also features the NPTEL logo on the left and the International Institute of Information Technology Hyderabad logo on the right. A small video inset in the bottom right corner shows a man in a yellow shirt.

Now, let us look at that. So, the two questions that we were interested in at that point in time was can we use twitter profile and some tweets that they have posted, can we successfully find PII of an individual in the electoral roles. Just, you just went and shared hashtag 'ivoted' at the rate the candidate. Is that enough to find more personal information about you?

Second, how can we detect the voter privacy leak from a tweet to inform the user and prevent it from happening in the future? So, what does this mean? So, we also saw 'nudges', in identity resolution content we also saw 'nudges' where you can actually build browser extensions to tell, give the user saying do not do this, even in the privacy nudge that we saw, timer nudge, So, picture nudge or the sentiment nudge, we saw all of this.

So, those things, similar method, can you build nudges, can you build interventions, So, that when somebody is posting like this you stop them. You nudge them saying what is possible, what could happen because they are actually sharing this information online. So, those were the two questions that we were interested in.

(Refer Slide Time: 23:31)

The slide features the NPTEL logo on the left and the International Institute of Information Technology Hyderabad logo on the right. The central content is a screenshot of a Twitter post and a voter registration form. The Twitter post, from user NaXXXa.XXXe, mentions a political party. The voter registration form shows fields for Name, Address, and Voter ID. Red circles and arrows are drawn over the image to show how the information in the tweet is linked to the voter registration data. A man is visible in the bottom right corner of the slide.

**Linking, Leak**

**Fig. 3.** An example of cross-linking information. A Twitter user NaXXXa.XXXe tweeted about their vote, hence, revealing their preference toward a party and losing their voting privacy. The Twitter display name is successfully linked to their entry in the electoral rolls. We have censored their voter ID, names, and house number from the results.

So, here is what happened, So, with all that data we could actually go to take the user and actually go and figure out people driver's license number connected to that, connected to the polling data that is publicly available which will also have the unique address, all of that, and connected to some driver license information. I will show you later in this lecture, in this week itself, what information, in this week or probably next week.

Look at how you can actually use publicly available information from government services only to actually put things together, some work that was done to collect information that is publicly available on government services and can you put them all together to create a profile and to use or misuse the same information.

So, you will see that a tweet about a political party affiliation for casting the vote can actually be linked to your driver's license, to your address, all that. This is interesting but this is also a little dangerous, just imagine this morning you went and casted your vote and then a couple of days later somebody knocks at the door asking something about the affiliation political party that you voted for.

(Refer Slide Time: 24:52)

NPTEL

# Intervention / Nudge

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

(a) VPL (b) Non-VPL

Fig. 4. (Left) The extension creates a red background and shows a nudge message below while drafting the tweet if the classifier reports a leak. In this case, the user has revealed that his vote is for a specific candidate. (Right) The extension creates a green background in this case. This is because the tweet doesn't reveal any sensitive information and is safe to be posted.

20 / 30

So, this is the intervention that you can actually build, which is give them a nudge saying look the information that you are sharing with this tweet can be used for identifying some personal information about you, the affiliation and everything. This intervention can be actually done for many other topics also. I am just using elections and privacy as one example, but I am sure you can think of many other contexts where these kind of nudging can be very useful.

(Refer Slide Time: 25:27)

NPTEL

# #IVoted to #IGotPwned

Studying Voter Privacy Leaks in Indian Lok Sabha Elections on Twitter

INDO

social info

21

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

**WHAT IS A Voter Privacy Leak?**  
One of the leading principles in elections is to ensure that the parties or candidates to which a voter casts a vote, reach their political goals through an election of their own merit and qualifications. We respect such issues as voter privacy leaks.

**THE PROBLEM**  
A user's VPL can be cross-linked with electoral rolls to get an idea of eligible voters available publicly to reveal their personal identifiable information. The user did not even realize their right to voter privacy but is also vulnerable to identity and personal details leaks.

**THE SOLUTION**  
To safeguard users who post VPLs, we created a nudge using browser plug-in that averages machine learning-based classifiers to flag a post and on Twitter as a VPL or a Non-VPL, thereby helping users protect their voter privacy.

**DATA**

- 1. User Information from Twitter
- 2. Electoral Rolls from Election Commission of India

**Tweets classified as Voter Privacy Leaks (VPLs) can be cross-linked with Electoral Rolls to reveal personally identifiable information like voter ID, Family details, Age, Gender and Address, consequently posing a serious privacy concern.**

**CLASSIFICATION**  
As a first towards educating users and providing them a better VPL, we built a simple classification model that identifies if a given tweet contains VPL of the user. Machine Learning classifier with cross-sections gives best results.

**NUDGE**  
We designed a notification in Twitter's web client which is browser extension that shows short warnings to nudge users to consider the content of their tweet and make it a "Non-VPL" tweet.

Example: A Twitter user @XXXXX tweeted a VPL. The Twitter display name is successfully linked to their entry in the electoral rolls.

21 / 30

If you are interested more please feel free to take a look at this paper, it goes into details of more examples, statistics and then more technical details of how we ended up actually doing this analysis.

(Refer Slide Time: 25:41)

**Trends in voter surveillance in Western societies: privacy intrusions and democratic implications**  
C.J. Bennett · 2015 · [dspace.library.uvic.ca](#)  
... This paper surveys the various **voter surveillance** practices ... move from **voter** management databases to integrated **voter** ... between concerns for excessive **surveillance**, and the broad ...  
☆ Save ☆ Cite Cited by 56 Related articles All 7 versions 56

**The politics of privacy and the privacy of politics: Parties, elections and voter surveillance in Western democracies**  
C. Bennett · ... and **Voter Surveillance** in Western Democracies (June ... 2013 · [papers.ssrn.com](#)  
... This paper surveys the various **voter surveillance** practices ... Five interrelated techniques are analyzed: the development of **voter** ... **Voter surveillance** requires further comparative analysis, ...  
☆ Save ☆ Cite Cited by 26 Related articles All 6 versions

**The old in the new: Voter surveillance in political clientelism and datafied campaigning**  
I. Kluiche · *Big Data & Society*, 2020 · [journals.sagepub.com](#)  
This article compares political clientelism and datafied campaigning as two modes of relating politicians/parties and voters that are centered around **voter surveillance**. It contributes to the ...  
☆ Save ☆ Cite Cited by 2 Related articles All 4 versions

**Privacy, voter surveillance and democratic engagement: challenges for data protection authorities**  
C. Bennett, S. Qureshi · *International Conference of Data ...* 2019 · [papers.ssrn.com](#)  
... Without a high level of transparency – and therefore trust amongst citizens that their data is being used appropriately – we are at risk of developing a system of **voter surveillance** by ...  
☆ Save ☆ Cite Cited by 7 Related articles

V  
o  
t  
e  
r  
S  
u  
r  
v  
e  
i  
l  
l  
a  
n  
c  
e

22 / 30

So, I also looked at voter surveillance, So, broadly this topic is voter surveillance, which is which is this work would fit in. And voter surveillance seems to be also very popular word that people are working on, So, please take a look at, if you are interested in figuring out using this publicly available information whether you can get political affiliation of people I think you should spend some time on these papers.

(Refer Slide Time: 26:09)

**Browser Extension**

From Wikipedia, the free encyclopedia

A **browser extension** is a small software module for customizing a web browser. Browsers typically allow a variety of extensions, including user interface modifications, cookie management, ad blocking, and the custom scripting and styling of web pages.<sup>[1]</sup>

**Contents** (from)

1. Plugins
2. History
3. Ad-blockers
4. Unwanted behavior
5. References
6. External links

**Plug-ins** (edit)

Browser plug-ins are a separate type of module. The main difference is that extensions are usually just source code, but plug-ins are always executables (i.e. shared objects). As of 2021, plug-ins have been deprecated by most browsers, while extensions are widely used. The most popular browser, Google Chrome,<sup>[2]</sup> has over 100,000 extensions available but no longer supports plug-ins.<sup>[3]</sup>

**History** (edit)

Internet Explorer was the first major browser to support extensions, with the release of version 4 in 1998.<sup>[4]</sup> Firefox has supported extensions since its launch in 2004. Opera began supporting extensions in 2008, and both Google Chrome and Safari did so the following year. Microsoft Edge added extension support in 2019.<sup>[5]</sup>

Brow  
se  
r  
E  
x  
t  
e  
n  
s  
i  
o  
n

23 / 30

So, the next one, So, now publicly available information we could actually get with, within the context of elections. Now, let us look at browser extensions. This browser extension, I mean, how many of you use, I am sure many of us use browser extension; many of you use browser extensions. There are many reasons why we use browser extension. It just makes our life simple, some nudges that we can get from it. Some statistics that we can get from it, quotes that we can get from it, I mean. I think people use browser extension for many, many reasons.

(Refer Slide Time: 26:41)



Something that we got very excited about at some point in time was looking at browser extensions from a point of view of how it is revealing information about us. So, what we did? So, this is the slide which says browser extensions are spying on you, 218 Spying Chrome Extensions installed by 2.4 million users were discovered out of the 43,000 Chrome Web Extensions.

So, Chrome Extension, Google Chrome Extension at that point in time had 43,000 tweets or 43,000 extensions and we downloaded all the 43,000 extensions. We kind of created a sandbox in which we could see how, what these extensions are doing. And in the process we actually learned how these extensions were sending out requests, what information are they collecting, for example, some calendar extension asking for the location, lat-long of where the phone is all of that.

Why do we need a calendar app to know where I am, location? I mean, you can make an argument that, oh, look it will do the time zone all that but how many of us actually need

those kind of features, automatically detecting where my location is and changing the time zone of my watch, my laptop, everything accordingly, not many of us would need those kind of features.

So, this one, looking at the extensions we wanted to understand whether these extensions are spying, whether they are spying or sending out some personal information about me. How much are they doing, what are they doing, can we actually find it, can we quantify it, those are the questions that you can actually think about.

So, web of trust, web of trust is a service where you can send a request for a domain and it will actually give you some details about the domain which is how credible the domain is, what is the, how many people have said positive about this domain, such kind of information comes out, it is an API request again, triple it dot ac dot in, returns with some scores. You can decide how to use this course in designing what you want to do with it.

Spying extensions can steal browsing history. Next slide also we will show you some examples, browsing history, IP address, geolocation, online social media access tokens and domains visited. These browser extensions, you may be using an extension which is, let us take one of the popular extensions that we saw when during our analysis was this hug, virtual hug extension.

So, that was very popular and just imagine if you understand that this extension is actually taking your Facebook and Twitter username password and sharing it or it is looking at all the websites that you are looking at, browsing history, and sharing it to a third party.

(Refer Slide Time: 29:45)

The slide features the NPTEL logo on the left and the International Institute of Information Technology Hyderabad logo on the right. The main content is divided into three horizontal sections: red, green, and yellow. The red section at the top states '218 SPYING CHROME EXTENSIONS' and notes that '2.4 million users were discovered to use 11.2% Chrome Web Store Extensions'. The green section highlights a 'Web Of Trust' extension with 140 million users that was found to be spying and has since been removed by Firefox. The yellow section lists the types of data these extensions can access: browsing history, IP address, geolocation, online social media (OSM) access tokens, and domains visited. A small inset image shows a person in a video call.

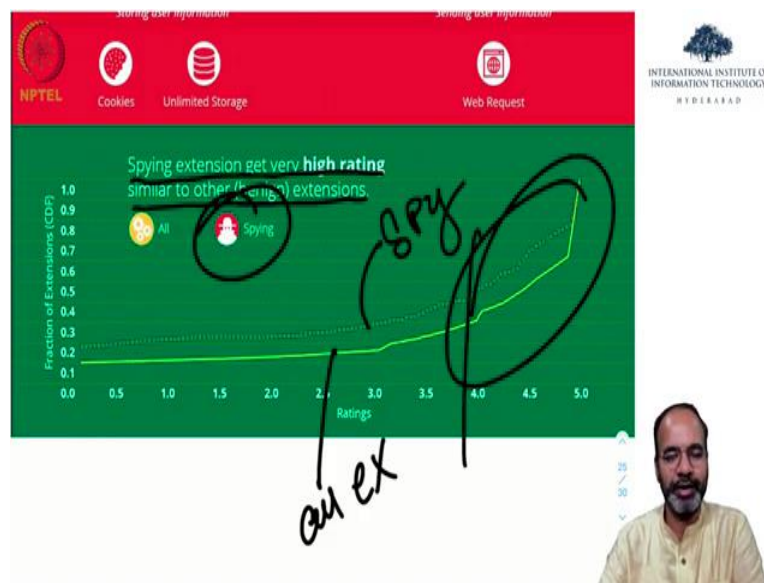
So, what we saw was browsing history, leakage of behavioral patterns like private documents, on services like Google docs, everything, those links can go away, geolocation, which is your lat-long, your home location, country location that can actually go away, or some access tokens, username password, API keys, those are also vulnerable.

(Refer Slide Time: 30:09)

This slide also features the NPTEL and IIT Hyderabad logos. It is titled 'Permissions which enable spying extensions to access, store and send user information'. It is divided into three categories of permissions: 'Accessing user information' (including Tabs, Cookies, Storage, All URLs, History, Geolocation, and Active Tabs), 'Storing user information' (Cookies and Unlimited Storage), and 'Sending user information' (Web Request). At the bottom, a line graph compares the 'Spying extension' (red line) with 'All' other extensions (green line) based on a 'Spies of Extensions (SDP)' rating. The graph shows that the spying extension has a significantly higher rating, indicating it is more malicious. A small inset image shows a person in a video call.

And some of the information that the extensions have access to is cookies, history, I mean, with the strength of these browser extensions they can actually get access to as much as information from your machine or from your browser.

(Refer Slide Time: 30:27)



This graph also was interesting because you will see that the spying extensions which are actually in red color, so, this spying extension gets very high ratings similar to other benign extensions, these are getting compared, the yellow one is all extensions, the red one is the spying one. And if you see, if you just take a look at it, there are extensions, let us take 4 ways to be called as oh, very useful, helpful.

I would use type of extension, there are like so, much of extensions which are rated more than four, which are also spying, it will be interesting to see who are these people who are rating it, what kind of collusion is happening there, but for now think about how many people are actually, how much of these extensions are getting rated, even though, they are rated high even though they are actually spying.



(Refer Slide Time: 31:24)

**Overall conclusions**

- Many spying extensions are reported. Only 12 out of 218 extensions received negative comments about spying happening on Chrome Web Store.
- Top developers also publish spying extensions. 2 of the top 10 developers (by volume of extensions) have published spying extensions.
- Using a neural network model based on permissions asked by extensions, client side behavior and meta information of extensions, we achieve a precision of 86% to effectively detect spying extensions.

NPTEL INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

So, these were some broader conclusions that we had which is 12 out of 218 extension received negative comments, 2 out of top 10 developers of browser extensions were actually part of this spying extensions. Yeah, So, you can also apply some machine learning model to understand the extensions, understand the user behavior all that.

(Refer Slide Time: 31:52)

**ABSTRACT**

Several studies have been conducted on understanding third-party user tracking on the web. However, web trackers can only track users on sites where they are embedded by the publisher, thus obtaining a fragmented view of a user's online footprint. In this work, we investigate a different form of user tracking, where extensions installed on a browser can capture the complete browsing behavior of a user and communicate the collected sensitive information to a remote server (trusted by the user). We conduct the first large-scale empirical study of 218 spying browser extensions on the Chrome Web Store. We observe that these extensions track a variety of sensitive user information, such as the complete browsing history of the users (e.g., the sequence of web resources), online social network, cookies, tokens, and significantly different type of user tracking—user spying (or spying for short) by browser extensions. Spying violates user privacy by collecting sensitive personal information without consensus from the user. This information can later be used towards targeted attacks, or can be traded in underground markets. Unfortunately, user tracking, which started as a means to enable better personalized advertising, has reached a state where even privileged software running on a user's browser is capturing sensitive personal information and sending it to remote servers (trusted by the user).

But one might still wonder why spying by browser extensions deserves special attention, when studies have mainly focused on third-party tracking on the web. First, it should be noted that today, the browser is the most widely used software and serves as a window to the web. In fact, plat-

[http://precog.iitd.edu.in/Publications\\_files/aa-spyingextensions.pdf](http://precog.iitd.edu.in/Publications_files/aa-spyingextensions.pdf)

NPTEL INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

So, if you are interested that is the paper, this paper got some super interesting attention when we did this work.

(Refer Slide Time: 32:00)

NPTEL

PII

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY  
HYDERABAD

- Personal information publicly available
- Can be used / mis-used

Information in the chart:  
Age, Gender, Source station, Destination station, Seat number, First Name, Last Name, Passenger Name, Record Number

So, the next part what I want to talk about is using publicly available information. Publicly available information is something that we already saw. Personally, personal information publicly available, I just showed you some examples of voter, in the context of voter surveillance, in the context of elections, all that. I am guessing that many of you would have taken a long-distance train.

I am going to use the train example to show you how we can actually use publicly available information to have a conversation with people when you are getting on board. For example, the train chart generally has information like age, gender, source, destination, seat number, first name, last name and a PNR number.

If you were somebody who is getting onto a train and seeing that I am on the train also I am sure you can start a conversation there. How can you actually get information? So, in this case probably if you are, you have taken this class So, there is already a connection, So, you probably would, hopefully, visually, at least, recognize me and therefore, we can have a conversation.

But imagine if you if you have somebody whom you do not know who is in the train, but you want to strike a conversation, is that possible, can you do it, what level of information can you actually get, that is what we are going to look at.