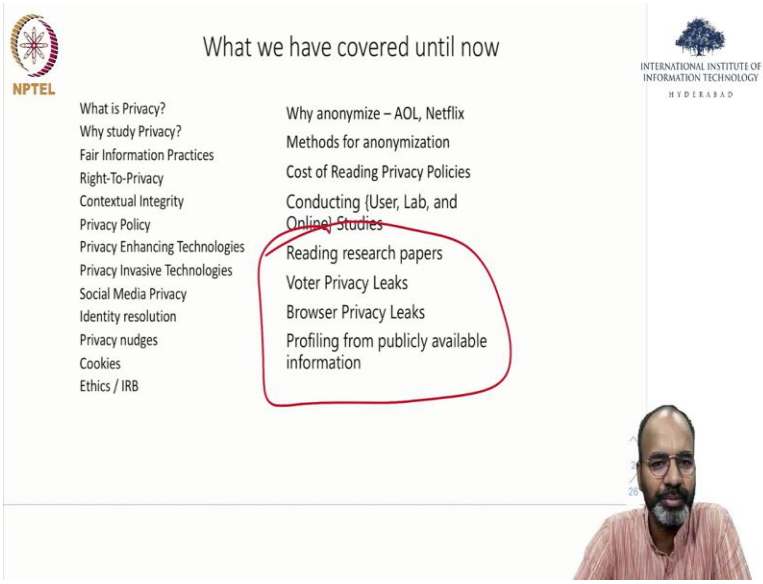
Online Privacy
Professor Ponnurangam Kumaraguru
International Institute of Information Technology, Hyderabad
Mobile numbers Home location LBSN

Welcome back to week 10. I hope you are enjoying the class and it is also nice to see some of you participate in the mailing list. But as always, I think, be more active, ask more questions, answer more questions. It will be fun. What we plan to cover in week 10, is how we can actually use mobile numbers that are online and what are the things that you can actually derive from data that is available on internet on social media for mobile numbers.

And then we will actually look at this one, I am guessing you would have you know, what LBSN is we actually did the location based social network in the first or the second week, what is location based social network, foursquare all of these connections you will probably remember. So, we will actually look at what one aspect of location based social network is, which can be used to find out where your home is, which city you working from? Where you actually go very frequently, all that.

(Refer Slide Time: 1:43)

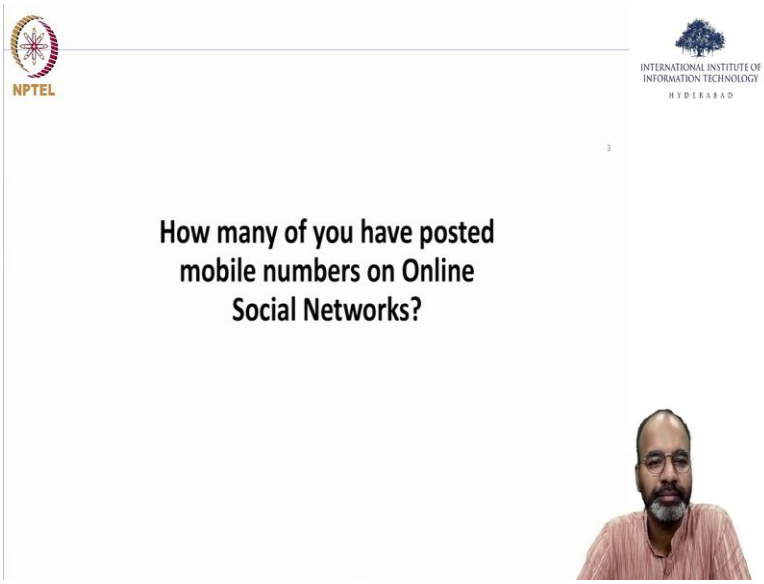


So, that is LBSN what we covered until now? So, again, every week, it is nice to see the topics that are getting added here. For last week, this these are the things that we added. But broadly, we have been talking about what is privacy, Fair Information Practices, particularly in the online

context, privacy policies, privacy enhancing technologies, social media privacy identity resolutions.

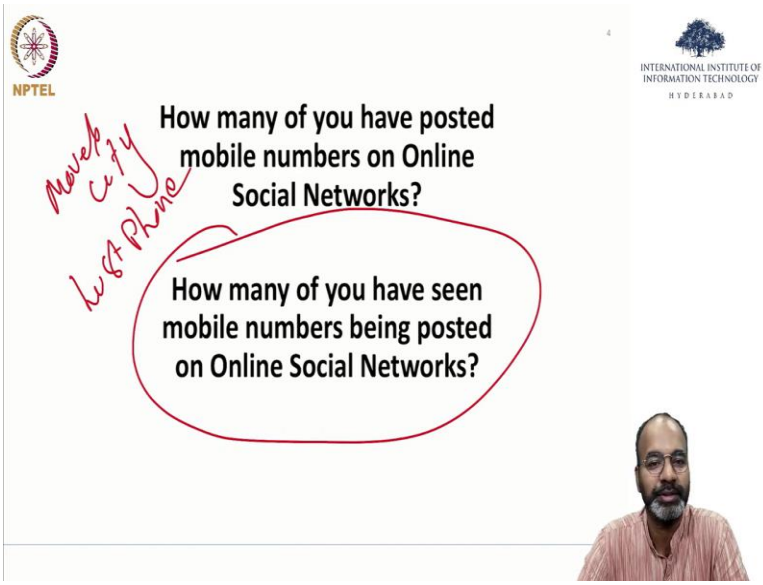
In the last week, we covered what is privacy, browser privacy leaks, and profiling from publicly available information, which, which I think will continue a different aspect of profiling is also this cell number that is publicly available. It would be really really nice if you are able to connect some of these topics at you are learning in the class to outside the class.

(Refer Slide Time: 2:45)



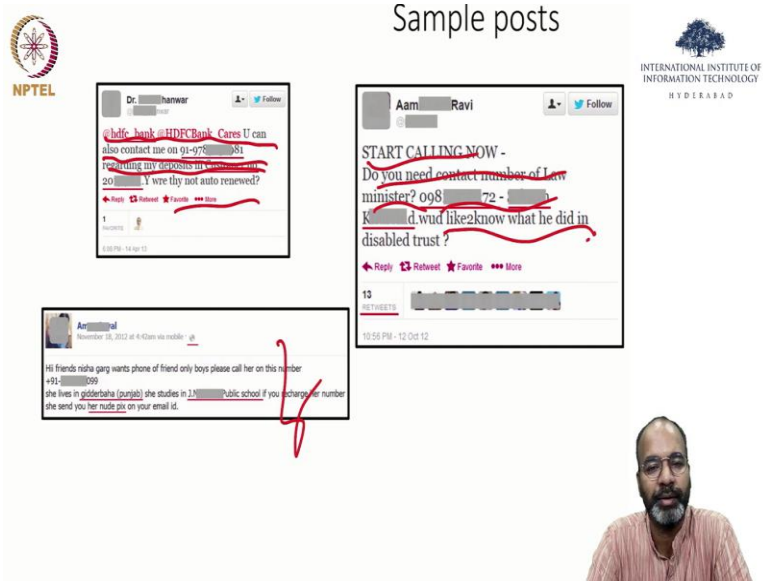
Have you ever shared your cell number online? Why would anybody share, There are people who could share it for various reasons. You may say that, look, I have not actually shared cell numbers publicly.

(Refer Slide Time: 3:04)



But have you seen others share cell numbers publicly, for possibly, yes, There are many, many times he would see the cell numbers being posted publicly and many reasons, move the moved city lost phone there are many reasons why cell numbers are posted online. And these reasons are probably valid, these reasons are necessary sometimes probably a customer service engineer is actually putting his number there So, that people could call him for any issues with let us take a washing machine or whatever he or she is able to fix.

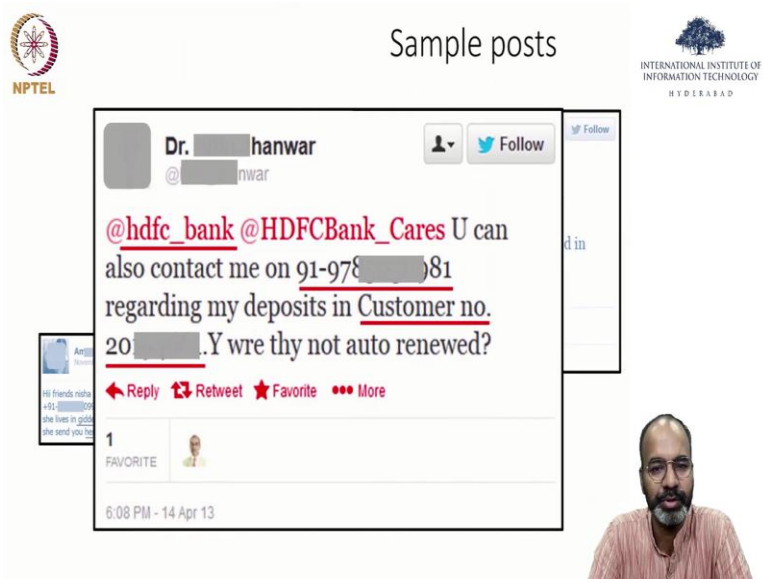
(Refer Slide Time: 4:09)



People post cell numbers for various reasons publicly, here are some numbers, some samples. But the thing to keep in mind is that it is critical to see that the cell numbers made public once people can actually take it from there and use it misuse at all that and that is the theme that we have been following in the class. How do you use the information that is publicly available for good or not for good purposes also.

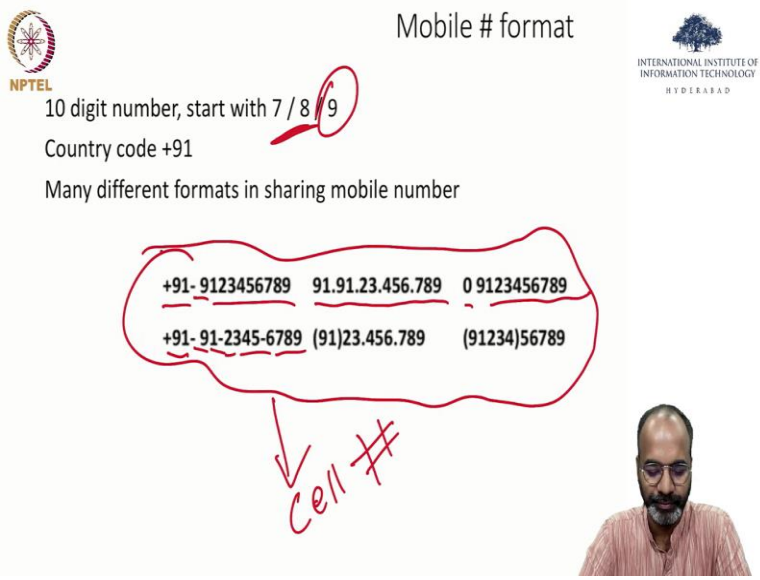
So, here if you see HDFC cares you can you can also contact me blah blah, blah, customers service number here, start calling you need contact number of law minister here and this. This is the more personal one post this number and you will get some pictures on your mobile all that, very, very intrusive posts.

(Refer Slide Time: 5:12)



Same post shown and slightly larger fonts.

(Refer Slide Time: 5:22)

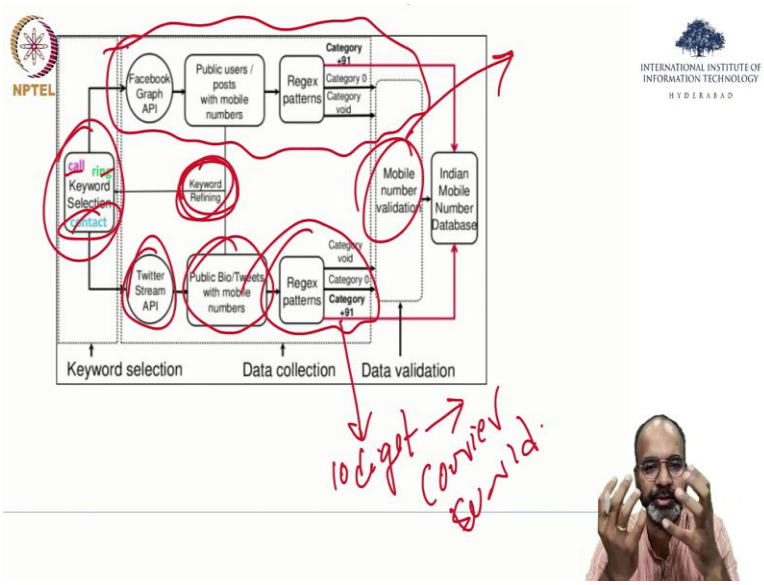


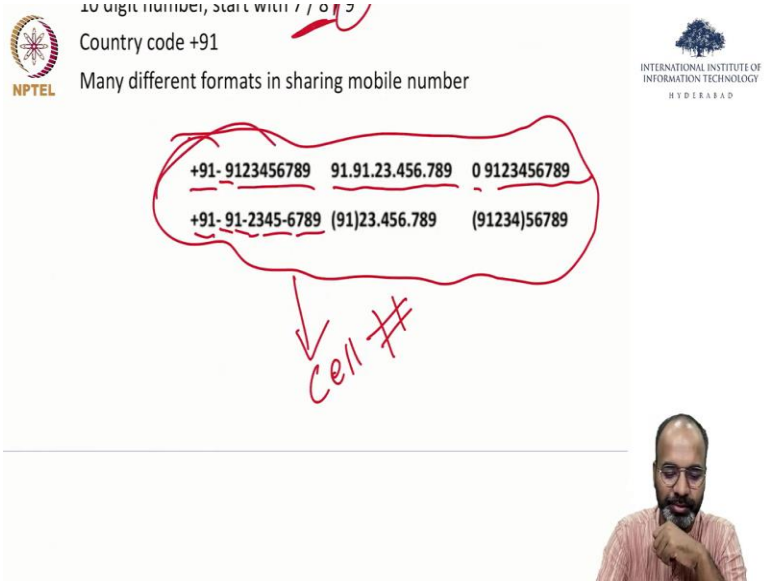
So, if you look at, in India, the cell number format, it is a 10 digit number. Earlier, it used to be only 9 starting now we have 7 and 8 also starting, probably there are other starting points, starting numbers also now. And then our code number, country numbers plus 91. And if you see the way by which numbers are being shared, I am sure you yourself would have seen writing these numbers up somewhere you would save it very differently.

Even if you just look at your phone right now, take your phone and look at the way that you have stored numbers of people that who have called you, you will have stored it in many different ways, +91 dash and this 91 dot blah, blah, blah, 091, whatever, 91 hyphen, 91 hyphen, 4 digits, 4 digits all of this can be many different ways in which the cell numbers are being stored.

And same variation shows up on social media also same variation social online, that is the point I am trying to get across in this slide that when you look at cell numbers online, you are going to have all these kinds of variations.

(Refer Slide Time: 6:58)





So, if the goal was to collect all the cell numbers online and do something with it, what would you do? How would you go about doing it? What would you get out of this when you have all this information collected to you? That is what we are going to see today. If any of you have already done some of these cool things, please let us know. It will be fun to know what you did.

What this diagram shows keyword selection, So, basically what meaning to collect data from social, you are going to start somewhere. So, here are the keywords are call, ring, contact, can

you take these keywords and look for in tweets, look for in Facebook, that any posts that has the word call. And So, let us go to the Twitter API now.

So, you search for call and then you say public bios tweets with mobile numbers. So, when you find this word call in it, then you say whether there is a 10 digit number, whether there is any number that is attached in that tweet. This is a critical part in this which is regular expressions and patterns to look for, in this tweet that you are getting. And it is important to think about these regular expressions, because these regular expressions will define the kind of numbers that you can actually get. So, for example, all of this could be converted into say regular expressions.

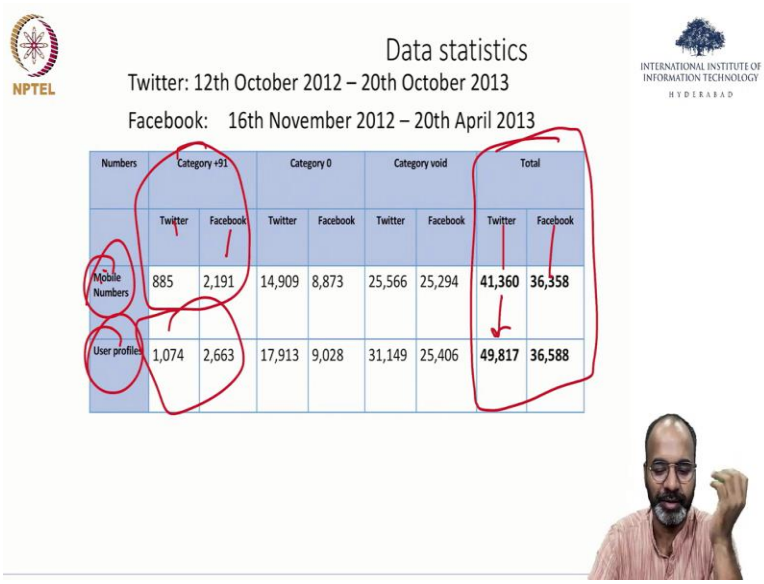
But the problem is that we have to know what we are looking for also because while doing this, we also realized that the we also realized that the 10 digit number can be many other things also can be courier tracking number, can be some other part of the world also. Plus 91 probably is harder, will say that it is in India. But other than that, if you just get a 10 digit number, that a possibility that it is numbers that are connected to some other cell number in the world also.

So, that is something to keep in mind while we are doing this data collection and finding the way to get the numbers from online, that is about the total site. So, once you get it, you get the mobile validation So, you need validation of these numbers also you cannot just take all the 10 digits from social and say that oh, they are all valid numbers.

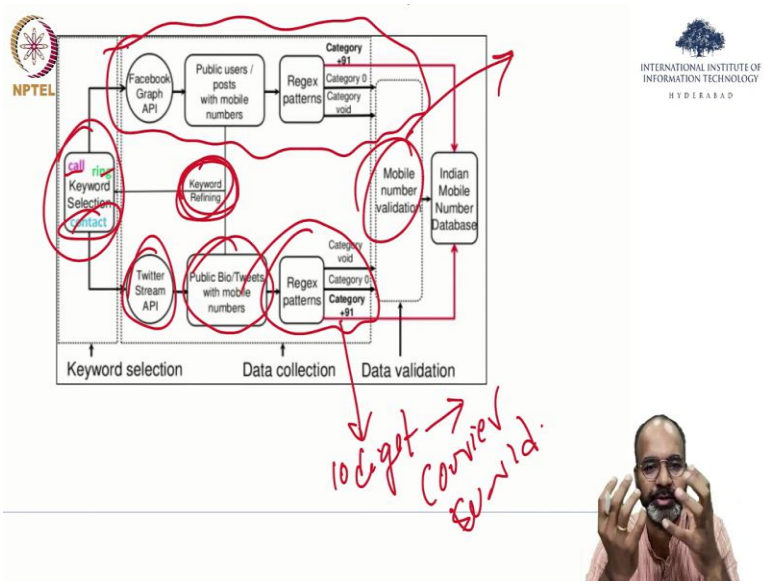
So, some level of validation needs to be done which is getting human beings to human to actually check the number, or find a way to check these numbers in a different manner. Same thing could be done on Facebook but unfortunately now, it is harder to do because search API for Facebook is deprecated. So, you cannot really search for these terms on Facebook and get it. But otherwise in the past, we could get collect data from Facebook also through search terms.

And one key theme here also is that when you start collecting contact or call or ring or something, tweets, you can also refine the search terms you get first call and then look for all the tweets and look for some specific words in there which probably call with that word possibly the cell numbers are always there, take that word and add it to your search to add it to your search API. During the process, your search results might be much more precise.

(Refer Slide Time: 11:18)



Numbers	Category +91		Category 0		Category void		Total	
	Twitter	Facebook	Twitter	Facebook	Twitter	Facebook	Twitter	Facebook
Mobile Numbers	885	2,191	14,909	8,873	25,566	25,294	**41,360**	**36,358**
User profiles	1,074	2,663	17,913	9,028	31,149	25,406	**49,817**	**36,588**



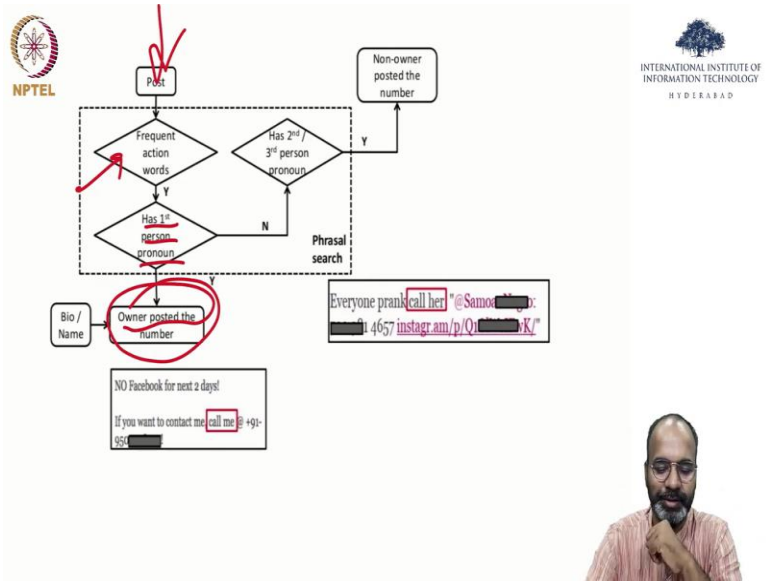
So, data was collected, but this data keep in mind that this such data can be collected even now you should any of you interested should explore this, take it, take it up, try collecting data, see whether you can actually do some interesting analysis with the cell numbers publicly available.

Some students actually recently did this also cell numbers that many even though the method that I am even though this diagram is specifically for cell numbers, I think even pre places to try anything. For example, you can think about looking at car registration number or vehicle registration number posted online.

Can you get that and find out whose car it is? Can you make some connections with the tweet, twitter handle that is posting this car picture? And can you connect to other publicly available information on the user, which we have already seen in the past. So, in this case, mobiles numbers, user profiles, this number of times mobile numbers appeared in Facebook and Twitter is here.

Number of people who posted that is here, category 0, category what… So, ultimately, what you want to look for is this this column, this says that total number of mobile numbers collected from Twitter is about 41,000. From Facebook is about 36,000 number of people who posted this 41,000 numbers are 49,000, user profiles, on Twitter, and on Facebook is about 36,000 users. So, that helps you to get a sense of what level of data can you access to the 1000s of cell numbers are being collected.

(Refer Slide Time: 13:38)



This ownership analysis is important is because you want to know whether the post that has come user himself or herself is posting it, or somebody else is posting it, here are two examples. So, post frequent action words, call, all that is there, yes, as a first person pronoun, yes, then on owner posted the number. And if I say that, for example, no Facebook next two days, if you want to contact me, call me here.

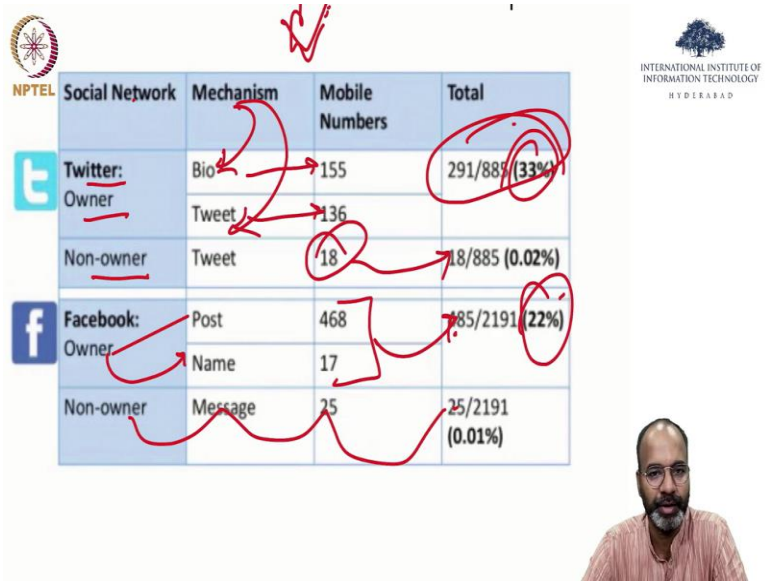
If this is there, then I think it is very clear that the user only posted the number. Whereas look at the other example. Everyone prank caller, blah, blah, blah on an Instagram link on the cell number. This is somebody else's number, this is also critically important information is because in many, if you go back to the slide that I asked how many of you have posted cell number yourself versus how many of you seen cell numbers posted you will clearly see that this number is much, much higher than the first number.

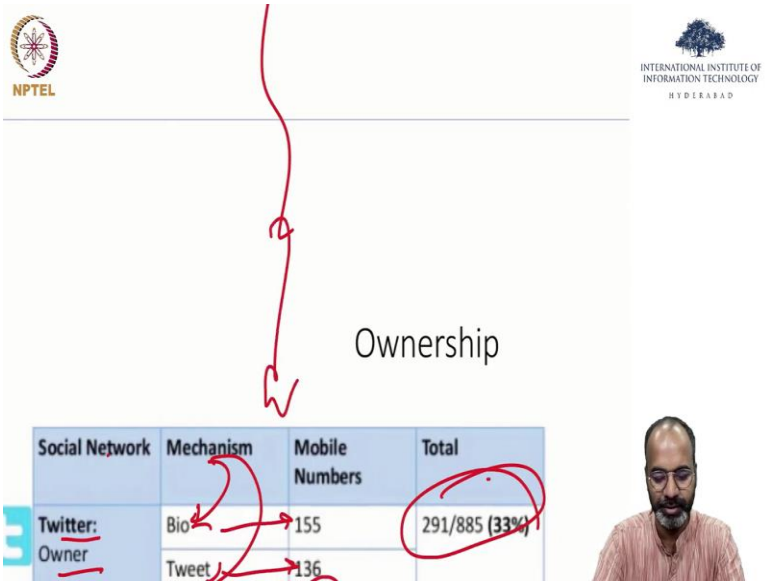
And knowing that others have posted is also important because you may be very, very privacy conscious, you may not be sharing your cell number you may not be sharing your cell number anywhere, but somebody else is posting your cell number online, that may not be a good idea, having your cell number online posted by somebody else, which can cause problems for you.

So, that is the reason why you need to know who was actually posting it and later, you will also see this part, which is how we actually get this information from the user itself. Now we are

collecting tweets and then seeing whether it is actually the user posted or somebody else posted it.

(Refer Slide Time: 16:06)





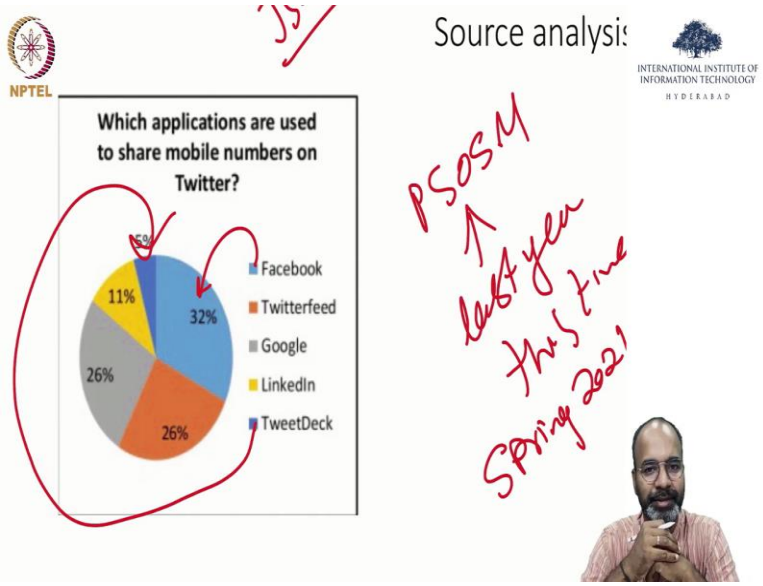
What we found is that in Twitter, So, how total number of posts you saw earlier, all this 885 number, but here this table is social network, Twitter, Owner, non-owner, mechanism bio, it is coming from tweet bio says the number, tweet says the number, therefore they both should be the owner, because even my profile, I am not going to put your number fully that is not the case, that

is what this 33 percent is, non-owner, tweet as somebody else's number, which is 18 of the posts that is all.

In Facebook, again, the same thing, Facebook owner post name this is total is 485, 485 posts has users have posted the number themselves. Non owner 25. And it is 0.01 percent. So, essentially, what this is, this table is showing us that one, the this mechanism, whatever you use, we can actually get this results.

And it also shows that majority of the times the numbers are being posted by the owner itself. Facebook has 33 percent and Twitter has 33 percent Facebook has 22 percent, hope that hope that helps to get a sense of what kind of data is getting, who is sharing the data and what kind of data is getting shared.

(Refer Slide Time: 17:50)

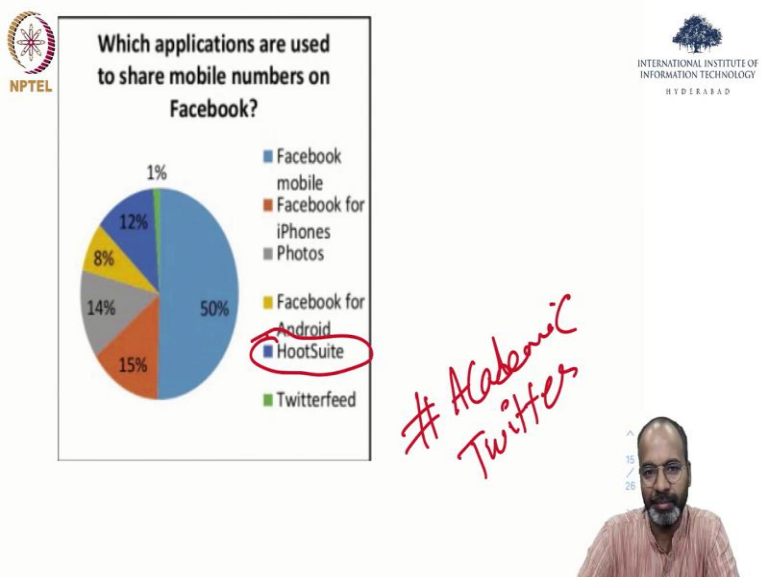


This is not this is not a directly social network analysis class, otherwise we could have actually done data collection from social media how to actually collect this data. If you are interested, please go look at the lab sessions of PSOS some course, which is also on NPTEL I thought it last year this time. And that is spring 2021 I think week two, week three, if you go look at it, you will find labs, where it talks about how to collect data or to do analysis with the data that you collect from social.

So, one of the ways one of the things that you can derive from these tweets that you get these Facebook updates you get is which applications are used to share this mobile numbers. Then lots of these kinds of information comes with the JSON when you collect from the social data JSON objects. So, this one is source analysis, which is it says basically the 32 percent of the people are actually posting on Twitter using sort of, say Facebook as a source.

Cross posting, that is what this is. This is showing a Tweet Deck people are about 5 percent Tweet deck is a platform which is used by people to manage multiple accounts Tweet Deck is a extremely powerful tool to manage twitter, which is you can have different slices of data that you can see I used to use it at some point in time but I never got to continue using it. But I know many people who manage multiple accounts used to adapt to powerfully use the features of Tweet Deck.

(Refer Slide Time: 20:12)



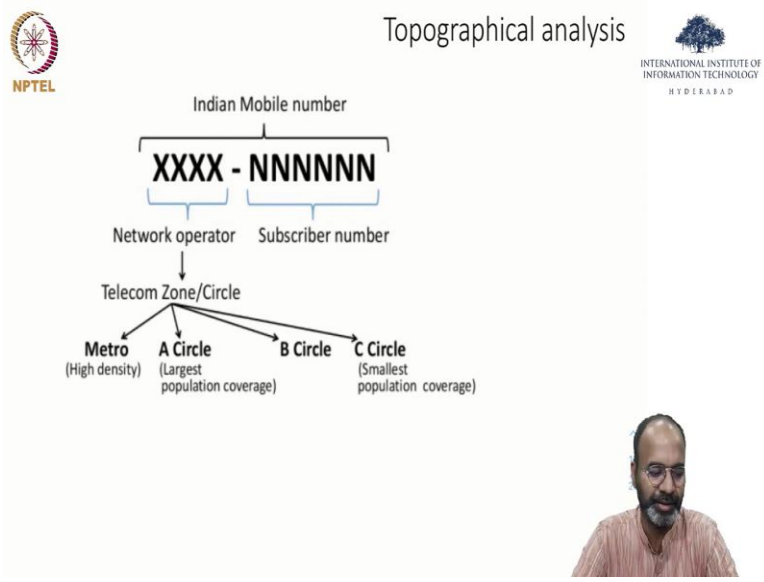
There is also I think, in the next slide, you will see Hoot Suite. So, this is a another platform that is actually very popular, which actually does show how to meaning it has mechanisms by which you can schedule post, set up multiple accounts, manage this multiple accounts together and views of the tweets that you get can be also controlled, for example, and with that, you can set up one particular hashtag stream in one view, another hashtag or probably only the user let us take you want to follow only you want to look at one particular user post whatever he or she does, there is a view only for that.

Ultimately, it is all the same data so, it is not that Tweet Deck or Hoot Suite is giving you a view of a data that is not something from Twitter. If you just go to Twitter directly it is cluttered in some sense, it is cluttered if you want to see all the people that your following tweets are going to show up in your newsfeed in your feed.

And there is also distraction you do not really get only I mean one meaning unless you actually unfollow people that you do not want to see the tweets and then get only the people that is probably a way but Tweet Deck allows you to declutter this part and then take a view of it only particular hashtags has been very helpful for example I at some point in time I used to keep this hashtag academic Twitter stream on Tweet Deck.

Because otherwise in Twitter you have to go search for hashtag Twitter and then look at the hashtag academic Twitter and look at all the status come there whereas what this in this stream would show you is all the posts in that Ackerman Twitter it will actually keep showing me in the one view of Tweet Deck. Take a look if it is of interest in on Facebook read in the mobile numbers came from is the source analysis here Facebook for mobile Facebook for I phone, Facebook for android, HootSuite all of that.

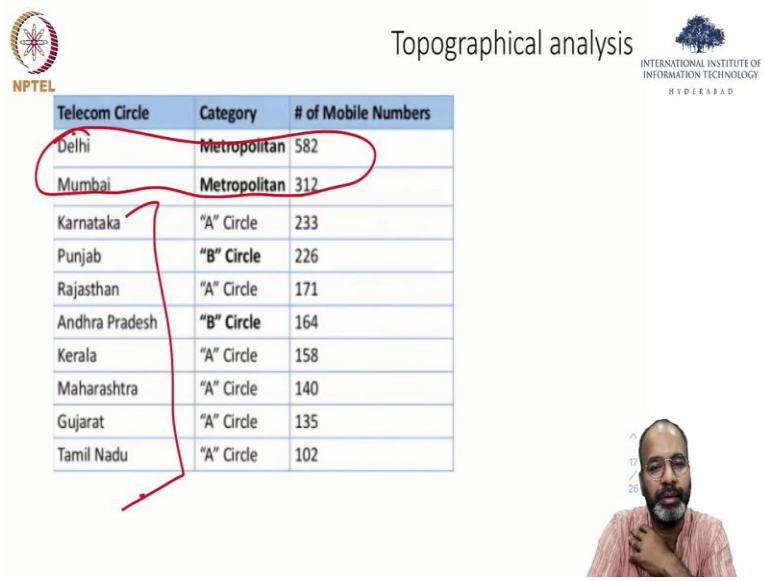
(Refer Slide Time: 22:30)



So, now let us look at the numbers itself, numbers itself having is what when a 10 digit number if you break the 10 digit number what all you can get, you can get network operators subscriber

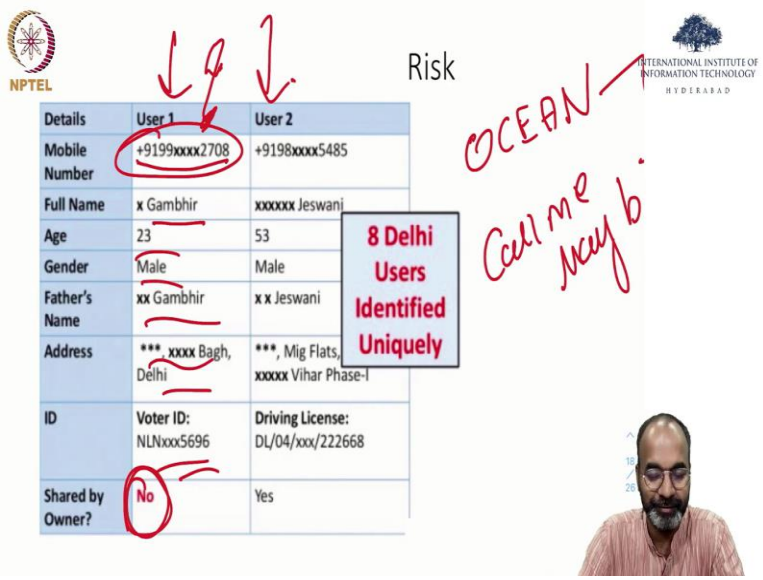
number. So, if you look at the first four digits, then you can say that okay, there is this Metro, large population coverage B, Circle C circle all of this. So, there are different pieces of information that you can get from the first part of the number itself. And these things we actually understood from some telecom provider, it is possible that some things may have changed, but this is where this is what we had when we were doing this work.

(Refer Slide Time: 23:25)



The goal for knowing the topographical information is this one, which is to know that from where the information is getting accurate from where the cell numbers are shared most number of times. So, if you look at this table, it basically says Delhi and Bombay are the highest it could be bias that the total number of tweets itself that is coming up on social total number of users on social itself, Delhi and Bombay may be high, keeping that caveat in mind, keeping that bias of users being hired from Delhi and Bombay on social desire, this is the table that we get.

(Refer Slide Time: 24:17)



Details	User 1	User 2
Mobile Number	+9199xxxx2708	+9198xxxx5485
Full Name	x Gambhir	xxxxxx Jeswani
Age	23	53
Gender	Male	Male
Father's Name	xx Gambhir	x x Jeswani
Address	***, xxxx Bagh, Delhi	***, Mig Flats, xxxxx Vihar Phase-I
ID	Voter ID: NLNxxx5696	Driving License: DL/04/xxx/222668
Shared by Owner?	No	Yes

And you will see other states from where the number is being posted. I am going to connect to something else that we saw earlier, but hopefully you will be able to go back and look at the lectures if you need we spoke about ocean at some point in time right publicly available information on government websites. How can you collect the data how can you put the data to create a profile all that.

So, now one thing that you can do is that this So, there is ocean and that is this work is called Call Me Maybe. Ocean and call me maybe, So, now what can you do is that you can actually put them together. Ocean is random name if you remember the context I was giving was railway train taking train from one location to another location finding out somebody that you know is in the train and trying to talk to them same background.

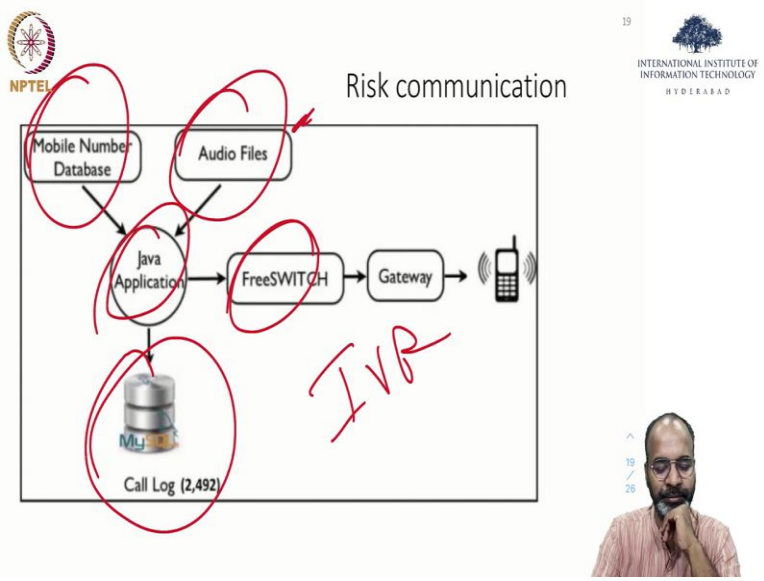
Ocean is that Srishti Rawat if you remember that is the name I kind of used that. But here in column E, maybe the starting point is a cell number. Oh, I get the cell number on Twitter, now what can I do with that? So, if you put these two together, now it becomes even more intrusive.

Srishti Rawat starting name random name and a cell number random name random number from the cell number, you could actually go to the name or from the name, you could come to the cell number, which is what we ended up doing. We ended up actually connecting multiple users exactly in terms of cell numbers the name age, all of this came from the ocean address voter id

all came from again, ocean. This one shared by owner this we got it from call maybe. And of course the cell number we got it from call me maybe.

So, thereby, we were able to uniquely identify users, exactly starting from two different random places, and both of it to be personally identifiable starting his name and starting cell number. I do not know if any of you feel like this is intrusive. But if you are from Delhi, definitely you could be identified because we did all this over the data from Delhi. So, therefore, if you are from Delhi data can be easily identified in terms of the information that is collected.

(Refer Slide Time: 27:22)



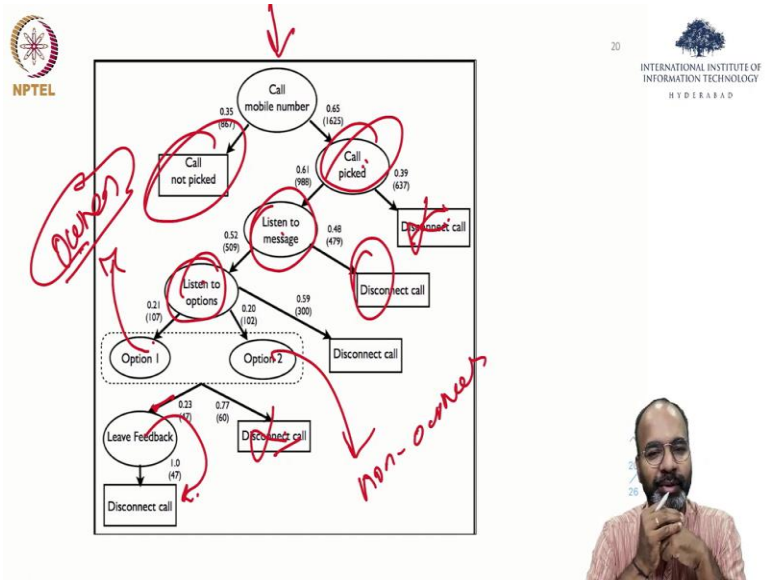
So, once we were at this point that okay, So, we are able to collect tweets cell number and know the person's voter ID or something. We wanted to actually do some communication of this that Oh, such kind of information could be collated, you saw this number of cell numbers that we had that is what was your mobile number, So, cell numbers, audio files or IVR. So, if you I am assuming all of you know what an IVR is, IVR is interactive voice response system, which is when you call up, a bank today, they will say press 1 for English, 2 for Hindi, 3 for Telugu.

After you have pressed 1, they will tell savings account, and then fix the deposits, all these kinds of things, it is presented to you by a voice it that is what is the interactive voice response system. Go look up if you need any more details there. But what we built was a simple mechanism by which take the audio files, which is audio files of asking the survey, I will show you what the

survey was but saying that oh we are calling for we are pre call researchers, we are calling you to tell you that we found your cell number online and and then ask them the questions about please click 1 if you posted the number yourself, please click 2 if you have press 2 So, to say press 2 if you have not posted the number yourself all that.

So, audio files free switches the platform that we used, and that is the last user that pivot column. We did this for all the cell numbers that we collected.

(Refer Slide Time: 29:39)



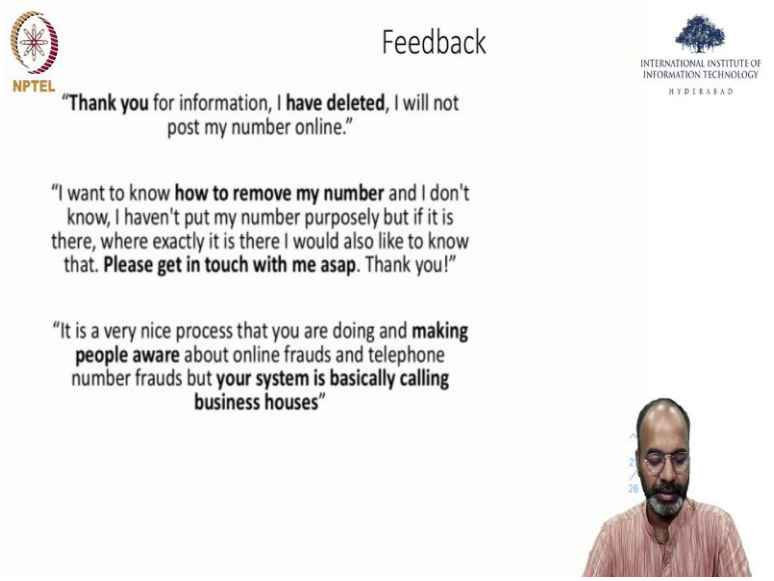
This is showing you how the total sort of decision tree went which is starting a call somebody gets the call, we take the we take the cell number from here, call the number free switch just the call it goes to the user. So, that is where the user gets the call. Let us assume that UK user is U, when you get the call what all can you do, you would not pick the call somebody picked the call

when you picked the call you are listening to the message which is research from pre-call blah blah blah all that and then you disconnect you picked and you realize that wrong it is actually I should not have picked you disconnect. Then listen to the message you listen to the entire thing of pre call thing and then you say listen to the options which is press 1 for if you uploaded the so, this is where we wanted to understand the ownership, option 2 is I am not the owner I did not post it click 1 if you posted it click 2 if you do not post it.

Then once that is done leave a feedback for what did you think about this whole data collection that we did, some people did not even give feedback say disconnected, somebody gave the feedback and then disconnected.

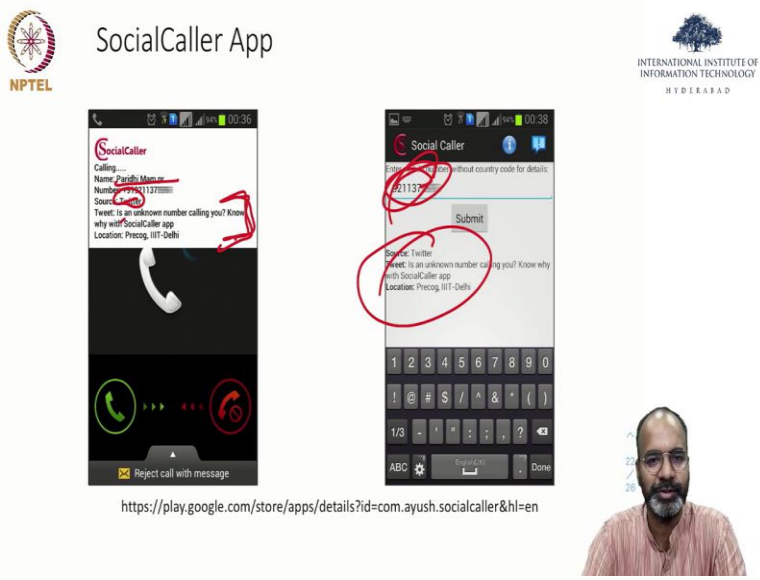
So, this graph helped us to understand quite a few things and it was a phenomenal insight into how users think about it people came back I have a slide which is about feedback but I was super excited to see how users reacted to this how people came back to me saying that how it helped them they did not know that their numbers were online they wanted help for deleting the numbers all of this went on at that time.

(Refer Slide Time: 31:55)



I said feedback and it is here, thank you for information I have deleted I will not post my number online anymore I want to know how to remove my number and I do not know I have not put my number purposefully but if it is there where exactly it is there I would also like to know that please get in touch with me say thank you. It is very nice process that you are doing and making people aware about the online frauds and telephone number frauds but your system is basically calling business houses

(Refer Slide Time: 32:30)



So, we took some of this feedback we were actually super excited about doing something with this, then we built the app called social caller motivated by truecaller idea where so, today when you get a call today when you get a call you see in truecaller who is calling. And the truecaller information is coming from how that number is being stored by somebody else who is using truecaller.

So, what we wanted to do was we wanted to imitate this but when the call is coming we wanted to add this information saying here it is Paridhi is calling but the information that we got from Twitter, Twitter has this information about her truecallers also could add this information of course if they know that the number where the number that is stored if they have the connect to the social media it is on play store feel free to take a look at it

So, a test case that happened was interesting was there were the way that the users ended up using most of the times was typing in their own number to see whether there is any publicly available tweet about them, that is from the test case that we built for we built it for users downloading the app and whenever a call comes they would actually get to see who is calling but the test case was majority of the times not majority sometimes it was used for searching their own number and taking it from there.

(Refer Slide Time: 34:14)

So, if you are interested in this, this is the paper gives you a lot of details about what all we did but my theme for connecting it today to in this course is that cell numbers. So, to say or PII are publicly available you can actually curate the cell numbers and use and misuse.

Of course there is a way to protect this so, you can have I mean I think the DND all of that is one way of reducing the impact of even though if the cell numbers are even though the cell numbers are public if your number is on DND you should not be getting the call all that but the goal is try and find not to have these numbers online.

If the numbers are online also how do you actually stop, how do you actually find ways that these numbers are online how do you actually delete them all that. I think more and more of us should think about helping people around us to reduce their publicly available information online.