

So, that that kind of things I think should be somehow found a way to deleted probably reporting the a tweet may be one solution, finding methods by which those that can be connected to the user whose number is being posted there. And that user is actually mentioned said that your number is online something like that. So, that is what this Call Me Maybe paper is all about.

(Refer Slide Time: 2:13)

The slide contains the NPTEL logo in the top left and the International Institute of Information Technology Hyderabad logo in the top right. The central image is a map of a city with several green location pins and lines connecting them, representing a location-based social network. Handwritten red annotations include 'LBSN' circled in red, 'Victims' written in red, and a list of '4 v's': Variety, Velocity, Volume, Veracity, with 'Value' written below them. A small video inset of a man is in the bottom right corner.

So, next thing that I wanted to cover is about the location based social network. I think in my first week, you saw one slide, which has mentioned about I think, foursquared we were having a lot of discussions about this 4 v's 4 v's of social media, and in that we talked about variety, velocity, volume and veracity.

So, and then I also mentioned that this 5th v that is being discussed, which is value. So, keeping this in mind, there was one mention about Foursquare, I think many of you did not know about Foursquare, or may not have used Forcesquare very frequently also it is not a very popular social network in India. But Foursquare has a feature which is actually very popular in India now which is this location based where you can, So, why do we need location based? Let us start from there.

Location based social networks are extremely useful to provide suggestions or recommendations to users, which are very, very timely, if I know that you are in Gachibowli now or you are in Connaught Place now in Delhi, or you are in Dadar in Mumbai, if I know those locations that you are going to be sure you are currently or that I think we can produce a lot of interesting

recommendations for you given that the assumption is that you are okay with getting those recommendations.

If that is acceptable, if that is disclaimer there that I think this location based works very, very well. And what kind of recommendations can be it could be saying that oh, look, you are in this mall in Chennai. There is an Adidas showroom and there is a 20 percent sale in this mall right now, or it is around during lunchtime, this if you go to this ice cream parlor and show give this code you will get 15 percent discount on the bill.

So, all of this could be very, very useful, again, please remember it is assuming that you as a user have accepted the condition that your information can be used to provide this advertisements. It can also be useful for crime detection. So, assuming that if I get access to all the victim victim's location where all he or she has been through this location based social network.

Then I can actually profile the user and see whether I can get to some other information using that. So, if the victim has moved from three different places inside in Hyderabad can I actually go to those places look at what collect some information from those places and use it for crime reduction.

So, many use cases are there if the location is available, of course, it is also privacy which is why we are studying in this course it is also privacy invasive if I keep the location information not public and still if that could be used for deriving where I am then it is actually invasive or probably I allowed only the service provider to make the choice using my location but that information is actually provided to everybody now.

(Refer Slide Time: 6:47)

NPTEL

LBSN

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Create Post

Ponnuram Kumaraguru is at International Institute of Information Technology, Hyderabad

Public

What's on your mind?

ADD TO YOUR POST

Post

check in

25 / 29

NPTEL

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Create Post

Ponnuram Kumaraguru is at International Institute of Information Technology, Hyderabad

Public

What's on your mind?

ADD TO YOUR POST

Post

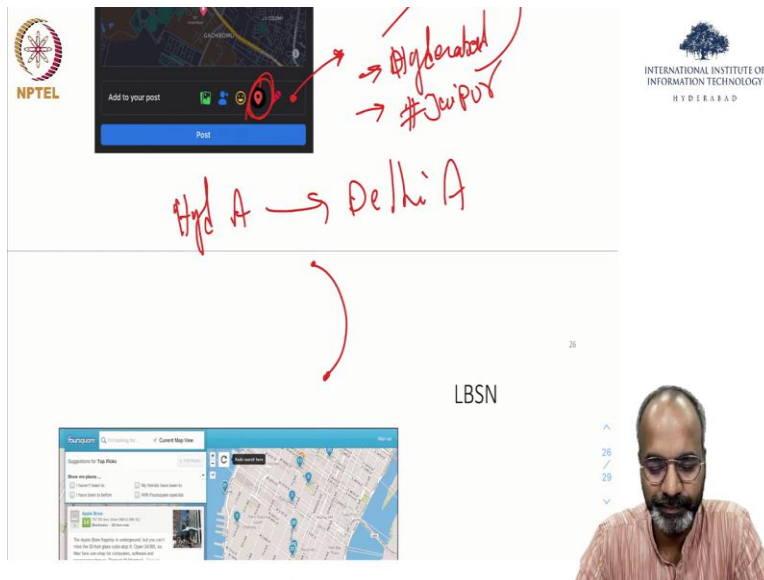
check in

Please Rob me .COM

Hyderabad #Jeeput

25 / 29

I RCM



So, these kind of privacy issues are there when the location based information is shared. So, this is the way I am sure many of you would have seen the location based social network right now. So, I when I was preparing for this lecture I was trying to check into Triple IT Hyderabad. So, in Facebook if you are going to create a post it if you go look for this particular three dots and then it will actually give you actually click on the three dots you will get this location.

And then from there you can say that I am going to check into Triple IT Hyderabad and then you do a post now this post is connected to this location and then if anybody goes looks at all the posts that are in that location they should they can get this post. This is also very helpful because I mean for example let us take if you are travelling to another city and you want all your friends to know that you are in that city.

You do not have to SMS every friend you do not have to email you do not have to basically the way I see it is that you do not need any buddies email addresses, cell number anymore if you are connected to them on social that is good enough. That is why if you see if any of you have been following me on social networks throughout this class, you would have seen me posting about the places I have travelled.

When I travel I take a picture of the airport saying that, oh I am in the city for so, many hours so, that people around here I can catch up. I am sure if I want I can send it to the specific people who might want to speak with which I did anyways in my travel which I do anyways. But this is more like a broadcasting of let us take for example an alumni and Triple IT Hyderabad, alumni of

Delhi is there we want to catch up I do not know the number of these people I do not know the email addresses all that.

But on the other side, this me saying that I am not in Hyderabad that I am travelling to a city A can also be misused against me, for example, there used to be a service called please rob me.com I think the site is still there if you go look at please rob me.com what they did was, So, for example, in my profile in my profile, you will see that I am in Hyderabad it says Triple IT Hyderabad all that.

So, they would look at my tweet where I say that oh, I am travelling to hashtag let us take Jaipur or I could have checked into airport in a Jaipur any of that location. Location information can be derived by three four ways, one, I have written it in my profile too, I post the tweet in geotag locations at 3am I am actually saying in the post itself, oh landed in Jaipur airport I am around for so, many hours.

All of this, please rob me.com looks at and then it says oh, look, he is from Hyderabad he is not saying that he is in Jaipur, which means he is not in Hyderabad. So, let us just post his profile on please rob me.com for burglars to get to my home. So, that is the idea of please rob.com and was basically using the location based information that users are sharing on social network.

And I am also guessing you would have seen on Facebook, this is only a location based post that I did. But if you were to do a post saying I am going from Hyderabad airport to Delhi airport, then there would be then there would be an India map, there would be a red line from like this, Hyderabad to Delhi and it will show that the post that I said saying 48 hours in Delhi, anybody around to meet.

(Refer Slide Time: 11:27)

NPTEL

LDJIN

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

NPTEL

LDJIN

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

moyan
Ship

NPTEL

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

NPTEL

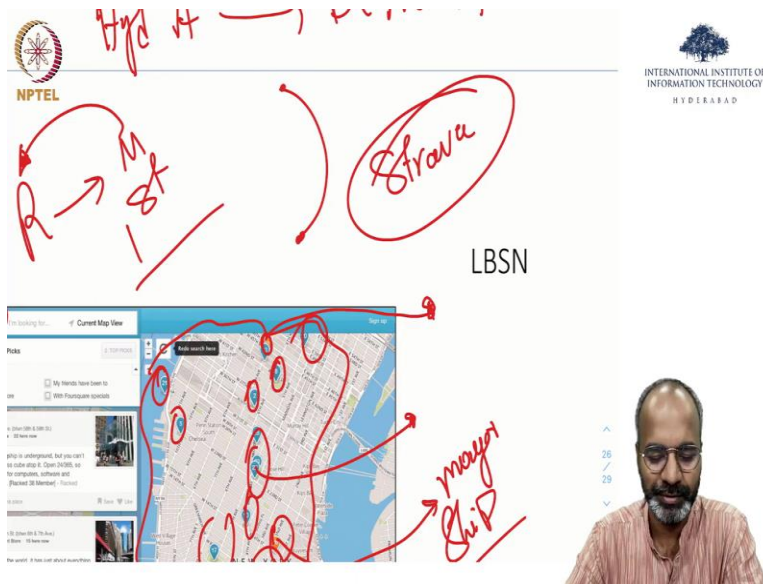
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

checkin

Please Rob .COM

Hyderabad

#JeePi



So, many of us would have seen I am sure you have seen these kinds of posts on your timeline. That is what I am talking about. Interestingly, there is a social network, which is Foursquare very, very popular in terms of only the location base, here, I mean, like the Facebook like the Insta, the minimum atomic level information on this platform is actually a location, whereas in YouTube, it is a video, in Twitter it is a microblog text, Facebook, it could be images, it could be text all that.

Foursquare is an interesting social network in that sense because just look at this part of the image. So, these are the check ins of the user in some part of the world, it is easy to derive which part of the world it is New York. So, the way it works is that you go you are travelling, you take your phone, you kind of go to the Foursquare app and you say that, I want to actually check into this location, for example, which is exactly what I was doing in Triple IT Hyderabad this slide, I am doing a post saying that that post is relevant to Triple IT Hyderabad, I am in Triple IT to Hyderabad all that.

Similarly, I could actually check into Triple IT Hyderabad on Foursquare. So, that is this blue marker, So, these are the locations that they use. So, if you notice there was a there are two different colors here there is this one orange here and another orange here. Before going ahead with the video, can you pause and think for a second what would that be?

Hopefully you thought about it and you had some answers but these orange ones are interesting information that it is basically called as mayor ship. Mayor ship is a concept that Foursquare had,

which is if you are the person that has come to that location and checked in, the most number of times in the last 60 days, then you become the mayor of that location.

Why is that useful? This is useful in many ways, again, I mean, I think all of these social networks, the features that they had or had some kind of motivation or incentive for users to use, and that is why they were getting they are getting popular.

So, the mayor ship is actually very useful is because if I know that you are the mayor in this location, for example, think of a restaurant, I run a restaurant and you are the mayor, and you come to my hotel for food and I know that you are inside the restaurant right now. Many useful, because I can actually come and talk to you thank you for coming more frequently. I can give you 10-15 percent discount, I can find my chef to come and talk to you, I mean all of this is happening by the way, I am not making it up.

All of this may, I have seen it myself to go to restaurants and there is a screen that is projected where they project all the Foursquare users who have checked into the restaurant right now and who are in the restaurant and the mayor would show up on the top. There can be only one mayor, because number of times somebody could be maximum, I do not know what they do when two people are the same number. But in last 60 days, same number, probably the probability is very low.

So, therefore, this mayor ship is very useful. Mayor ship also gives you a lot of information about the person's preference, I come to the restaurant and I am the last 6 days meaning also please remember, it could be that in the last 60 days, nobody has checked into the restaurant and I am the first person to check in and I become the mayor, that is also true. But if it is a popular place, and if there are more people checking in, then if you are a mayor just gives you more that Mayor feeling I guess.

This mayor ship has been also used for sort of giving parking spot free, for example malls in some locations actually provide if you are a mayor, you get a one week parking free. So, monetization of some of these things, has happened very well or is happening very well, using these social networks.

So, the mayor ship also can give you a lot more, I mean this is about restaurant and all that, but think about mayor ship of we could actually use the mayor ship, and for example, if you are at home, if you are checking in at home or if you are checking at your work, potentially we could find out that if you are going to that location more and more frequently could be something that you are interested in.

Let us take a place that you go jog every day it looks like, there is also another network called Strava which has this which is designed for doing it for exercise, jogging and cycling all that. But if you were to say Lodhi garden, if you go in Delhi for walking every day, and you check in. So, we know that something you do that in that location very frequently.

(Refer Slide Time: 17:50)

**We Know Where You Live:
Privacy Characterization of Foursquare Behavior**

Tatiana Pontes*, Marisa Vasconcelos*, Jussara Almeida*,
Ponnurangam Kumaraguru†, Virgilio Almeida*

*Universidade Federal de Minas Gerais, Brazil
†Indraprastha Institute of Information Technology, India
*{tpontes,marisav,jussara, virgilio}@dcc.ufmg.br
†pk@iiitd.ac.in

ABSTRACT
In the last few years, the increasing interest in location-based services (LBS) has favored the introduction of geo-referenced information in various Web 2.0 applications, as well as the rise of location-based social networks (LBSN). Foursquare, one of the most popular LBSN, since its inception, has

INTRODUCTION
Online social networks (OSN), such as Facebook, Twitter and the recent Google+, are currently very popular. Some reasons for their great popularity include the easiness at which users can communicate and share content at large scale, the

PICS

That is the details about what our location based social network is. What we will do now is we will actually look at the flip side, that is the basics of social location based social work. Now I will flip around and then see how we can use the same information to actually find something that you are not making public. Two parts of this work one I will go in very detail the other one I will let you actually look at it yourself.

Skimmed through but I will actually let you to look at it yourself more closely. The first one is we know where you live privacy concerns of Foursquare behavior. The second one is beware of what you share inferring home locations and social networks. Both idea are the same, these ideas

are also built on some concepts like so, there are others who have studied if you upload a picture on social network without saying what location it is, still the location can be inferred roughly.

You take a picture in front of a Eiffel Tower, you take a picture of a of you standing in front of let us take Taj Mahal, what is difficulty in figuring out what location it is, but if you do it in front of Triple IT Hyderabad, when the buildings are not known, or the location is all everything is trees, it is hard to find, So, that is the sort of spectrum or inside your home, you take a picture and upload it, how do we find out where those locations.

Outdoor pictures I think is possibly easier to get which locations are particularly when they are popular when there are more and more pictures around these locations it is probably easier to find out which location this is. But we will actually use the location based information to derive where you live and where you work.

(Refer Slide Time: 20:13)



information in various Web 2.0 applications, as well as the rise of location-based social networks (LBSN). Foursquare, one of the most popular LBSNs, gives incentives to users who visit (check in) specific places (venues) by means of, for instance, mayorships to frequent visitors. Moreover, users may leave tips at specific venues as well as mark previous tips as done in sign of agreement. Unlike check ins, which are shared only with friends, the lists of mayorships, tips and dones of a user are publicly available to everyone, thus raising concerns about disclosure of the user's movement patterns and interests. We analyze how users explore these publicly available features, and their potential as sources of information leakage. Specifically, we characterize the use of mayorships, tips and dones in Foursquare based on a dataset with around 13 million users. We also analyze whether it is possible to easily infer the home city (state and country) of a user from these publicly available information. Our results indicate that one can easily infer the home city of around 78% of the analyzed users within 50 kilometers.

reas
user
opp
as th
lot c
rela



Due
Glo
prev
mur
[20]
to C
tion
stud
of it
lo a
spe
th





one of the most popular LBSNs, gives incentives to users who visit (check in) specific places (venues) by means of, for instance, mayorships to frequent visitors. Moreover, users may leave tips at specific venues as well as mark previous tips as done in sign of agreement. Unlike check ins, which are shared only with friends, the lists of mayorships, tips and dones of a user are publicly available to everyone, thus raising concerns about disclosure of the user's movement patterns and interests. We analyze how users explore these publicly available features, and their potential as sources of information leakage. Specifically, we characterize the use of mayorships, tips and dones in Foursquare based on a dataset with around 13 million users. We also analyze whether it is possible to easily infer the home city (state and country) of a user from these publicly available information. Our results indicate that one can easily infer the home city of around 78% of the analyzed users within 50 kilometers.

Author Keywords



patterns and interests. we analyze how users explore these publicly available features, and their potential as sources of information leakage. Specifically, we characterize the use of mayorships, tips and dones in Foursquare based on a dataset with around 13 million users. We also analyze whether it is possible to easily infer the home city (state and country) of a user from these publicly available information. Our results indicate that one can easily infer the home city of around 78% of the analyzed users within 50 kilometers.

Author Keywords

Location Prediction, Privacy, Foursquare

ACM Classification Keywords

K.4.1 Computing Milieux: Computers and society—Public policy issues

General Terms

Experimentation, Measurement

op
as t
lot
rel

Du
Glc
pre
mu
[20
to
tio
suc
of t
loc
spe
tha
wit
in I



mu
[20
to
tio
suc
of t
loc
spe
tha
wit
in I

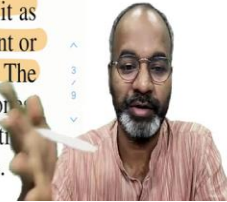


In l
ren
virt
c
e
a
r
f





Users can post *tips* at specific venues, commenting on their previous experiences when visiting the corresponding physical places. Tips can also serve as feedback, recommendation or review to help other users choose places to visit. Examples of tips include the best option of a menu in a restaurant, the best place to have lunch in an airport, or even a complaint about a service. With a limitation of 200 characters, tips nourish the relationship between users and real businesses and may be a key feature to attract future visitors [17]. Each user has a history of all tips she posted, with associated venue and timestamp. When visiting a venues' page, after reading a previously posted tip, a user may mark it as *done* or *to-do*, in sign of agreement with the tip's content or intention to visit that location in the future, respectively. The history of mayorships as well as the list of tips and dones along with corresponding venue and timestamp information of a user are publicly available at the user's profile page.



So, let us go to the paper, So, that is the paper and in the process of this, I am also letting you to look at paper because if you remember the post condition for the course, one of it is to actually to look at papers and understand how to review a paper, how to understand what is in the paper, all that.

So, this is Foursquare its using Foursquare as the network, before we get into the details of the paper, we need to know some specific technical terms, which is part of Foursquare I just used to check in word very casually check in as a mechanism in which in Foursquare, you say that I am in this location now.

So, here, check ins are there so, unlike check ins, which are shared only with friends, the list of mayor ships tips and dones of a user are publicly available to everyone, thus raising concerns about disclosure of the users movements patterns and interest. So, we will see, do not worry, we will see what a tip is what a done is we have already we know what this is, we know what this is.

So, the goal of the paper was to characterize the use of a mayor ships tips and dones in Foursquare based on a data set of 30 million users. At that point in time, whenever this work was done, 13 Million users of Foursquare was the largest data that was used, it is potentially arguably the entire Foursquare at that point in time.

So, the results was 78 percent of the users within 50 kilometers which could be derived, which is I could actually accurately find out that where your home location is, with the error rate of 50

kilometers 78 percent of the times. So, I did no other that is intrusive enough but think about it, if I can easily find out 50 kilometers is a little long.

But when you see the graph, you will actually see that this 50 kilometers is only the max to look at one set of users, but there are users who you can easily find out where their home is where they work, probably 0 kilometers point in time. I am going to look at only the specific parts of the paper just to give you the insights on important parts of the paper and to get a sense of what the analysis is, feel free to read the paper if you are interested if you have any questions.

Please post it, I will be happy to actually answer. This is one of the super cool work we did, meaning I think all the all the work that we are discussing in this course are actually cool, that we had done at some point in time. But it is definitely one of the interesting pieces of work that we did.

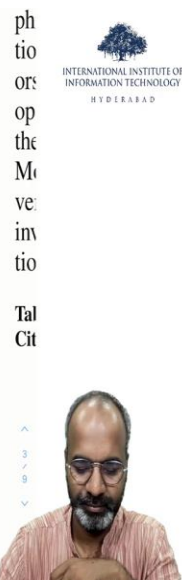
So, what is the tip? Users can post tips at specific venues commenting on their previous experience when visiting the corresponding physical places. So, I come to Triple IT Hyderabad I see that I have already checked into the location and I want to give a tip saying that, oh the campus is very beautiful, the buildings are very new, the classrooms are very modern. So, these kinds of inputs if you give those are the tips.

(Refer Slide Time: 23:58)

acters, tips nourish the relationship between users and real businesses and may be a key feature to attract future visitors [17]. Each user has a history of all tips she posted with associated venue and timestamp. When visiting a venues' page, after reading a previously posted tip, a user may mark it as *done* or *to-do* in sign of agreement with the tip's content or intention to visit that location in the future, respectively. The history of mayorships as well as the list of tips and dones, along with corresponding venue and timestamp information, of a user are publicly available at the user's profile page.

Crawled Dataset

Our study is based on a large dataset collected from Four-square using the system API. We crawled user profile data consisting of user type, user home city, list of friends, mayorships, tips, dones, total number of check ins, Twitter screen names and Facebook identifiers. Our crawler ran from August to October 2011, collecting a total of 13,570,060 users,

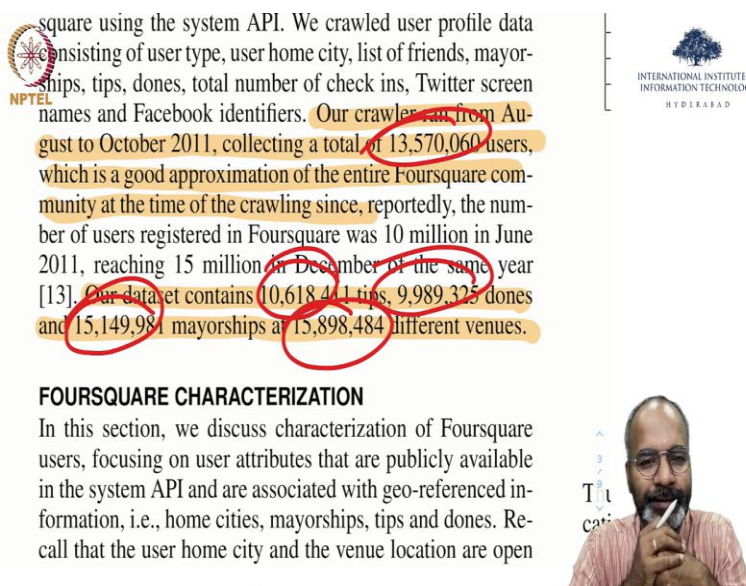


What is the venue? Venue is Triple IT Hyderabad itself is an venue, Taj Mahal, Eiffel Tower, Connaught place these venues on Foursquare. When visiting a venue page after reading a previously posted tip, which is I posted saying that Triple IT Hyderabad is a beautiful campus you use a market as a done or to do in sign of agreement with the tip content or intention to visit that location in the future respect. So, now we are defining what a tip what done or to do is.

I posted Triple IT Hyderabad is a beautiful campus you come and see it you also agree to it saying yeah, I also agree to it or you have never been to Triple IT Hyderabad campus and you say that look, I want to go to this campus. So, both this what is a done, which is you agree to do is you want to actually do you want to actually visit campus I do not know but it may be the case that you can think about it for you can think about it for let us take a restaurant.

I ate in a restaurant and then I liked the restaurant and you want to, I say that the restaurant is pretty good, the menu is very good, they have a diverse set of menu, the food was for high quality, the hygiene was good, all that I say and then you can you can actually done or to do with it.

(Refer Slide Time: 25:31)



square using the system API. We crawled user profile data consisting of user type, user home city, list of friends, mayorships, tips, dones, total number of check ins, Twitter screen names and Facebook identifiers. Our crawler ran from August to October 2011, collecting a total of 13,570,060 users, which is a good approximation of the entire Foursquare community at the time of the crawling since, reportedly, the number of users registered in Foursquare was 10 million in June 2011, reaching 15 million in December of the same year [13]. Our dataset contains 10,618,411 tips, 9,989,325 dones and 15,149,981 mayorships at 15,898,484 different venues.

FOURSQUARE CHARACTERIZATION

In this section, we discuss characterization of Foursquare users, focusing on user attributes that are publicly available in the system API and are associated with geo-referenced information, i.e., home cities, mayorships, tips and dones. Recall that the user home city and the venue location are open

NPTEL

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

Tu car

So, this was entire data of Foursquare at some point in time. So, this is the turning 13.5 million users that was mentioned earlier, a dataset contains 10 million tips, 9 million dones, 15 million mayor ships and 15 million venues. Huge data actually and that is the fun part about studying social networks also everything is huge.

(Refer Slide Time: 26:04)

of as-
On the
ms of
NPTEL

public
ships,
asso-
ocation

cations, the tool's response depends on the "quality" of the query, which, in turn, is related to the spatial granularity (e.g., street, city, state, country) of the location information provided in the query. For instance, for a query "New York", *Yahoo! PlaceFinder* returns that the query's quality is at the granularity of city, and provides the corresponding geographic coordinates, a standardized city name as well as the state and country names. *Yahoo! PlaceFinder* may also identify locations at the finer granularity of streets. Moreover, note that the use of standardized city name allows us to

<http://developer.yahoo.com/geo/placefinder/>

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

Computational
Social Science

I will give you a sense of the magnitude of the data any conclusions you are making from this, it may be very, very appropriate you should also look up. So, if any of you are interested in just as large and everything you should also look up the course that I am teaching on campus this semester called the syllabus is online computational social science, there we actually discuss, do you really need such large data to make large inferences and is it appropriate even to keep collecting more and more data to answer the same kind of questions, but that is for a different course different Google there.

(Refer Slide Time: 26:59)

identifiers. Our crawler ~~ran from Au~~ collecting a total of 13,570,060 users, ~~an~~ ~~imation~~ of the entire Foursquare com- ~~the~~ ~~rawing~~ since, reportedly, the num- ~~l~~ in Foursquare was 10 million in June ~~illion~~ in December of the same year ~~tains~~ 10,618,441 tips, 9,989,325 ~~done~~ ~~ships~~ at 15,898,484 different venues.

CHARACTERIZATION

Discuss characterization of Foursquare ~~er~~ attributes that are publicly available ~~l~~ are associated with geo-referenced in- ~~cities~~, mayorships, tips and ~~done~~. Re- ~~ie~~ city and the venue location are open ~~id~~ is not enforced by the system. In- ~~noise~~ and invalid locations. Thus, we ~~lyz~~ing the amount of valid location in- ~~set~~. Next, we analyze the use of tips, ~~s~~, focusing on the distribution of ~~as~~ around the globe. Finally, we perform a

Table 2. Quality of Geographic In-

Quality	# Users
Continent	107
Country	602,932
State	390,224
County	251,383
City	10,354,058
Neighborhood	981,139
Area of Interest/Airport	27,307
Street	326,751
Point of Interest	5,607
Coordinate	61

Thus, in order to standardize the home ~~ation~~ fields, we created a dictionary ~~c~~ ~~the~~ *Yahoo! PlaceFinder*, the *Yahoo! PlaceFinder*'s ~~get~~ tool was used to verify the validity of ~~the~~ ~~ge~~ For a given query (text), the tool either ~~a~~ ~~g~~ ~~raphic~~ data, in case the query consists ~~s~~ or an error, otherwise. For queries ~~c~~ ~~at~~ ~~ions~~, the tool's response depends ~~e~~ ~~query~~ which in turn is related to

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD



l in
pen
. In
, we
1 in-
tips,
as-
m a
s of

bl ic
ips,
sso-
ation

Thus, in order to standardize the home city and venue location fields, we created a dictionary of city names using the *Yahoo! PlaceFinder*, the Yahoo's geo-coding API.⁵ This tool was used to verify the validity of the data in both fields. For a given query (text), the tool either returns some geographic data, in case the query consists of a valid location, or an error, otherwise. For queries consisting of valid locations, the tool's response depends on the "quality" of the query, which, in turn, is related to the spatial granularity (e.g., street, city, state, country) of the location information provided in the query. For instance, for a query "New York", *Yahoo! PlaceFinder* returns that the query's quality is at the granularity of city, and provides the corresponding geographic coordinates, a standardized city name as well as the state and country names. *Yahoo! PlaceFinder* may also identify locations at the finer granularity of streets. Moreover, note that the use of standardized city name allows us to

⁵<http://developer.yahoo.com/geo/placefinder/>



c
s,
)-
n

ocations, the tool's response depends on the "quality" of the query, which, in turn, is related to the spatial granularity (e.g., street, city, state, country) of the location information provided in the query. For instance, for a query "New York", *Yahoo! PlaceFinder* returns that the query's quality is at the granularity of city, and provides the corresponding geographic coordinates, a standardized city name as well as the state and country names. *Yahoo! PlaceFinder* may also identify locations at the finer granularity of streets. Moreover, note that the use of standardized city name allows us to

⁵<http://developer.yahoo.com/geo/placefinder/>

Computational
Social Science

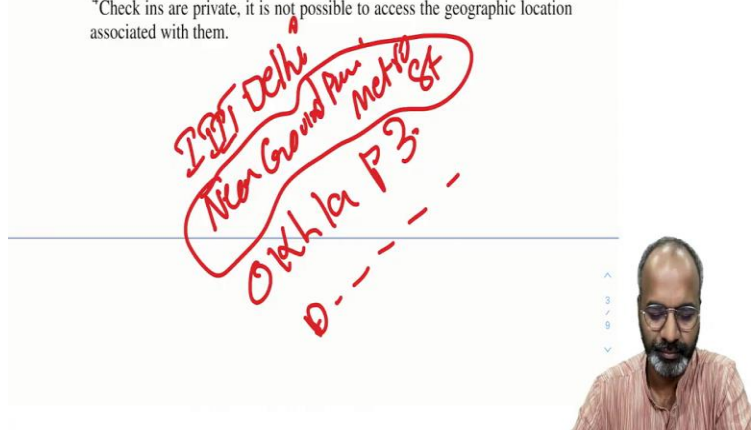




Location Information in Foursquare

We here discuss the location information available in public attributes of Foursquare users, i.e., in home city, mayorships, tips and dones.⁴ Since mayorships, tips and dones are asso-

⁴Check ins are private, it is not possible to access the geographic location associated with them.



With these data, tips, dones, checkins all that that is available, what can we do? What else can we create? So, we create a dictionary of city names using the Yahoo place finder and use the Yahoo geocoding API to go ask for a location given you checked into let us take Connaught place or Sarna Bhawan in Connaught place you give this information Sarna Bhawan Connaught place back to Google Search to Yahoo place finder.

And then it will return a location saying it is Connaught place or turn a location saying this is the lat long, which is what is written later place finder returns that the query quality is that granularity of the city and provides the corresponding geographic coordinates. Standardized city name as well as the state and the country names.

Yahoo plays finder may also identify locations at the final granularity of streets moreover note that the use of standardized city names allows us to uniquely identify the city despite the existence of multiple name variations. Unfortunately, in India it may be much harder to achieve this because for example, Triple IT Delhi addresses Triple IT Delhi, near Govind Puri Metro Station, Okhla Phase 3, Delhi blah, blah, blah.

So, now this part, how do you actually decipher? How do you disambiguate? Very hard, right? So, if the addresses are much more cleaner, it is easy to actually triangulate find the lat long, I am supposing a sort of a technical challenge here in terms of finding the location given the address, particularly in India, it is very hard to get the address given the constraint that I just mentioned.

(Refer Slide Time: 29:20)

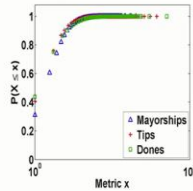


Figure 1. Cumulative Distribution of the Number of Mayorships, Tips and Dones per User.

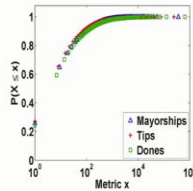
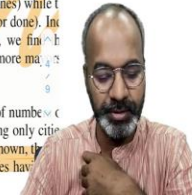


Figure 2. Cumulative Distribution of the Number of Mayorships, Tips and Dones per City.

Mayorships, Tips and Dones

In this section, we analyze the mayorships, tips and dones of users in our dataset. Since our goal is to extend the venues associated with these attributes to a user's home city, we start by showing an overview of mayorships, tips and dones among users in our dataset. We observe that almost 4.2 million users, or about 90% of all users in our dataset, have at least one of these attributes. Out of these, around 1 million have only mayorships, 367 thousand have only tips and 367 thousand have only dones, whereas 890 thousand users have all three attributes. Thus, exploiting these attributes to infer a user's home city is promising as the required information is available in a large fraction of all users. Moreover, as shown in Figure 1 and consistent with previous analyses of Foursquare [14, 17], the distributions of the numbers of mayorships, tips and dones per user are very skewed, with a heavy tail, implying that few users have many mayorships (tips or dones) while the vast majority have only one mayorship (tip or done). Indeed, for users that have one of these attributes, we find that 69% (59% and 56%) of the users have 2 or more mayorships (tips and dones).

Figure 2 shows the distributions of numbers of mayorships, tips and dones per city, considering only cities with at least one instance of the attribute. As shown, the distributions are also very skewed, with a few cities having as many as 100 mayorships, tips or dones.



thousands have only tips and 367 thousand have only dones, whereas 890 thousand users have all three attributes. Thus, exploiting these attributes to infer a user's home city is promising as the required information is available in a large fraction of all users. Moreover, as shown in Figure 1 and consistent with previous analyses of Foursquare [14, 17], the distributions of the numbers of mayorships, tips and dones per user are very skewed, with a heavy tail, implying that few users have many mayorships (tips or dones) while the vast majority have only one mayorship (tip or done). Indeed, for users that have one of these attributes, we find that 69% (59% and 56%) of the users have 2 or more mayorships (tips and dones).

Figure 2 shows the distributions of numbers of mayorships, tips and dones per city, considering only cities with at least one instance of the attribute. As shown, the distributions are also very skewed, with a few cities having as many as 100 mayorships, tips or dones.

Mayorships, Tips

of multiple

various indicat-

Next, we analyzed the correlation between the number of mayorships, tips and dones per city. We found that there is a high correlation between the number of mayorships and the number of tips in cities with a significant number of



Let us look at some analysis, figure 1 so, that is this, So, this is figure 1, this is figure 2 so, this is showing the cumulative distribution number of mayor ships, tips and dones per user cumulative distribution number of mayor ships tips and dones per city. So, essentially the goal why do we have to draw this graph, these graphs are necessary in terms of saying what is the pattern of the data that you have collected, particularly in this one, it is like this power law, it is like this large number of check ins are created by small number of people.

And this behavior is consistent across many social networks. Facebook, many people do less of posts, but a small set of proportion people do a large chunk of post power law, Pareto

Principle, there are many names for this behavior. In drawing this graph and having a confirmation that this is the how the data is, is very, very useful that is what this data says.

Shown in for consistent with previous analysis of Foursquare, the distribution of the numbers of mayour, tips and dones per user are very skewed with a heavy tail, implying that few users have many mayor ships tips and dones while the vast majority have only one mayor ship tip or done. Indeed, for users that are one of these attributes, we find that 69 percent of users have two or more mayor ships tips and dones.

(Refer Slide Time: 31:19)

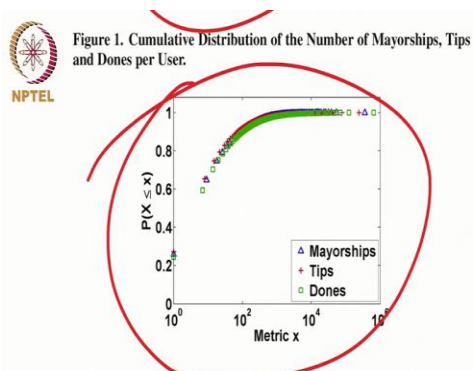


Figure 1. Cumulative Distribution of the Number of Mayorships, Tips and Dones per User.

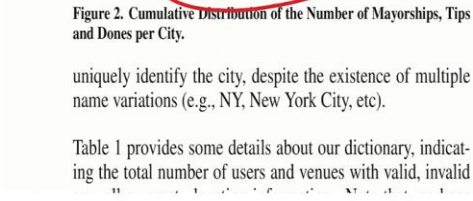


Figure 2. Cumulative Distribution of the Number of Mayorships, Tips and Dones per City.

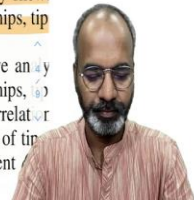
uniquely identify the city, despite the existence of multiple name variations (e.g., NY, New York City, etc).

Table 1 provides some details about our dictionary, indicating the total number of users and venues with valid, invalid

exploiting these as the requirement of all users. Moreover, with previous analysis, the distributions of the numbers of mayorships, tips and dones per user are very skewed, with a few users having many mayorships, tips and dones while the vast majority have only one mayorship tip or done. Indeed, for users that are one of these attributes, we find that 69 percent of users have two or more mayorships tips and dones.

Figure 2 shows the distributions of numbers of mayorships, tips and dones per city, considering only cities with at least one instance of the attribute. As shown, the distributions are also very skewed, with a few cities having as many as 100 mayorships, tips or dones.

Next, we analyzed the correlation between the number of mayorships, tips and dones per city. We found that there is a high correlation between the number of mayorships and the number of tips across cities, with a Spearman's correlation coefficient ρ [21] equal to 0.78. Similarly, the correlation is also high between the number of mayorships and the number of dones ($\rho = 0.72$). Moreover, we found that the cities with the largest numbers of mayorships tend also to have large numbers of tips and dones, although some interesting differences are worth noting. For instance, mayorships are



NPTEL

, Tips

tiple

licat-
valid
haps
pro-
r, in
users
nues

and 69% of the users have 2 or more mayorships (tips and dones).

Figure 2 shows the distributions of numbers of mayorships, tips and dones per city, considering only cities with at least one instance of the attribute. As shown, the distributions are also very skewed, with a few cities having as many as 100 mayorships, tips or dones.

Next, we analyzed the correlation between the number of mayorships, tips and dones per city. We found that there is a high correlation between the number of mayorships and the number of tips across cities, with a Spearman's correlation coefficient ρ [21] equal to 0.78. Similarly, the correlation is also high between the number of mayorships and the number of dones ($\rho = 0.72$). Moreover, we found that the cities with the largest numbers of mayorships tend also to have large numbers of tips and dones, although some interesting differences are worth noting. For instance, mayorships are





ber of dones ($\rho = 0.72$). Moreover, we found that the cities with the largest numbers of mayorships tend also to have large numbers of tips and dones, although some interesting differences are worth noting. For instance, mayorships are more concentrated in Southeast Asia, in cities like Jakarta, Bandung and Singapore, which are the top three cities in number of mayorships, jointly having more than 500,000 mayorships. Tips, in turn, are concentrated in different locations around the Earth: the top three cities in number of tips are New York, Jakarta and São Paulo, with a total of 600,000 tips. Dones, on the other hand, tend to be concentrated in venues in the United States, in cities like New York, Chicago and San Francisco, which jointly received around 1 million dones.



We note that, although other studies [1, 8, 11] have exploited textual features to analyze user location, we here chose not to exploit the tip's content as they are often targeted towards



So, it is very small, the same thing for figure two, this is figure 2. This is per city figure 2 the distribution of the number of mayors and tips and dones per city considering only cities with at least one instant of the attribute which is your check ins, tips and dones mayors are available, as shown in the distributions are also very skewed with a few cities having many as 100 mayor ships tips and dones.

So, this is the power law behavior. Any analysis that you do in social it is good to there is also arguments that you will see online about, oh is this even a power law behavior is it even necessary the social network power behavior itself is wrong all that but having a verification of what kind of pattern distribution your data is, is very useful.

So, for instance, mayor ships are more concentrated, So, now this is looking at where the mayor ships are, I will show you a graph full map where this is displayed. For instance, mayor ships are concentrated in Southeast Asia in cities like Jakarta, Bandung and Singapore, which are the top three cities in the number of mayor ships jointly having more than 500,000 mayor ships just three cities having 500,000 memberships,

Tips and dones are concentrator in different locations around the Earth. The top three cities in the number of tips are New York, Jakarta and Sao Paulo with a total of 600,000 tips, dones on the other hand, tend to be concentrated menus in the United States in cities like New York, Chicago and San Francisco. Together they have a million dones. Again, the same thing few cities are having large number of dones, tips and mayor ships.

(Refer Slide Time: 33:29)

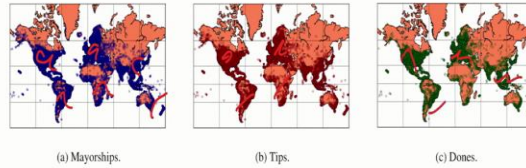


Figure 3. Global Distribution of Mayorships, Tips and Dones.

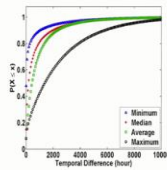


Figure 4. Cumulative Distribution of Time Interval Between Consecutive Tips/Dones Posted per User.

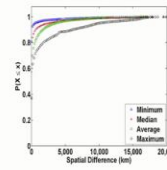


Figure 5. Cumulative Distribution of Displacements Between Consecutive Tips/Dones Posted per User.



Figure 4. Cumulative Distribution of Time Interval Between Consecutive Tips/Dones Posted per User.

Figure 5. Cumulative Distribution of Displacements Between Consecutive Tips/Dones Posted per User.

We only consider attributes associated with venues that have valid cities (validated by *Yahoo! PlaceFinder*) as location. Figure 3 shows these distributions in maps of the globe, with each point representing a city with venues with at least one mayorship, tip or done.⁶ As the maps show, Foursquare venues are spread all over the world, including remote places such as Svalbard, an archipelago in the Arctic Ocean, with coordinates (78.218590,15.648750). Moreover, all three maps are very similar, with most incidences of points in America, Europe and Southeast Asia. The distribution of mayorships, shown in Figure 3(a), is denser, with a total number of unique cities (79,194) much larger than in the distributions of tips and dones, which cover a total of 54,178 and 30,530 unique cities, respectively. The somewhat sparser tip map (Figure 3(b)) indicates that there are many cities, particularly in Canada, Australia, central Asia and Africa, where, despite the existence of venues and mayors, users do not post tips. The distribution of dones, shown in Figure 3(c), reveals an even sparser map, with most activity concentrated in touristic or developed areas, such as USA, western Europe and southeast Asia. We note that a similar map was produced for check ins in [2]. Besides both datasets were collected at different times, we can see that their main areas of concentration overlap.

We start by investigating the frequency of tips and/or mark previous tips as done (the time interval between consecutive tips or a done) of the same user. Thus, with at least two activities, covering users. We summarize user activity by median, average and maximum inter-activity times. The cumulative distributions computed for all considered users. A short period of time between consecutive tips/dones (around 1 hour) is common, with almost all users doing long periods of time between consecutive tips/dones. For instance, around 50% of the users have a maximum inter-activity time of at least 450 hours (roughly a week).

Next, we analyze the displacement between consecutive tips and/or dones. We consider only users with at least two activities associated with these activities with "quality" of city level or



Here is the world map I said so, this is showing your world map this is the mayor ship. So, these are all the blue dots are all the major ships that we have data for. The darker red dots here are the ones that are for tips in green is for the dones here. A quite sort to say spread out but concentrated on some parts of the world which is understandable. You can see even though many of us do not use Foursquare you can see all almost in all the three we have India.

So, please find out if you are interested in exploring Foursquare try out and let us let us know. What do you find? How do you find using it?

(Refer Slide Time: 34:29)

Figure 3 shows these distributions in maps of the globe, with each point representing a city with venues with at least one mayorship, tip or done.⁶ As the maps show, Foursquare venues are spread all over the world, including remote places such as Svalbard, an archipelago in the Arctic Ocean, with coordinates (78.218590,15.648750). Moreover, all three maps are very similar, with most incidences of points in America, Europe and Southeast Asia. The distribution of mayorships, shown in Figure 3(a), is denser, with a total number of unique cities (79,194) much larger than in the distributions of tips and dones, which cover a total of 54,178 and 30,530 unique cities, respectively. The somewhat sparser tip map (Figure 3(b)) indicates that there are many cities, particularly in Canada, Australia, central Asia and Africa, where, despite the existence of venues and mayors, users do not post tips. The distribution of dones, shown in Figure 3(c), reveals an even sparser map, with most activity concentrated in touris-

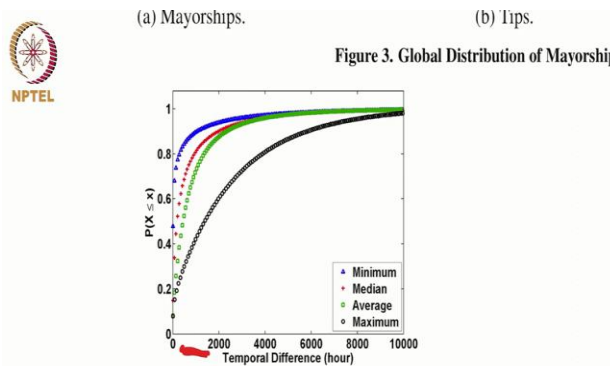


Figure 4. Cumulative Distribution of Time Interval Between Consecutive Tips/Dones Posted per User.

We only consider attributes associated with venues that have valid cities (validated by Yahoo! PlaceFinder) as location. Figure 3 shows these distributions in maps of the globe, with each point representing a city with venues with at least one

Figure 5

We s... tips and in



So, as the map shows, same world map Foursquare venues are spread all over the world, including remote places such as Svalbard, archipelago in the Arctic Ocean with coordinates is a bit when doing this it is interesting now you will find some locations look at all this look at these kinds of locations I do not know whether there are any land also in these places.

Which is interesting but that is we just point, we are just pointing it out here. 50 percent of the users have an average interactivity time. So, this is now looking at the graph of four and five, four is cumulative distribution of time interval between consecutive tips dones and posted per

user which is this is analyzing if PK does a post now PK does a post in another one hour later that is what this is capturing.

So, this would be time difference between two different ports. So, this would help us to know how frequently I am actually using the platform, frequently doing an activity in the platform to be more precise.

(Refer Slide Time: 36:01)

The slide features a graph titled '(c) Dones.' showing the cumulative distribution of spatial differences between consecutive tips/dones. The x-axis is 'Spatial Difference (km)' from 0 to 20,000, and the y-axis is 'P(X ≤ x)' from 0 to 1. Four curves represent Minimum, Median, Average, and Maximum values. The Minimum curve rises sharply, while the others rise more gradually. The NPTEL logo is in the top left, and the International Institute of Information Technology Hyderabad logo is in the top right. A video feed of a speaker is visible on the right side of the slide.

Figure 5. Cumulative Distribution of Displacements Between Consecutive Tips/Dones Posted per User.

We start by investigating the frequency at which users leave tips and/or mark previous tips as done. We do so by analyzing the time interval between consecutive activities (be it a

... minimum, average and maximum inter-activity times. Figure 7 shows the cumulative distributions of these four measures computed for all considered users. We note that the distribution of minimum inter-activity times is very skewed towards short periods of time, with almost 50% of the users posting consecutive tips/dones 1 hour apart. However, on average, median and maximum, users do tend to experience very long periods of time between consecutive tips and dones. For instance, around 50% of the users have an average inter-activity time of at least 450 hours, whereas around 80% of the users have a maximum inter-activity time above 167 hours (roughly a week).

Next, we analyze the displacement between two venues visited in sequence by the user, as indicated by consecutive tips and/or dones of the user. For this analysis, we consider only users with at least two activities, provided that the venues associated with these activities have valid locations, with "quality" of city level or finer granularity. Our dataset contains almost 1.5 million users in this group. For these

And then this one is cumulative distribution of displacement between So, now I know that because I did a check in after three days from the last 1am I how far these two locations were

two venue were, that is what this graph is showing. For instance, over 50 percent of the users have an average interactivity time of at least 450 hours.

Whereas around 80 percent of the users have a maximum interactivity time of around 167 hours, which is roughly 178 hours is 178 something around that is number of hours 24 to 7 so, yeah, 168 hours per week. So, that is roughly a week. So, therefore, there is some pattern in the maybe there is some pattern in the user checking in on Foursquare.

(Refer Slide Time: 37:09)

users, we computed the displacements between consecutive
ps/dones by taking the difference between the coordinates
of the associated venues. Once again, we summarize user
activity computing the minimum, median, average and max-
imum displacement per user. Figure 5 shows the distribu-
tions of these measures for all analyzed users. Around 36%
of the users have average and maximum displacements of
0 kilometer, indicating very short distances (within a few

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

5

So, this is figure 5, the one that we saw here, this is figure 4, this one, and now what we're going to see is actually figure 5. So, this is a figure 5 shows the distribution of these measures of for all the analyzed users around 36 percent of the users have an average maximum displacement of . kilometers, you know, earlier I said there may be that I am in the same location I am checking in the same location even after a week apart.

So, that is 0 kilometers 36 percent, which is my error rate can be pretty high, pretty low rate in terms of saying that this is a home this is a office is something a place that they like all of that. Indicating very short different short distances, within a few meters.

(Refer Slide Time: 38:02)



Figure 6. Distribution of Returning Times.

Moreover, 70% of the users have an average displacement of at most 150 kilometers, which could be characterized as within the metropolitan area of a large city. Also 60% of the users have a maximum displacement of at most 100 kilometers, possibly the distance between neighboring cities. Thus, overall, consecutive tips/dones of a user are often posted at places near each other. However, there are exceptions. About 10% of the users have a maximum displacement of at least 6,000 kilometers.⁷

Finally, we analyze how often users return to the same venue for tipping or marking tips as done. That is, we analyze the returning times, defined as the time interval between consecutive tips/dones posted at the same venue by the same user. This analysis is focused on 813,607 users, who have at least two tips/dones in the same venue, and cover more than 3 million returns. We here choose to show the distribution

of evidence present pl a user o frequent reveal pl users are also pro although conjectur ical plac We note assumpti cussed in the users least 6,00

As a first s ple app attribut hom



Moreover, 70 percent of the users have an average distance, displacement of at least 150 kilometers, which could be characterized as within a metropolitan city of a metropolitan area of a large city. Also 60 percent of the users have a maximum displacement of at most 100 kilometers, possibly the distance between neighboring cities. Over 10 percent of the users have a maximum displacement about 6000 kilometers. And this could be that look, I travelled from Delhi to Kanyakumari or Delhi to Dubai I checked into the airport in Delhi I checked into the airport in Dubai 6000 kilometers.

(Refer Slide Time: 38:59)



them per user first, so as to compare our results against previous findings of check in patterns [2]. Figure 6 shows the distribution, focusing on returning times under 360 hours, which account for 69.7% of all measured observations. The curve shows clear daily patterns with returning times often being multiples of 24 hours, which is very similar to the distribution of returning times computed based on check ins [2]. We note, however, that 50% of the measured returning times are within 1 hour, which cannot be seen in the Figure as its y-axis is truncated at 1% so that the rest of the curve could be distinguished. Moreover, out of these observations, 90% of them are at most 10 minutes. Thus, returning times, in general, tend to be very short. If we analyze the behavior per user (omitted more details, due to space constraints), we note that most users have very short minimum returning times, which is below 1 hour for 62% of the users. However, consistently with results in Figure 4, on average, median and maximum, users do tend to experience longer returning times. For instance, 52% of the users have average

mach simpl poten

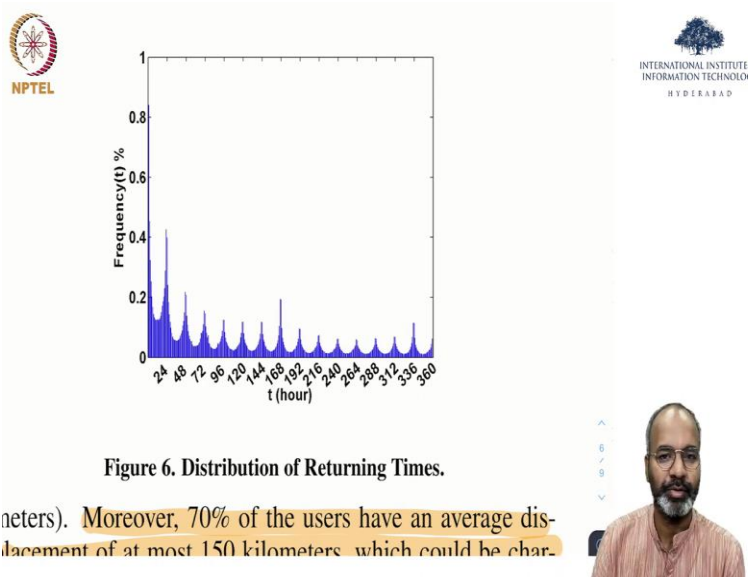


We co the at only t locati tions Mayc only attrib able t ferenc Tip of the the c ence m



So, I hope this is giving you a sense of what the platform is in details of the platform. The curve shows clear daily patterns with returning times, often multiples of 24, which is a very similar distribution of returning times computed based on checkin. So, what this graph is showing, and this is talking about figure 6, I will show you a figure 6 in a second.

(Refer Slide Time: 39:15)



So, distribution of return times. So, this is a very interesting graph I kind of alluded earlier, I made that comment, I had this graph in mind that is you see a pattern this pattern is happening for 24 hours, every 24 hours there is a spike in the location that you are checking in. Every 24 hours, the check ins of for the location, not for the location for the user, the locations that they check in, is actually higher and higher every time.

You can make some guesses, I am going to office every day from office, I am coming back home, I came back home today at 6pm, tomorrow also roughly, I will come back around 6pm I get into office at 10am today, potentially, I will get into office 10am tomorrow. So, that is a 24 hour cycle between the check ins in that location that is what this is highlighted.

(Refer Slide Time: 40:21)

NPTEL

vious findings of check in patterns [2]. Figure 6 shows the distribution, focusing on returning times under 360 hours, which account for 69.7% of all measured observations. The curve shows clear daily patterns with returning times often being multiples of 24 hours, which is very similar to the distribution of returning times computed based on check ins [2]. We note, however, that 50% of the measured returning times are within 1 hour, which cannot be seen in the Figure as its y-axis is truncated at 1% so that the rest of the curve could be distinguished. Moreover, out of these observations, 90% of them are at most 10 minutes. Thus, returning times, in general, tend to be very short. If we analyze the behavior per user (omitted more details, due to space constraints), we note that most users have very short minimum returning times, which is below 1 hour for 62% of the users. However, consistently with results in Figure 4, on average, median and maximum, users do tend to experience longer returning times. For instance, 52% of the users have average returning times of at least 168 hours.

simp
potel



We c
the a
only
locat
tions
May
only
attrit
able
ferer
Tips
of
the
ene
m

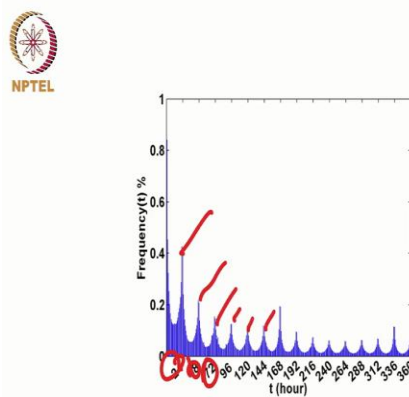


Figure 6. Distribution of Returning Times.

eters). Moreover, 70% of the users have an average displacement of at most 150 kilometers, which could be characterized as within the metropolitan area of a large city. Also

We he
ology
sults a

Methc
The k
have i
cation
that th
of evi
rese
a us
fre
reve
us



The curve shows clearly daily patterns with 24 hours of difference 24 hours with multiple so, if you see here, that is what is right, 24, 48, 172 so, all of this.

(Refer Slide Time: 40:41)

NPTEL
variations, 90% of them are at most 10 minutes. Thus, returning times, in general, tend to be very short. If we analyze the behavior per user (omitted more details, due to space constraints), we note that most users have very short minimum returning times, which is below 1 hour for 62% of the users. However, consistently with results in Figure 4, on average, median and maximum, users do tend to experience longer returning times. For instance, 52% of the users have average returning times of at least 168 hours.

INFERRING USER'S HOME LOCATION

In this section we investigate whether one can infer, with reasonable effectiveness, the location where a user lives based only on information that is publicly available on her Foursquare profile page, notably the lists of mayorships, tips and dones.

⁷Note that the maximum displacement between two points in the Earth is the distance between antipodes (two diametrically opposed points) that is about 20,000 kilometers.

NPTEL
that more sophisticated methods could be applied such as classification algorithms (e.g., k-nearest neighbor) and other machine learning techniques [8, 11, 1]. Instead, we chose a simple majority voting approach as it allows us to assess the potential for effective inferences of this type in Foursquare.

We consider seven inference models which differ in terms of the attributes used for inference. The *Mayorship* model uses only the locations of the mayorships to infer the user's home location. Similarly, the *Tip* and *Done* models use only locations of tips and of dones, respectively. The *Mayorship+Tip*, *Mayorship+Done*, *Tip+Done* models use information from only two attributes, whereas the *All* model takes all three attributes jointly. By comparing alternative models, we are able to assess the potential of each attribute as source of inference. Moreover, recall that, as discussed in Mayorships, Tips and Dones section, there are non-negligible numbers of users that only have one or two of the attributes. Thus, the combination of multiple attributes may enable the infer-

attributable to
ference
Tips and
of user
the cor
ence fo
mainly
sider it



To evaluate
format
truth. /
this att
do e



⁸Although
about or
that



52 percent of the users have an average return time of at least 168 hours, which I said earlier also which is about a week. So, there are many strategies, we could have actually used to find the location, exact location, which is if you are looking at how many of your friends are in the same location, we could have actually derived whether this is the location of your home office all that your own check ins, for sure majority voting is what we ended up actually using but we could actually derive, we could actually add more information to this analysis by knowing that your friends are also from the same location.

Your friends are also in the close vicinity, the same user is actually coming back, going back and forth probably to you and your spouse, you and your parents, things like that.


(Refer Slide Time: 41:46)

ity, as validated by various researchers.

NPTEL In our evaluation, we group users into three classes. *Class 0* consists of users who have a single activity, either a mayor-ship, a tip or a done. In this case, the unique choice is to set the user's home location equal to that of her activity. *Class 1* consists of users who have multiple activities with a predominant location across them. For these users, the inferred location matches the most often location of their activities. *Class 2*, in turn, consists of users with multiple activities in which there is no single location that stands out (i.e., there are ties). Our current inference approach cannot be applied to *Class 2* users.

Thus, we evaluate the proposed models by assessing their accuracy on users of both *Class 0* and *Class 1*. The accuracy corresponds to the percentage of correctly inferred locations out of all users of each class. Moreover, we also report the overall accuracy of each model, considering all users that are eligible for inference by the given model (i.e., users who

Fig
De
We
res
of
th
the

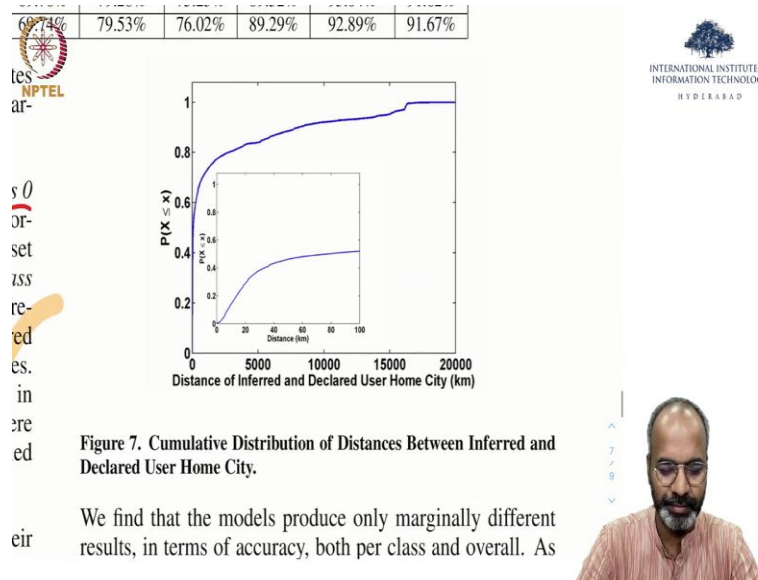


So, the details of how the different types of errors different types of inferences may be grouped users with into three classes, class zero, consists of users who have a single activity, either a mayor ship, or a tip or a done because again, if you look at the data that we have the data can be of per user, that could be many different pieces of information, mayor ship, tip, dones so, we cannot put all of them together.

So, at least one data if for a user is there, he or she goes into class zero. In this case, the unique choice is to set the user's home location equal to our activity. Class one consists of users who have multiple activities with the predominant location across them. I see one lat long coming back again and again. For these users in for location matches the most often location of their activity.

Class two in turn consists of users with multiple activities in which there is no single location that stands out or current inference approach cannot be applied to class to users, because there is many we cannot it is like a fuzzy we cannot really infer and we were more interested in, let us just precisely get the location of the user.

(Refer Slide Time: 43:19)



That is the most interesting graph, you will find in this paper, which is, actually there is one more graph cumulative distribution of distances between inferred and the declared users home city. So, this is a distance between what is there and what was inferred.

(Refer Slide Time: 43:41)

To better understand the models' errors, we computed for each incorrect inference the distance between the inferred city given by the *All* model and the declared user home city. Figure 7 shows the distribution of these distances. We found that around 46% of the distances are under 50 kilometers, which is a reasonable distance between neighboring (twin)

class 0
 mayor-
 to set
 Class
 a pre-
 ferred
 cities.
 ties in
 there
 applied

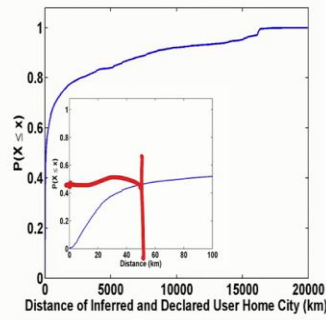


Figure 7. Cumulative Distribution of Distances Between Inferred and Declared User Home City.

their
 accuracy
 actions

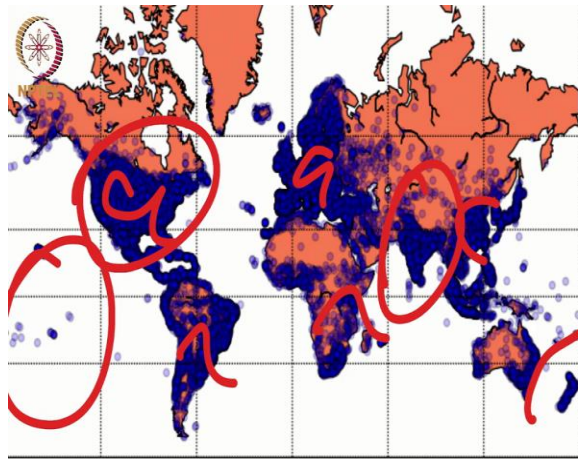
We find that the models produce only marginally different results, in terms of accuracy, both per class and overall. As expected, mayorships are the best single attribute to infer



cities. Thus, combining these results with the correct inferences produced by our model, we find that we can correctly infer the city of around 78% of the users within 50 kilometers of distance.

We now turn our attention to the inference of a user's home state, whose results are also shown in Table 3. We note that, in comparison with the home city inference, all models improved for home state inference, reaching an overall accuracy around 75%. Once again, mayorships arise as the single attribute that produces the highest accuracy, for home state inference, followed by tips and donees. Nevertheless all models lead to very similar accuracies, both per class and overall. Thus, once again, due to the larger user coverage, the *All* model is able to correctly infer the home state of the





(a) Membership

So, we found that around 46 percent of the distances around 50 kilometers, So, this one is the 50 kilometer graph. So, that is the 46 percentage mark I am guessing you know how to read a I mean I think we also discussed this before please remember how to read a CDF graphs all that 46 percent is about 50 kilometers which is reasonable distance between neighboring cities that is combining these results with the correct inferences produced by our model, we find that we can correctly infer the city around 78 percent of the users within 50 kilometers of distance.

So, it is interesting that I meaning if the location based information is publicly available or not available only to friends, it can be actually used to infer a lot of information from that that is what the purpose of this paper was to argue that look in a small distance one can argue that 50 kilometers itself is big.

But I am sure if I am sure today if you get more than so, I think the 50 kilometers could also have been because the data that we had is quite diverse probably if you just turn this around and say, okay, let us just analyze this only for the US, probably the resistance may have been much lesser, I guess, because why I am making that inference is if you look at the if you look at the mayor ship here, it is like crowded know, lots of data is there.

So, if you only use US to do the analysis, probably the distance may have been much lesser, if you are interested in that you should try. So, that is what this paper is the idea, the idea of this paper was to show that we can actually infer where you live, just by taking this publicly available information.

(Refer Slide Time: 45:55)

TP-SM_MV_AG_JA_PK_VA_Pinsoda_2012
2022 at 4:24 PM

NPTEL

**INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD**

Beware of What You Share: Inferring Home Location in Social Networks

Tatiana Pontes*, Gabriel Magno*, Marisa Vasconcelos*, Aditi Gupta¹,
Jussara Almeida*, Ponnurangam Kumaraguru¹, Virgilio Almeida*

*Universidade Federal de Minas Gerais, Brazil
{tpontes,magno,marisa,jussara,virgilio}@dcc.ufmg.br

¹Indraprastha Institute of Information Technology, India
{aditi.g.p.k}@iitd.ac.in

Abstract—In recent years, social media users are voluntarily making large volume of personal data available on the social networks. Such data (e.g., professional associations) can create opportunities for users to strengthen their social and professional ties. However, the same data can also be used against the user for viral marketing and other unsolicited purposes. The invasion of privacy occurs due to privacy unawareness and carelessness of making information publicly available. In this paper, we perform a large-scale inference study in three of the currently most popular social networks: Foursquare, Google+ and Twitter. Our work focuses on inferring a user's home location, which may be a private attribute, for many users. We analyze whether a simple method can be used to infer the user home location using publicly available attributes and also the geographic information associated with locatable friends. We find that it is possible to



data associated with location information could be even more invasive [20]. The collation of public location based attributes of a user aggregated over time may reveal her behavioral patterns and habits, emphasizing her preferences. Despite the privacy threats of sharing location, this is arising as a common behavior among users in Foursquare, which is currently the most popular LBSN, and even on the traditional OSNs, such as Google+ and Twitter.

Motivated by the possible privacy breaches due to the increased sharing of location information in social networks, here we perform a large-scale study on inferring the user home location in three of the currently most famous systems, namely Foursquare, Google+ and Twitter. Foursquare is a LBSN

making large volume of personal data available on the social networks. Such data (e.g., professional associations) can create opportunities for users to strengthen their social and professional ties. However, the same data can also be used against the user for viral marketing and other unsolicited purposes. The invasion of privacy occurs due to privacy unawareness and carelessness of making information publicly available. In this paper, we perform a large-scale inference study in three of the currently most popular social networks: Foursquare, Google+ and Twitter. Our work focuses on inferring a user's home location, which may be a private attribute, for many users. We analyze whether a simple method can be used to infer the user home location using publicly available attributes and also the geographic information associated with locatable friends. We find that it is possible to infer the user home city with a high accuracy, around 67%, 72% and 82% of the cases in Foursquare, Google+ and Twitter, respectively. We also apply a finer-grained inference that reveals the geographic coordinates of the residence of a selected group of users in our datasets, achieving approximately up to 60% of accuracy within a radius of six kilometers.

Keywords-Location; Privacy; Social Networks; Location Infer-

invasive of a us patterns privacy behavior most po as Goog
Motiv increase here we location Foursqu geared t check to the also loc at speci with



The follow up paper, which I am not going to discuss in detail, but I will let actually you to read and come back if you need any help in understanding the paper it says beware of what you share inferring home locations and social, very similar strategy very similar networks Foursquare was used here. But Foursquare, Google Plus and Twitter, all the three networks was used in this study again, one graph, which will infer is this one.

(Refer Slide Time: 46:28)

of 3.25% of the users, which is expected, since we are using attributes of places where the user studied or worked. Thus, these higher distances errors for Google+ suggest that people may live and work in different cities.

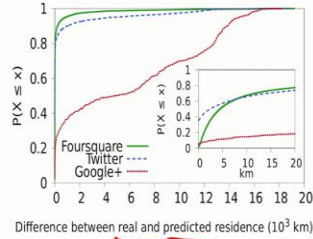


Fig. 5. Distance Between Real and Predicted User Residence.

VI. CONCLUSIONS

In this paper, we addressed the problem of privacy invasion using publicly shared attributes on three popular social networks: Foursquare, Google+ and Twitter. For each system, we considered models based on attributes of the medium to

in *GIS*, 15(6):735-740, 2007.

[8] R. Gross and A. Acemoglu. Social Networks. In *Handbook of Economic Geography*, pages 1-24. North-Holland, 2009.

[9] B. Hecht, L. Hong, and S. E. Heart. The Dynamic Heart: the Dynamic Heart of a City. *CHI '11*, 2011.

[10] I.-F. Lam, K.-T. Chan, and S. C. Chen. Social Network Analysis in Social Network Mining. *ACM SIGSPATIS*, 2010.

[11] N. Li and G. Chen. Social Network Mining. *ACM SIGSPATIS*, 2010.

[12] G. Magno, G. Corbelli, and M. L. S. Costa. On the Block: Exploring the Block. *ACM SIGSPATIS*, 2010.

[13] J. Mahmud, J. Nie, and S. E. Heart. Inferring Home Location from Social Networks. *ACM SIGSPATIS*, 2010.

[14] A. Mislove, B. Viswanath, and S. E. Heart. You Know: Inferring Home Location from Social Networks. *WSDM '10*, 2010.

[15] A. Noulas, S. Scellato, and C. L. S. G. L. Social Network Analysis of Geographic User Location. *ACM SIGSPATIS*, 2010.

[16] J. Pesce, D. Casas, and S. E. Heart. Media Using Privacy. In *Proc. PSOSM '10*, 2010.

[17] T. Pontes, M. V. C. S. We Know Where You Live: Inferring Home Location from Social Networks. *ACM SIGSPATIS*, 2010.

[18] T. Sakaki, M. Ohno, and S. E. Heart. Real-Time Event Detection from Twitter. *ACM SIGSPATIS*, 2010.



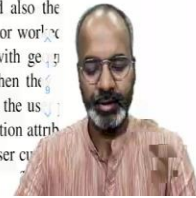
popular social networks: Foursquare, Google+ and Twitter. Our work focuses on inferring a user's home location, which may be a private attribute, for many users. We analyze whether a simple method can be used to infer the user home location using publicly available attributes and also the geographic information associated with locatable friends. We find that it is possible to infer the user home city with a high accuracy, around 67%, 72% and 82% of the cases in Foursquare, Google+ and Twitter, respectively. We also apply a finer-grained inference that reveals the geographic coordinates of the residence of a selected group of users in our datasets, achieving approximately up to 60% of accuracy within a radius of six kilometers.

Keywords-Location; Privacy; Social Networks; Location Inference; Foursquare; Google+; Twitter

I. INTRODUCTION

Online Social Networks (OSN) are one of the most popular web applications amongst Internet users. Initially, they were designed to connect close friends, but gradually new social networks were created with diverse purposes attracting users with different needs and reasons to sign up to this kind of system. Thereby, users are voluntarily making more personal information available such as their favorite places to visit, professional interests, personal views and reviews of company

Motivated by the increased sharing of geographic data, here we perform a large scale study of user location in three of the most popular social networks: Foursquare, Google+ and Twitter. Foursquare is geared towards sharing location check ins, which are to the most frequent users. Google+ also leave notes (tips) at specific venues, and Twitter present geographic data. A user's address, and also the location (geographic) where the user has studied or worked, can be tagged with geographic coordinates. When a user shares data, the user's home location attribute where the user cr

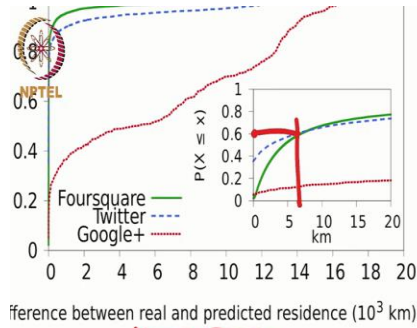


The last graph that I showed you in the last paper is exactly this difference between real and predicted, inferred and the real, was what was used there. So, in Foursquare, there is a high chance of actually getting your location more precisely, particularly because of its location based itself.

So, that is what location based networks can be used for. So, the one conclusion here was, we also applied the fine grained inference and reveals that geography coordinates of residence of a selected group of users in our data set approximately 60 percent of the users within a radius of 6

kilometers 60 percent within 6 kilometers, which, which I think you can go back to this graph again.

(Refer Slide Time: 47:26)



Distance Between Real and Predicted User Residence.

VI. CONCLUSIONS

In this paper, we addressed the problem of privacy invasion by publicly shared attributes on three popular social media platforms: Google+, Twitter, and Foursquare.

[10] I.-F. ...
in Sc
[11] N. Li
Netw
[12] G. M
on th
[13] J. M
Infer
[14] A. M
You
WSD
[15] A. N
of Ge
[16] J. Pe
Medi
In
[17] T.
We
P



So, 6 kilometers if you refer to draw here, it will be. So, that is what is for location based social network, particularly privacy connected to it. That is another paper I will send week 10, thank you for again listening for listening to this lecture hopefully you are enjoying the class. Let us please reach out if you need any help in terms of understanding the content.

The exams are also coming closer. So, if you need if you need any specific sessions or something from me, let me know doubt clearing sessions I will be happy to actually set them up and we can actually interact. Take care. Good luck.