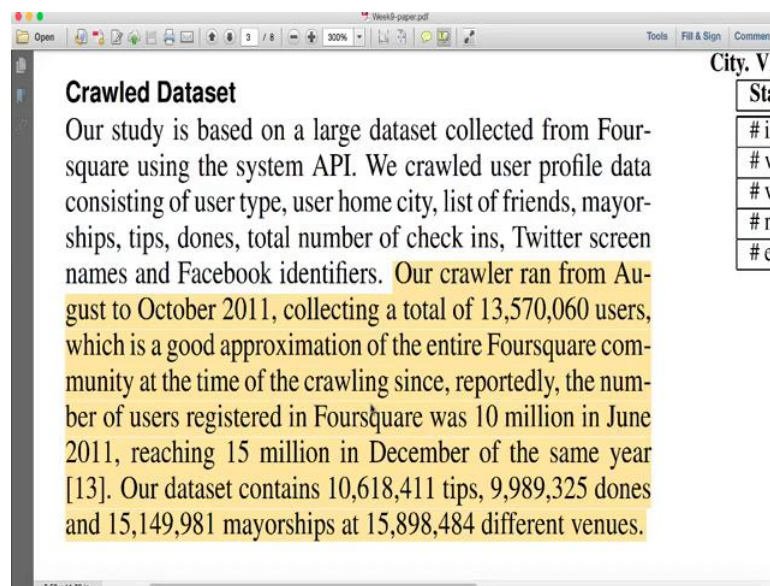**Privacy and Security in Online Social Media**
Prof. Ponnurangam Kumaraguru ("PK")
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Week - 9.2
Lecture - 30
Privacy in Location Based Social Networks Part 2

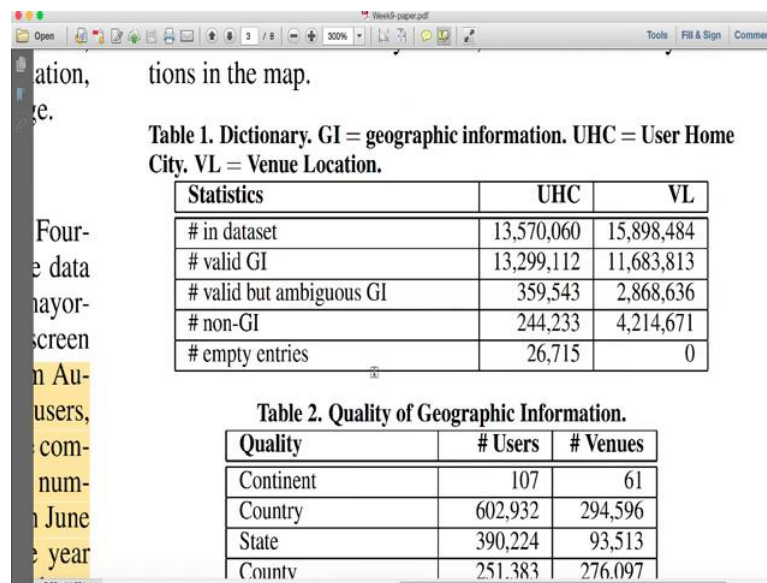Now, let us look at the dataset that was collected.

(Refer Slide Time: 00:12)

So, the dataset is about total of about 13 million users collected ran from August to October 2011, the total of 13 million users. It is almost close to the entire Foursquare in terms of the number users that are using it, and the total dataset contains 10 million tips and what we are interested in the data particularly the questions that we are asking as I said, we are interested in mostly the tips, dones and mayorships, because that is the information that is publicly available.

What can you use this information for in terms of the actually getting the location of the particular user. It is the total number of tips that are available in the dataset is about 10 million, total number of dones is about 9 million, and mayor ship is about 15 million and different venues that are available are about 15 million again.

(Refer Slide Time: 01:56)

So, let us just look at just characterization of this data, which is what kind of in generally about the dataset that is available. So, we will look at every table and every figure in this paper. Also here if you look at the number in the dataset is 13 million UHC. UHC stands for user home city; VL stands for venue location, the number of venues that are available; and GI stands for geographic information right.

Of course, there is going to be some information, some locations which are not going to be valid right for example, something in the middle of sea, you are not going to get any location, and there are locations that may be generated which is somebody's heart right, h e a r t. So, these kind of locations has to be removed that is what happened between number in the dataset and valid GI; valid, but ambiguous which is it is valid, but we are not able to figure out the exact location, reverse look up, and find out the location that falls into the third row.

Non- geographic information and empty entries, so essentially the dataset was pruned to get data which the researchers can actually use to do the analysis right. This is what even you would do for the home works that you did you collected some data, but you probably did not do the way to actually prune the data to get more accurate, more specific data.

Also the quality of geographic information is continent, countries, state and this information that is available in this dataset is number of users, number of venues. Country, state, county, city, neighborhood, area of interest or airport, street, point of interest and coordinates. So, all this pieces of information you will get in your json when you collect data from foursquare. And this was basically pruned to get more quality data which can be used for analysis, is that making sense. So, these are called exploratory data analysis, here we just explaining the data itself describing the data in terms of what is available in the data that was collected.

Multiple tools were used. So, let me just show you one of them which are developer dot yahoo dot com slash geo slash placefinder. These kind of tools lets you actually reverse look up a place and find out where they are in the map; if you give them location, it can actually give you the latitude, longitude even the other way round, you give the latitude longitude it will give you the location.

(Refer Slide Time: 03:35)

Figure 1. Cumulative Distribution of the Number of Mayorships, Tips and Dones per User.
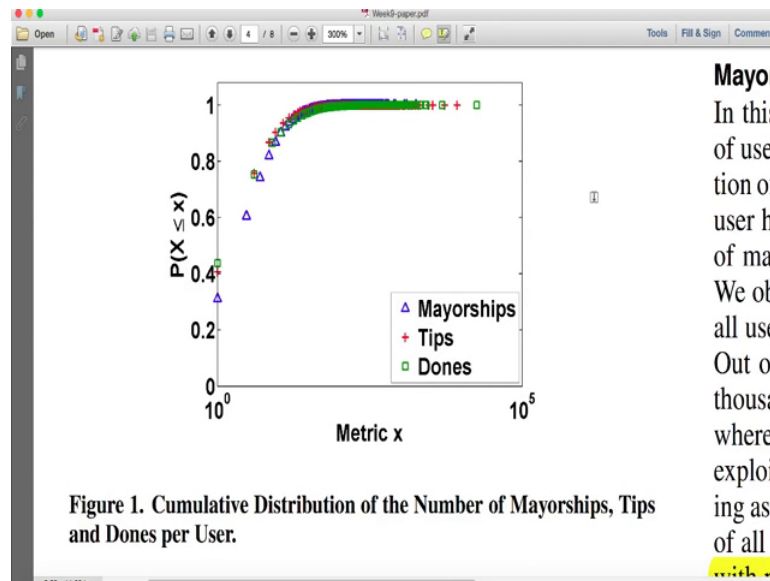
Many of such tools were used in this particular research to find out the location of the check ins, location of tips, and other information that was collected. So, if you look at the rest of the analysis, so here is the two figures that we will also talk about; first let us talk about the figure 1, which is shown on the left, but let us look at the content.

(Refer Slide Time: 06:03)
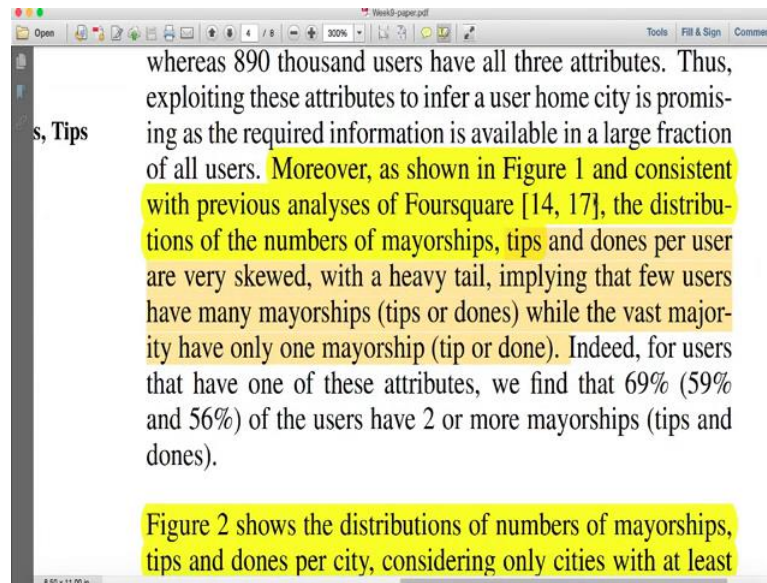
For the figure 1, as shown in the figure 1 and consistent with previous analysis of Foursquare the distribution of number of mayorships, tips and dones. So, this is what you do first when you collect such data. So, one other things that you would have generally seen in social media analysis is to show whether the data is a power law or show over that the data that you have collected from social media follows the pareto principle. which is to say that 20 percent of the users only actually contribute to the 80 percent of the content that is generated on the social media.

One of the similar type of graph was drawn here which is to see the distribution of mayorships, tips and dones per users and the inferences that they are skewed, with a heavy tail, implying that few users have many mayorships - tips or dones, while vast majority of them have only one mayorship, tip or done. So, that is essentially what a Pareto principle is that is essentially, what power law is also.

(Refer Slide Time: 07:30)
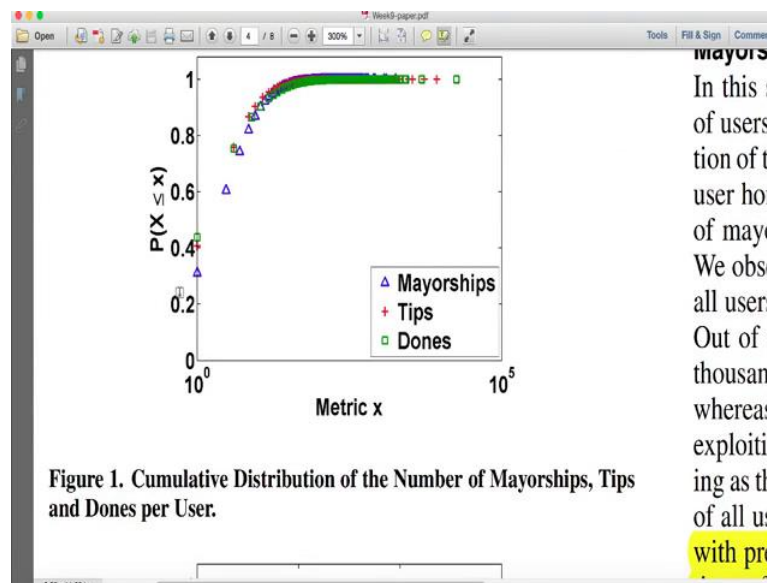
Figure 1. Cumulative Distribution of the Number of Mayorships, Tips and Dones per User.

So, if you look at the graph which is on the left for the figure 1, you will see the same thing which is large amount of, so large amount of users actually have small amount of users have so that is what this here. So, let us take figure 90 percent of the data is getting generated by small set of users; majority of the users over here do not contribute to any of the mayorships, tips or dones, so that is the graph that you would read.

(Refer Slide Time: 08:07)
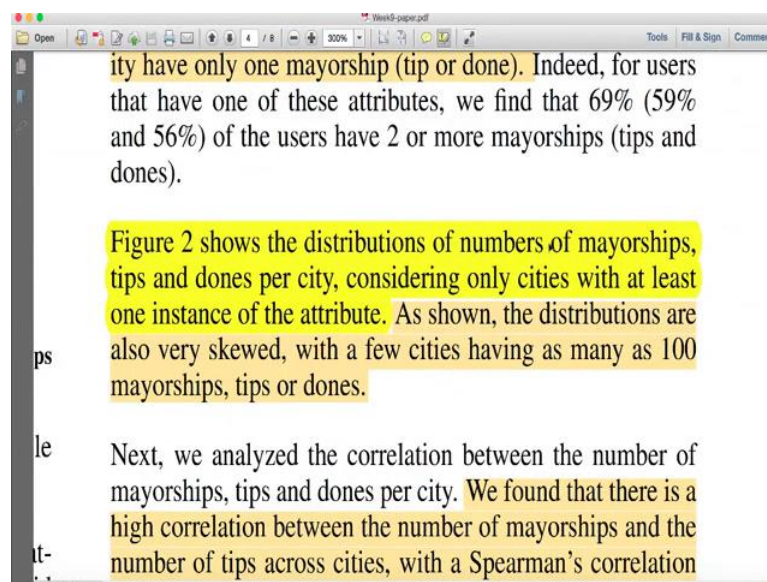
Similarly, let us go to figure 2. Figure 2 shows the distribution of number of mayorships, tips and dones per city, considering only cities with at least one instance of the attribute.

So, it is actually wanted to see whether from a city, whether you are able to get a lot of mayorship, tips and dones, this can actually help some, help find out how the data is. As shown the distributions are also very skewed, with a few cities having as many as 100 mayorship tips and dones.

(Refer Slide Time: 08:47)
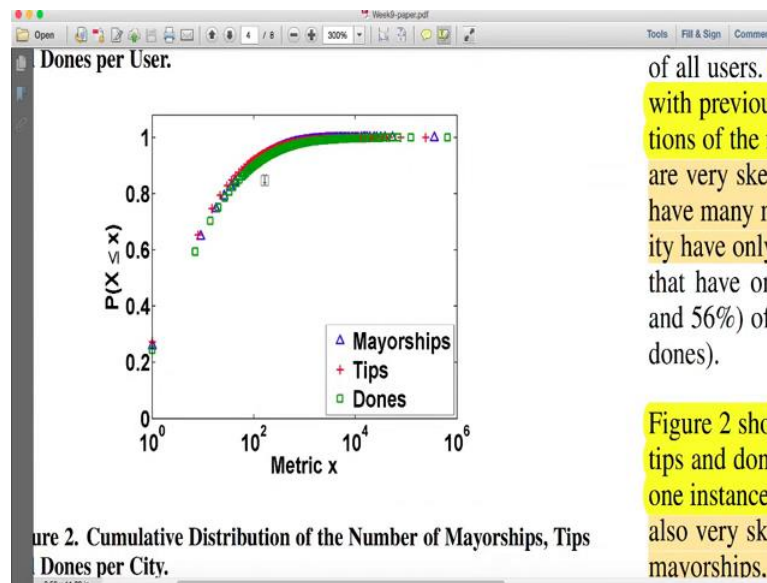
Figure 2. Cumulative Distribution of the Number of Mayorships, Tips and Dones per City.

So, again if you look at figure 2, the distribution is very similar in terms of the small number of cities having large number of mayorship, tips and dones; and large number of cities, do not have these. So, these are all large, so if you look at it somewhere around 80 or 90 years something that is only a small set of cities here. Large amount of cities do not have mayorship tips and dones or very little as many little mayorship, tips and dones because the condition was that they were considering only cities with at least one instance of the attribute, which is they should have had one tip, mayorship or done.

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 more concentrated in Southeast Asia, in cities like Jakarta, Bandung and Singapore, which are the top three cities in

So, then looking at correlation between the number of mayorship, tips and dones per city, they found that a high correlation between the number of mayorship and the number tips across cities, with the coefficient of 0.78. Similarly, the correlation is also high between number of mayorships and the number of dones, which is if there are more mayorships there in a city that is high chance that there will more of tips and dones also. This is helping us to understand that where if I find cities which I have a high mayorship, I should be able to find, there should be more of tips and dones also there.

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

So, if you look at the tips, tips are concentrated in different location around the Earth; which is if you remember tips of the content that people post for a particular venue. The top 3 cities in the number tips are New York, Jakarta and Sao Paulo with the total of 600,000 tips. Dones, on the other hand, tend to be concentrated in venues in the US, in cities like first New York, Chicago and San Francisco, and total they have about 1 million dones, so which is to show that some cities, some popular cities have a lot of these tips and dones New York being common in both tips and dones. Once again to say that cities generate a lot of these tips mayorships and dones and of course, this would also probably lead in to check ins also.

(Refer Slide Time: 11:55)
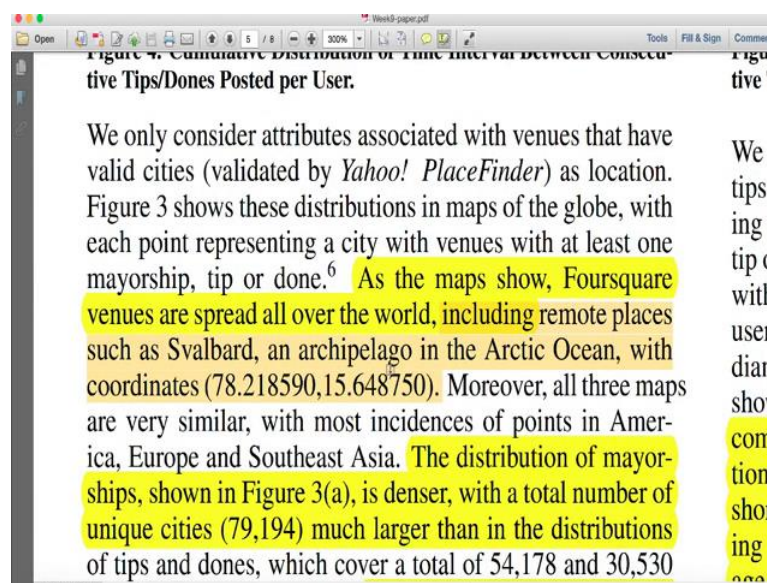
So, here is another graph which we will see in figure 3 and figure 3 for that we will see. So, figure 3 shows these distributions in maps of the globe with each pointer representing a city with venues, with at least one mayorship tip and done. So, essentially until now we only saw per city what is happening? Now when you look at it in a map, the figure 3 actually shows the results. As the maps show, Foursquare venues are spread all over the world including remote places such as Svalbard, an archipelago in the Arctic Ocean.

(Refer Slide Time: 12:41)

(a) Mayorships.   (b) Tips.

Figure 3. Global Distribution of Mayo
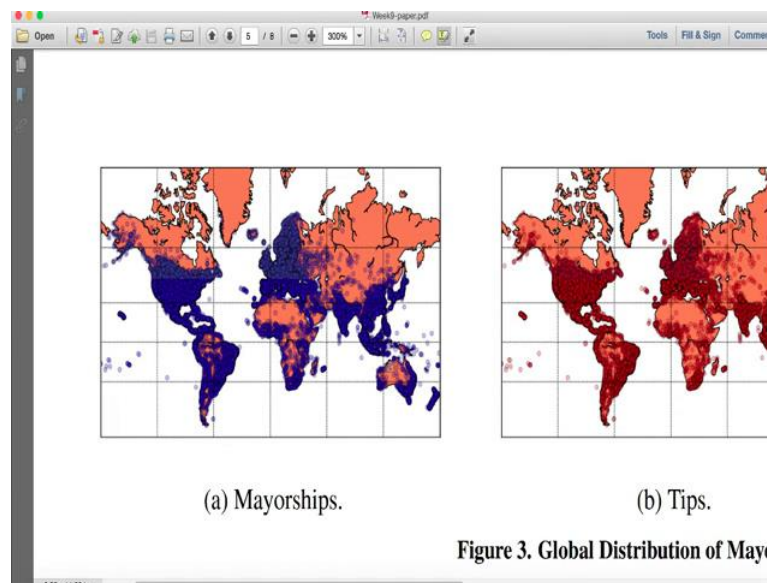
So, for example, let us go look at the figure 3 a, b and c. So, this is the mayorship. This is basically showing you every dot in this graph, every blue dot in this graph, figure 3(a) shows you the distribution of the mayorships that are available around the world, that were done around the world.
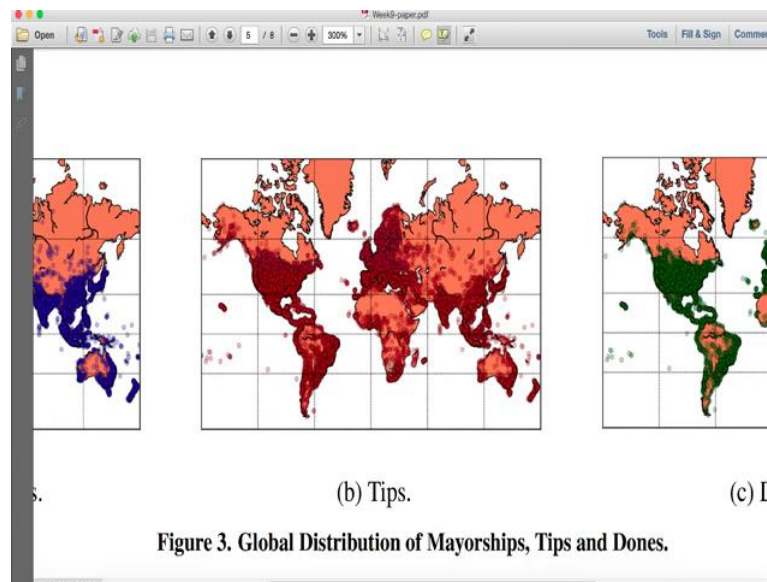
(Refer Slide Time: 13:00)

(b) Tips.   (c) D

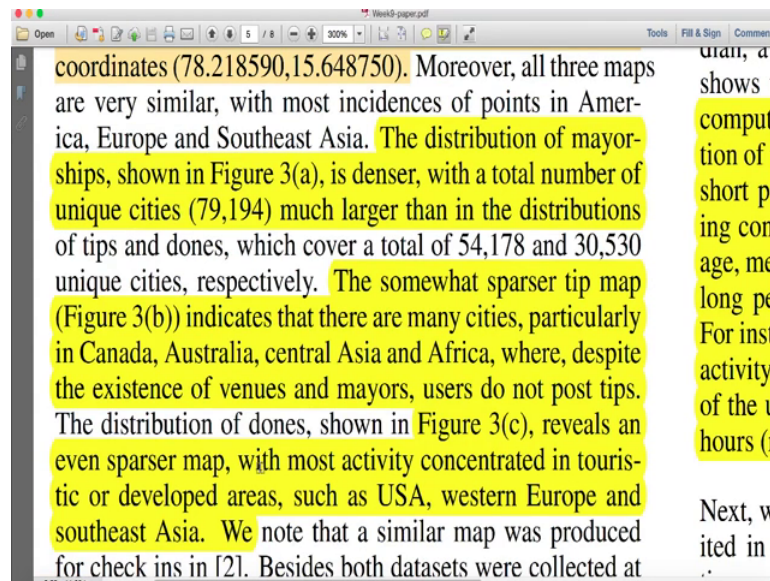Figure 3. Global Distribution of Mayorships, Tips and Dones.

Figure 3(b) shows you the tips that were done. So, if you see earlier, we saw that the there is high correlation between mayorship, tips and dones. So, therefore, when mayorships are high there is going to be tips and dones also which are high. So, you can

clearly see heavy concentration on many places in the world. And the last one is dones, the figure 3(c) shows you green dot, every green dot is a done from that particular locations. So, this basically shows you mayorships of the blue dot, tips - the red dot, and the green dot being dones. So that is figure 3 is denser with the total number of unique cities.

(Refer Slide Time: 13:48)

So, the distribution of mayorships shown in figure 3(a) is denser with a total number of unique cities being 79,000. So, 79,000 cities have higher check ins, have mayorships in the figure 3(a) in the total data. And if you look at the figure 3(b), somewhat sparser tip map for figure 3(b) indicates that there are many cities, particularly in Canada, Australia, and Central Asia, and Africa, where despite their insistence of venues and mayors' users do not post tips.

So, therefore, there is a chance that there are mayorships in that location, but not tips. 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. So, essentially even though there is a correlation between mayorships, tips and dones, there is actually some places which are sparser for a tips and dones.

(Refer Slide Time: 15:33)
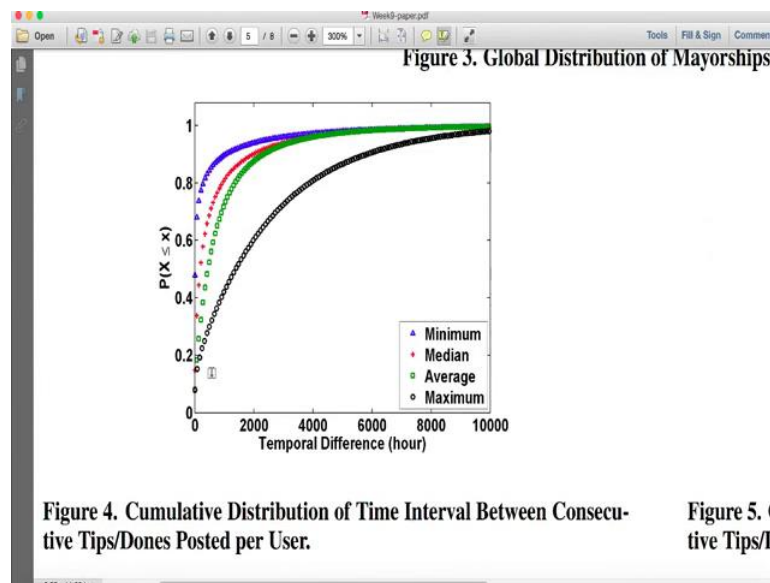
So, now let us look at figure 4; figure 4 here, so this is figure 4. Figure 4 is showing you the cumulative distribution of time interval between consecutive tips and dones posted per user. Why is this interesting? This is interesting to find out about the activity or the frequency of activity of the users.
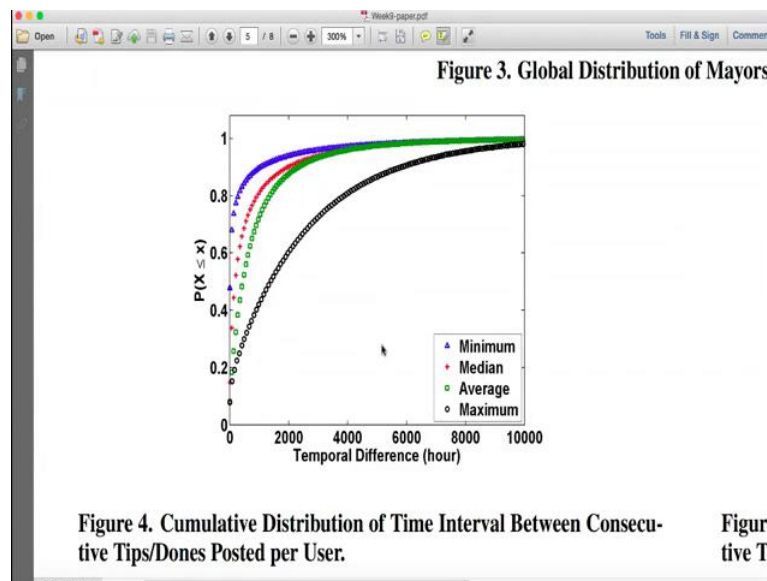
(Refer Slide Time: 16:07)
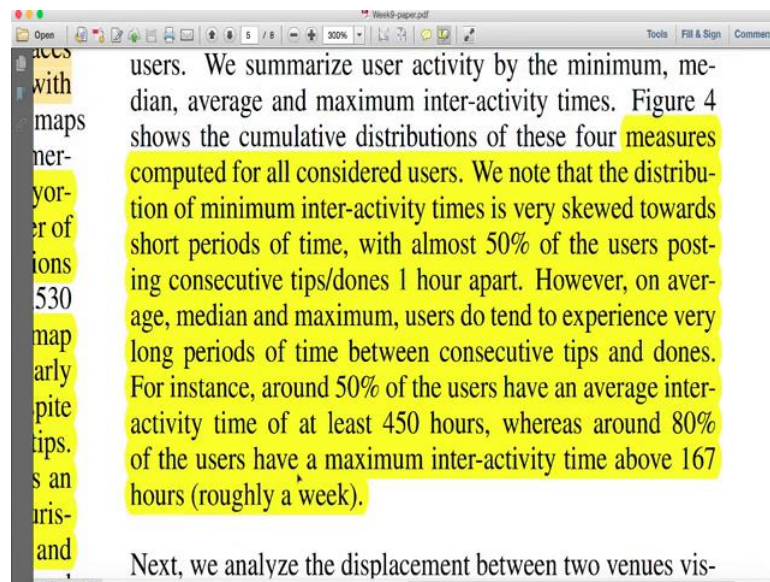
This is figure 4, figure 4 shows the cumulative distributions of these four measures. We note that the distribution of minimum inter-activity times is very skewed towards short periods of times, with the almost 50 percent of the users posting consecutive tips and

dones 1 hour apart, got it. So, that is it shows the there is a lot of content that are generated, lot of tips are generated by users, tips and dones are generated by users within apart 1 hour. However, an average, median and maximum users do tend to experience very long periods of times between consecutive tips and dones.

So, essentially what this shows is again it is going back to the same power law concept, there are some set of users where there is consecutive tips and dones are done very frequently. There is set of population where this distribution is actually pretty skewed, which is long set of long time taken between two consecutive tips and dones. For instance, around 50 percent of the users have an average interactivity time of at least 450 hours that is close to about 20 days, whereas around 80 percent of the users have the maximum interactivity of 167 hours - roughly a week.

(Refer Slide Time: 17:51)

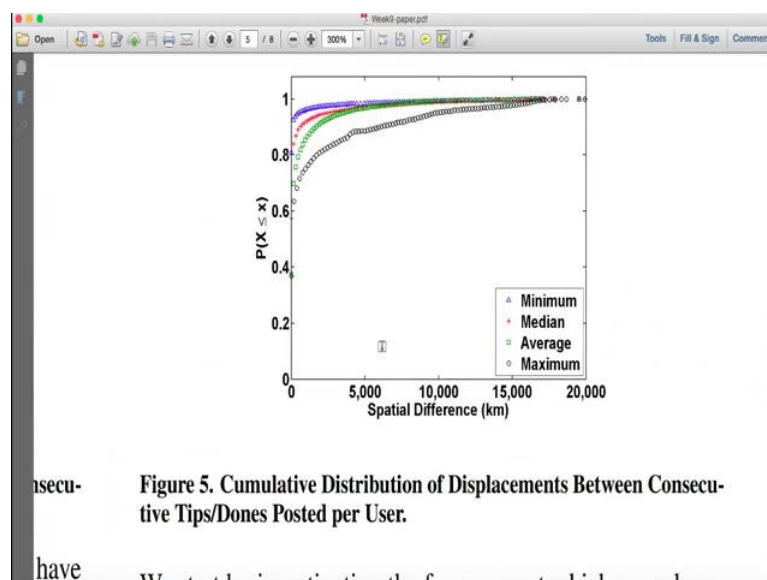Figure 4. Cumulative Distribution of Time Interval Between Consecutive Tips/Dones Posted per User.

So, here is the graph for cumulative distribution of time interval between consecutive tips and dones posted per user, it's the same thing as what we saw in the text.
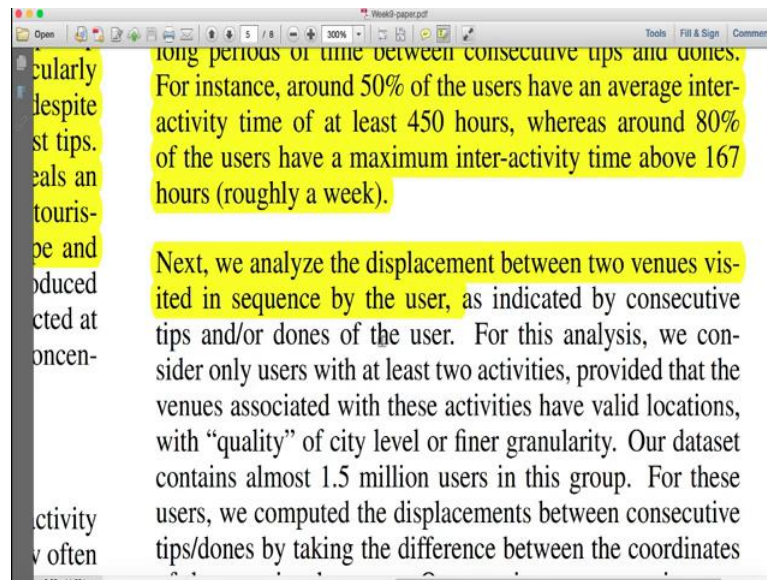
Now, let us look at the next figure, figure 5. So, this part we already saw. Figure 5 is basically looking at the same question which is now we are analyzing the displacement between two venues. The last figure that we saw was looking at two different timing in which the post was done. Now we are looking at two different venues, which are done by the same user consecutively.

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

So, let us look at the figure 5, first and I will tell you what the figure 5 all means. So, figure 5 is the cumulative distribution of displacement between consecutive tips and

dones posted per user. So, on the x-axis it is showing you the distance; on the y axis, it is showing you the number of the distribution.
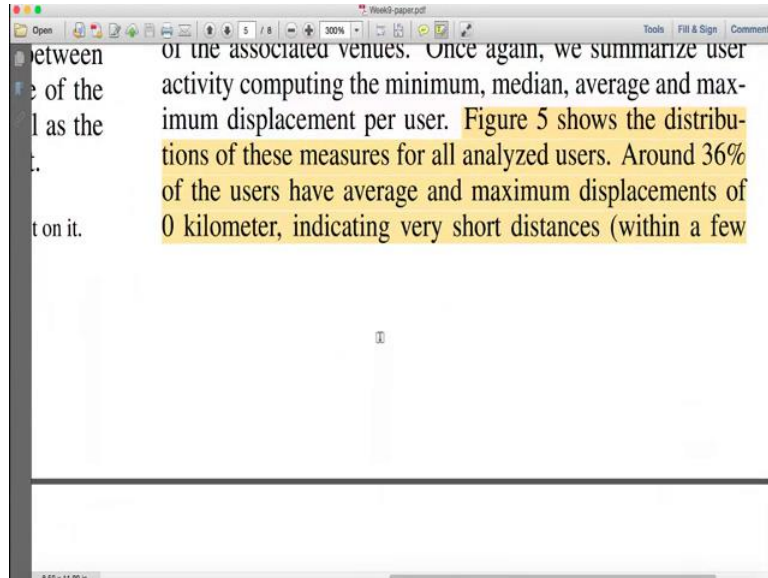
(Refer Slide Time: 19:27)

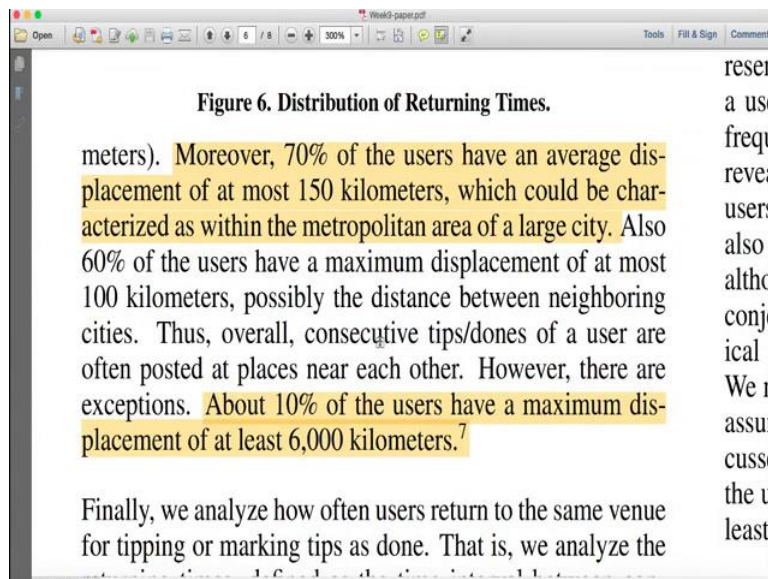Figure 5 shows the distributions of these measures for all analyzed users. Around 36 percent of the users have average and maximum displacements of about 0 kilometers, right, indicating very short distances - within a few meters.

(Refer Slide Time: 19:49)

70 percent of the users have an average displacement of at most 150 kilometers, which is basically somebody moving between cities. So, I probably from Delhi I go to Agra, and

then I do it check in all Mathura, I do check in Mathura, I do a tip or a done, that is what is actually capturing within 150 kilometers.

And about ten percent of the users have a maximum displacement of about 6000 kilometers, this is probably international travel between two consecutive tips or dones, so that shows what is the distribution of the users who we have in the dataset, the consecutive tips and dones that they do on Foursquare. Let us go to the figure again. So, this is basically showing you that 70 percent of the users are about 150 kilometers and the 10 percent is about the 7000 kilometers that is what you will see in this figure.

(Refer Slide Time: 21:09)

Figure 5. Cumulative Distribution of Displacements Between Consecutive Tips/Dones Posted per User.
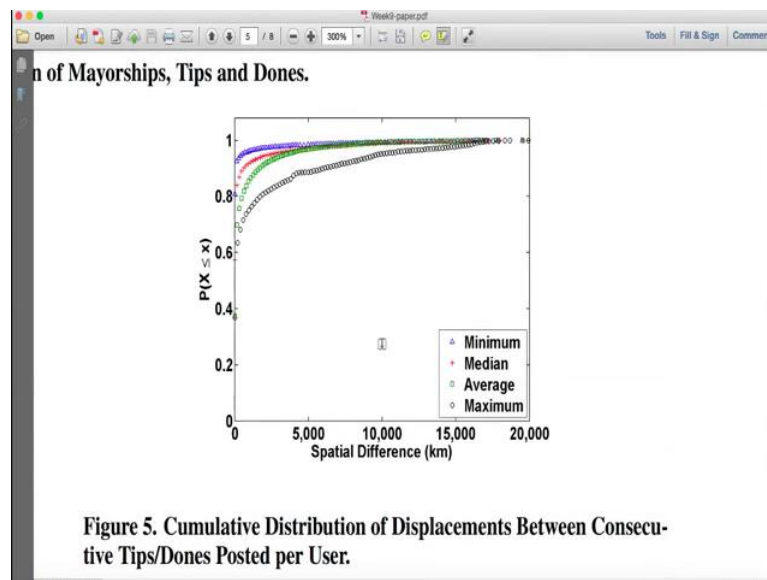
So, if you seen here 7000 kilometers, so it is about the last 10 percent of the users, who are about 7000 kilometers and then very short is about 70 percent. Average 70 percent is that the green one is the average, the green square is the average. The red plus symbol is a median, triangle is the minimum, and the circle is the maximum right, so that gives you a sense of.

(Refer Slide Time: 21:51)
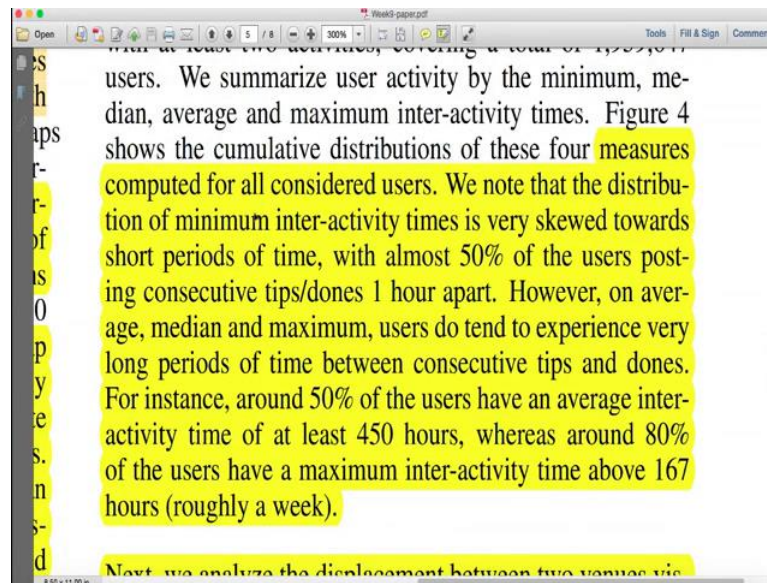
So, now, what all we have seen, we have seen the time, consecutive, tips or dones that is done with respect to time, consecutive tips and dones with respect to distance.

(Refer Slide Time: 22:18)

Figure 6. Distribution of Returning Times.
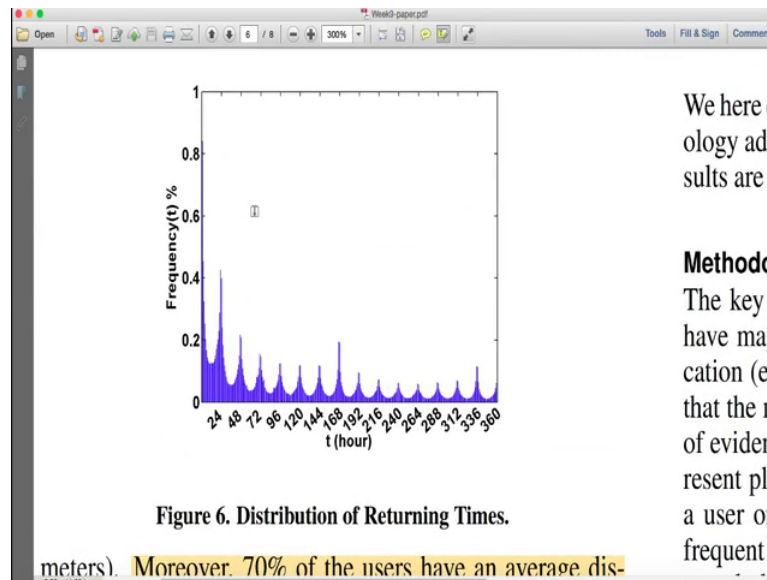
Now we will see the next figure, next figure is actually very interesting. Next analysis is actually a very interesting analysis, where they saw how frequently that the check ins, the tips or the dones are coming back for that particular location. So, this is distribution of the returning time. So, if I do tip or a done in IIIT, how frequently do I actually do a tip or a done in that location.
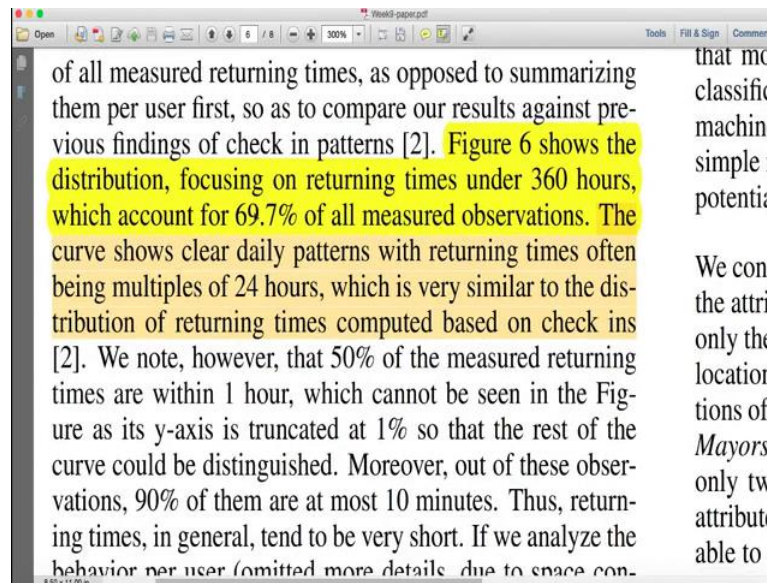
And the figure 6 shows the distribution focusing on returning times under 360 hours, which account for 69.7 percent of all the measured observations. The curve shows clearly daily patterns with returning times often being multiples of 24 hours which is very similar to the distribution of returning times computed based on the check ins.

So, if you really look at what does it mean why is it 24 hours? If somebody checks into office in the morning today that they have got into the office, shop, institute and everything they do the same check in the next day, so that is what this means right. Check ins, the references is given to another research where the check ins where seen but in this case we are also looking at the tips or the dones. That is a very interesting conclusion to know or basically it is complimenting the real world behavior that you could expect from the users.
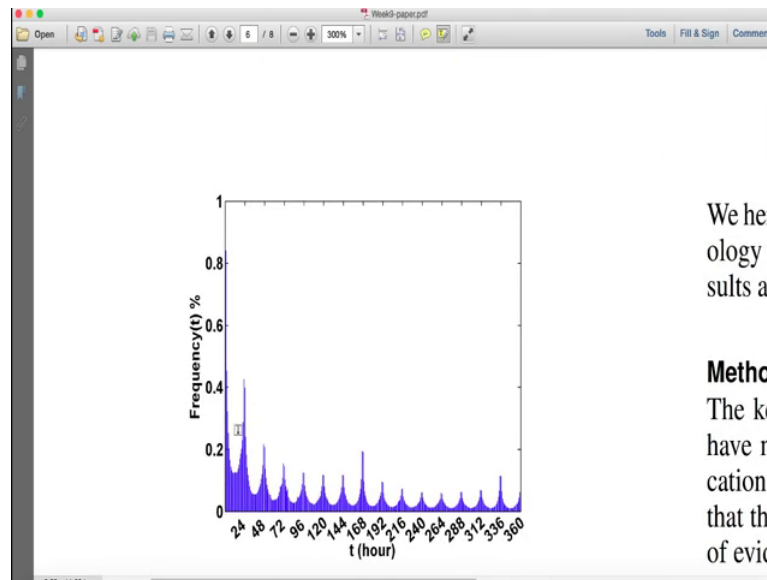
Here is a graph, which you see here this is about 360 hours. And you can clearly see that it's coming back. So, this is for every day 24, 48, 72, 96 120, so it is kind of coming back every 24 hours and sometimes the frequency is also increasing for something happens. So, this is 168 should be the week. So, therefore, there is a slighting increase from the day that which is which is the 7th day of tips and done. I hope that is making sense essentially the conclusion there is that people come back to that same location with the examples like me doing it in IIIT Delhi, that is figure 6 that is an interesting analysis.

So, now what we will do is, now we will attack the question that we started off with which is to find out how much can be actually inferred about the users' home using this data. So, in this case, we are going to actually use data for most popular location among mayorships, tips and dones of a user or home location using a majority voting scheme. I am going to explain to you at this voting scheme is there are two tables that we will look at and,

(Refer Slide Time: 26:01)

Table 3. Home Location Inference.
Classes Distribution

	Home City			Home State			Home	
	Class 0	Class 1	Class 2	Class 0	Class 1	Class 2	Class 0	C
	727,179	847,876	239,129	707,953	913,166	110,110	727,179	1,0
	725,073	671,576	192,781	702,583	727,219	99,672	725,073	83
	546,815	541,795	106,297	524,137	561,165	55,115	546,815	63

We will also see what mechanisms the authors followed in terms of generating this information about what is the possible location that this person's home would be.

(Refer Slide Time: 26:15)
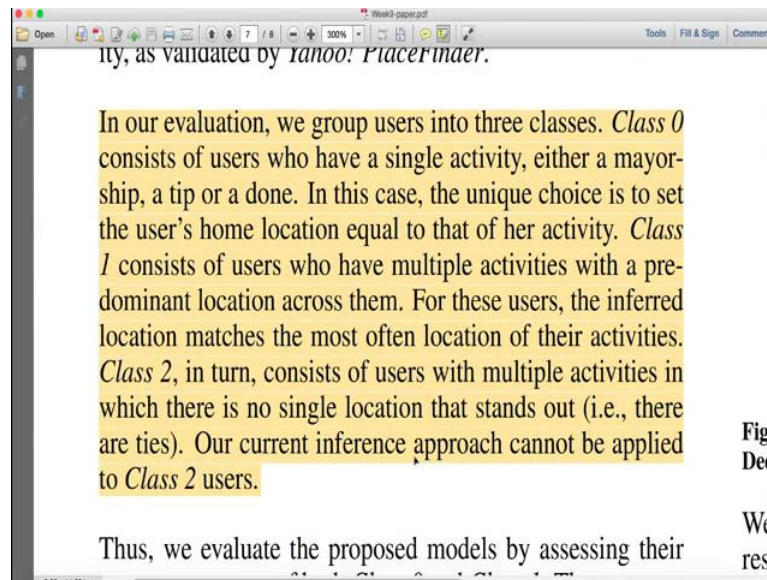
Here is the scheme that they followed. They actually put the data into 3 buckets, class 0, class 1 and class 2. The class 0 is of the users who have single activity either the mayorship, the tip or done. They only have one activity in the dataset, whereas class 1 consists of users who have multiple activities with predominant location across them. So, for example, I have multiple tips and dones, mayorships in my account, but there is one which is very, very high which is IIIT Delhi for that matter. Class 2 is consists of users with multiple activities in which there is no single location that stands out.

So, again just to understand, if you understand this I think the inferences become much simple, the logic the authors followed is that take all the users who are been doing this tip, dones and mayorships in the data. Class 0 or the people who have only single activity either a tip, or a mayorship or a done, class 1 is set of people who were where this one majority location that shows up for them. Class 2 is a set of people where multiple activities are done, but no single location is actually predominant in their activity, so that is the kind of classification that they made with the users.

Table 3. Home Location Inference.

	Home City			Home State		
Classes Distribution						
Features	Class 0	Class 1	Class 2	Class 0	Class 1	Class
Mayorship	727,179	847,876	239,129	707,953	913,166	110,1
Tip	725,073	671,576	192,781	702,583	727,219	99,67
Done	546,815	541,795	106,297	524,137	561,165	55,11
Mayorship+Tip	898,293	1,322,214	300,831	878,578	1,398,351	146,5
Mayorship+Done	825,009	1,213,917	270,974	805,029	1,278,784	130,4
Tip+Done	831,759	1,038,268	223,093	807,091	1,089,638	116,5
All	939,888	1,573,471	310,045	919,938	1,643,825	153,9
Accuracy						
	Home City			Home State		
Features	Class 0	Class 1	Total	Class 0	Class 1	Tota
Mayorship	51.61%	67.41%	60.12%	71.27%	80.92%	76.70

Now I will show you two tables, one is the number of people, number of the data points from this dataset. If you just consider the class 0, class 1 and class 2, how many people are actually where you can infer home city, home state, and home country. So, how do I read this graph this how do you read this table. This table is this column it showing you features mayorships, tip and done, mayorship plus tip, mayorship plus done, tip plus done and all of them, right. So, which is if you take only the mayorship, what is the class 0, which is only if I consider mayorship and there is only 1 activity by this user, there are about 727,000 data points in class 0; 847,000 where that are users where one location is actually predominant; and 239,000 there is no location that is predominant, that is how you read this table, correct.

So, if you look at mayorships 127,000; mayorship plus tip 898,000; obviously, mayorship plus tip will be higher, ==all will be higher for all of them==, bigger than all of them and that is for the city. So, for the state 700,000 is for class 0; 900,000 are for class 100,000 is for class 2. Similarly, for the country, so that is giving you a sense of in the data points or the pieces of information that is available for each of these features.
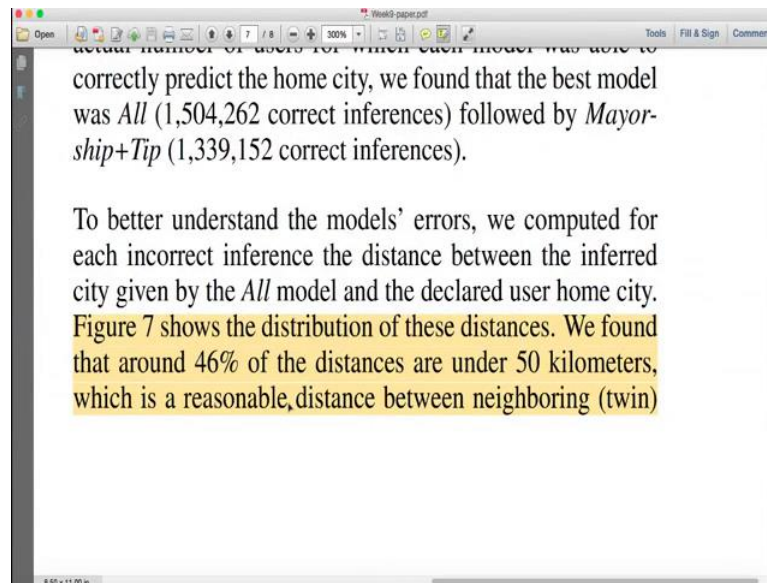
(Refer Slide Time: 30:09)

	546,815	541,795	106,297	524,137	561,165	55,115
Done	546,815	541,795	106,297	524,137	561,165	55,115
Mayorship+Tip	898,293	1,322,214	300,831	878,578	1,398,351	146,526
Mayorship+Done	825,009	1,213,917	270,974	805,029	1,278,784	130,439
Tip+Done	831,759	1,038,268	223,093	807,091	1,089,638	116,549
All	939,888	1,573,471	310,045	919,938	1,643,825	153,955

	Accuracy					
	Home City			**Home State**		
Features	Class 0	Class 1	Total	Class 0	Class 1	Total
Mayorship	**51.61%**	**67.41%**	60.12%	**71.27%**	**80.92%**	**76.70%**
Tip	51.52%	67.29%	59.11%	70.29%	80.59%	75.53%
Done	50.09%	61.74%	55.89%	70.16%	78.38%	74.41%
Mayorship+Tip	51.57%	66.24%	**60.31%**	70.21%	80.27%	76.39%
Mayorship+Done	51.05%	65.27%	59.51%	70.01%	79.89%	76.07%
Tip+Done	51.18%	64.16%	58.38%	69.76%	79.28%	75.23%
All	51.46%	64.86%	59.85%	69.74%	79.53%	76.02%

we consider only users whose home city attributes

Now, let us look at accuracy, which is if I were to use this information and find out that mayorship, only using the mayorship, I am able to find the home city 51.6 percentage of the times, wherever the percentage is higher it is been actually kept in bold. So, if you look at mayorship it is class 0 on class 1, we cannot do class 2 because it is actually the places where a particular location cannot be, one single location cannot be inferred, which is because our goal is to infer actually the home location.
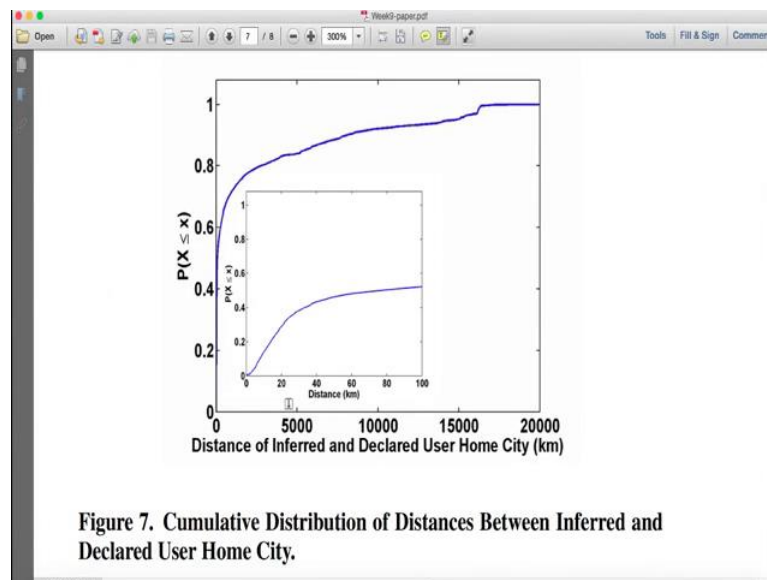
So, if I have multiple locations, I am not going to use that column, that is why class 2 does not exist in high accuracy. So, which 67 percentage of the times, home city can be inferred, if I look at class 1 category of people. Which is I do a lot of tips and dones, but my predominant place where I do a tip or a done, tip or a done is actually my home location. Home state becomes higher, seventy percent and home countries even higher, of course, the percentage for the country is going to be, so the percentage of city will always be lesser than state, will always be lesser than country. Because here to get the country that I am from India more difficult to get that I'm from Tamilnadu as a state it is even more difficult to get that I am from Chennai, that is about the inference of the home location.

(Refer Slide Time: 32:09)

There's another interesting graph that authors have which is figure 7, which I show you the graph and then I will try to explain it.

(Refer Slide Time: 32:17)

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

So, here figure 7 is cumulative distribution of distances between inferred and the declared home city, which is that because people actually declared that what the home location in these services also. If you go to my account, you will find that is, if you go to my facebook account I probably say that I am from the current location is from New Delhi and my home location is from Chennai, so that information you can use to make

the difference which is what did we predict from the table that I showed you know, which is prediction of my home location with class 0 or class 1.

And then use it for finding the difference between what did I say and what I actually have, that is the graph here. So, x-axis is the distance of inferred and declared user home city, and y-axis is the probability. So, this is how you will read the graph.

(Refer Slide Time: 33:28)
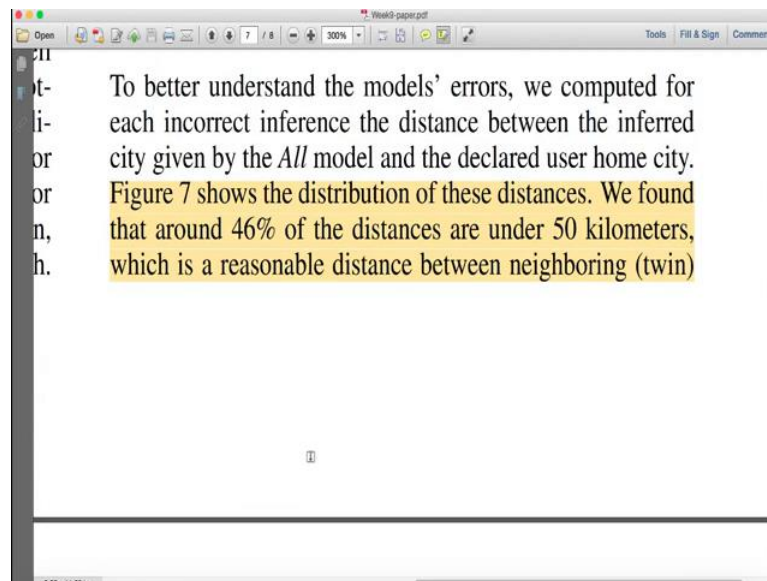
Which is, distribution of these distances are shown in figure 7. We found that 46 percent of the distances are under 50 kilometers that is what the authors did.

(Refer Slide Time: 33:40)

Figure 7. Cumulative Distribution of Distances Between Inferred and Declared User Home City.
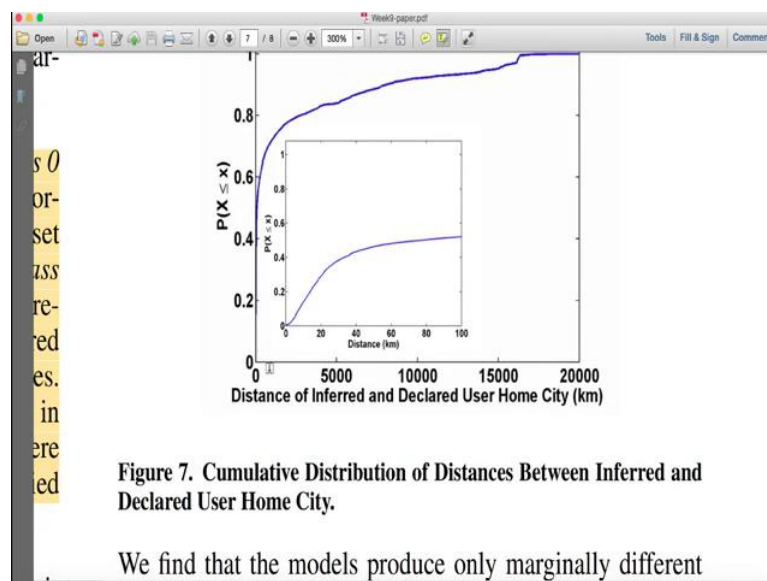
They actually zoomed in. So, this is meaning, this is 5000 kilometers, but as if you look at this graph the inside graph is only for 100 kilometers. So, you can clearly see that 46 percent of the distances are 150 kilometers. So, here is 50 kilometers and if you see 46 percent should be somewhere here correct. So, 46 percent of the users are actually having the error between finding the home location and the actual location is about 50 kilometers, that is pretty small, if I were able to actually use this information with only 50 kilometers of error which is I just getting it from the general behavior tips and dones, that's quite effective.

(Refer Slide Time: 34:37)
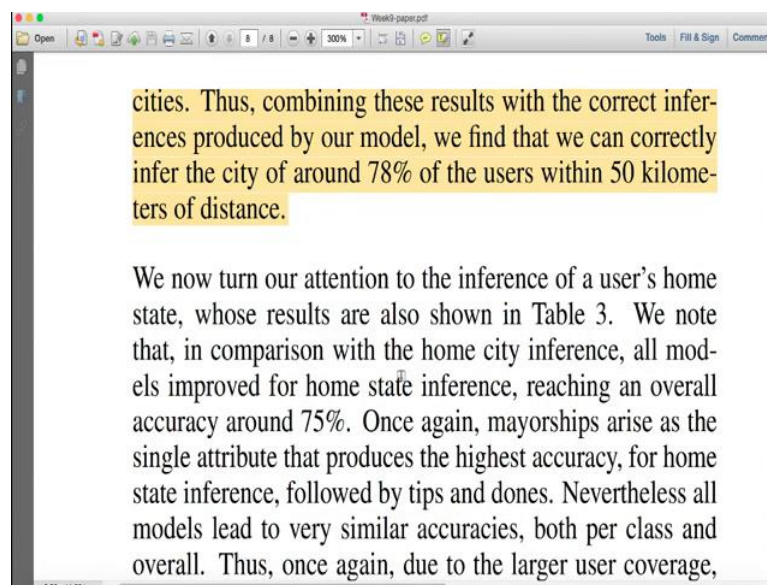

So, if you just take this model, and then if you look at the author's claim that 78 percent of the users within 50 kilometers of distance, which is what they are saying is combining these results with the correct inferences produced by our model, we find that we can correctly infer the city of around 78 percentage of the users within 50 kilometers. So, whatever your city is we will be able to make an inference of that city of about 78 percent within the 50 kilometers of distance, correct.

(Refer Slide Time: 35:27)
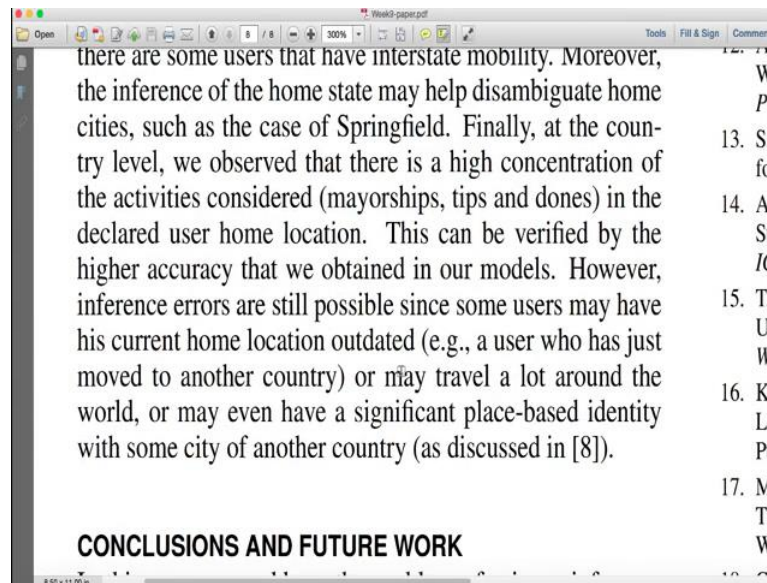
So, that gives you a sense of how also if you go if you remember even in the abstract I showed you this 78 percent of accuracy that the authors claimed that you can find the home location. And of course, in the paper structure, you finish off with the conclusions and future work and probably have some limitations if there are any data limitations in data methodology, any limitations in the paper, right.

(Refer Slide Time: 35:49)
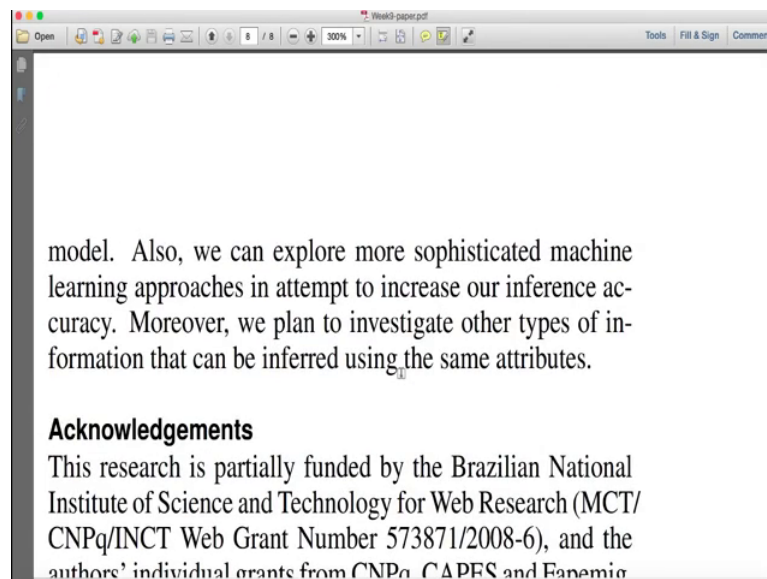
So, that gives you a sense of how a paper meaning the things that you have seen in the class until now which is to look at take some data do some analysis, make some inferences, how these inferences are put into paper is what we saw in this particular lecture. And the focus was actually taking foursquare and finding the home location. With that, I will stop here for this paper, and I will see you soon.