

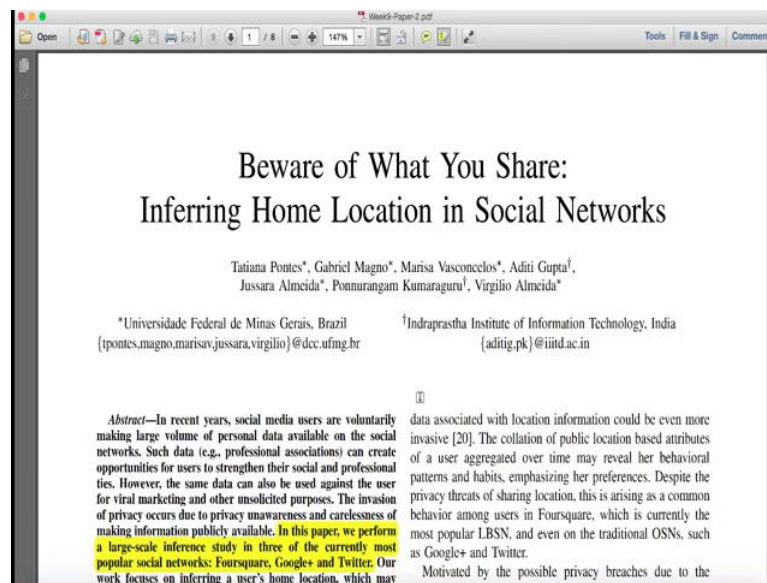
Privacy and Security in Online Social Networks
Prof. Ponnuram Kumaraguru (“PK”)
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Week - 10.1

Lecture – 32

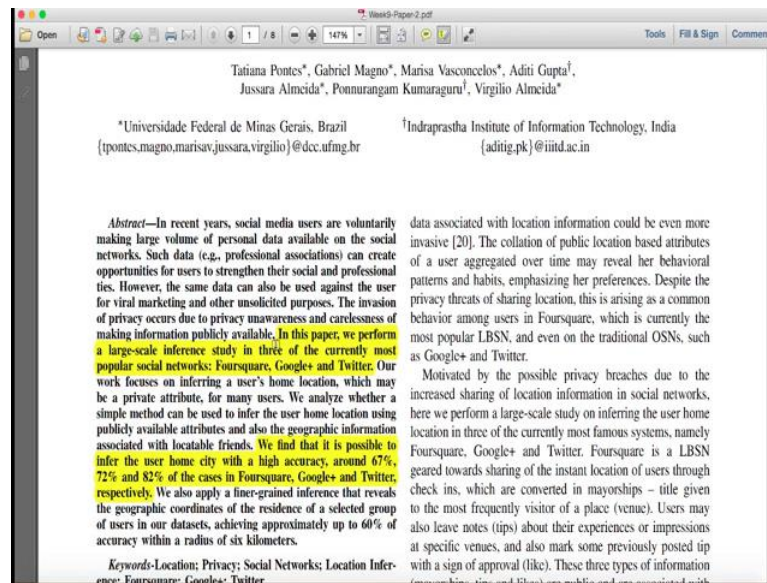
Beware of What You Share inferring Home Location in Social Networks

(Refer Slide Time: 00:12)



Welcome back to the course Privacy and Security in Online Social Media. Continuing the trend that I did last lecture, I am going to continue actually looking at some papers which are basically addressing the problem of privacy leakages from location based services. If you remember last lecture, we had paper which looked at Foursquare, and the paper analyzed, how they can actually identify, where a person lives. That was only using the Foursquare mayorship, tips and dones. And what we are going to see now is almost the same topic, but we are going to actually compare it with different social networks.

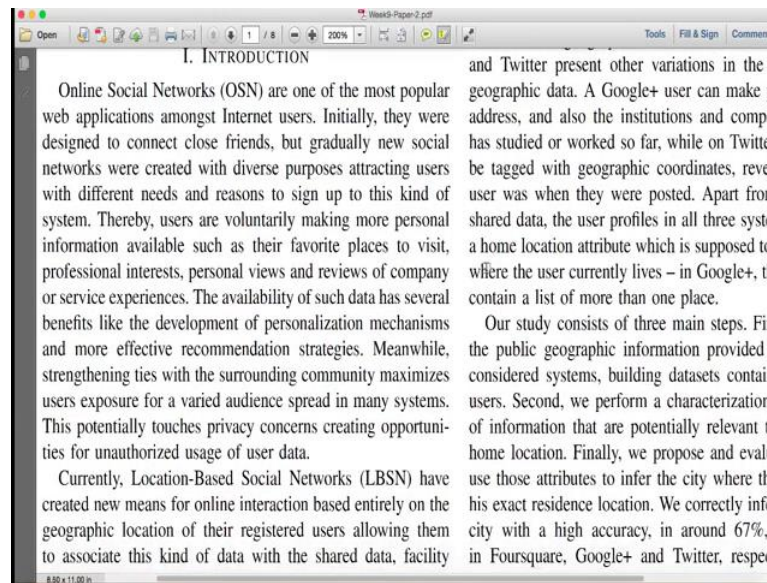
(Refer Slide Time: 01:11)



So, if you see here in this paper the authors perform a large scale inference study in three of the currently most popular social networks like Foursquare, Google plus and Twitter. So, the goal in this paper is very similar to the paper that we saw last time, but it is going to be looking at different social networks not just only Foursquare. So, in this authors looked at Foursquare, Google plus and Twitter. You know as part of this course, you have already seen all three social networks in terms of their content, in terms of the data collection that is done and information that you can actually collect from the social networks.

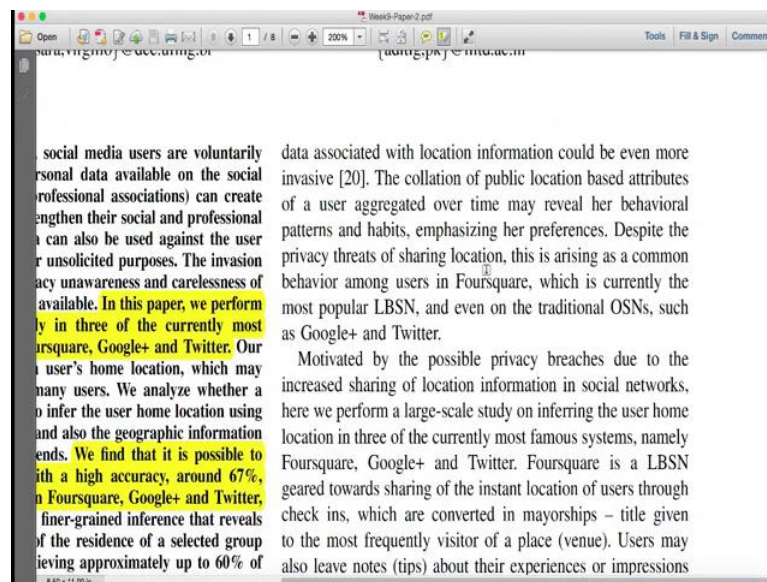
The authors actually find that it is possible to infer the user home city with the high accuracy around 67 percent, 72 percent and 82 percent in the case of Foursquare, Google plus and Twitter, which is 67 percent for Foursquare, 72 percent for Google plus and 82 percent for Twitter. I am sure as we move along; you will actually understand why Twitter is actually high in terms of finding out the home location.

(Refer Slide Time: 02:55)



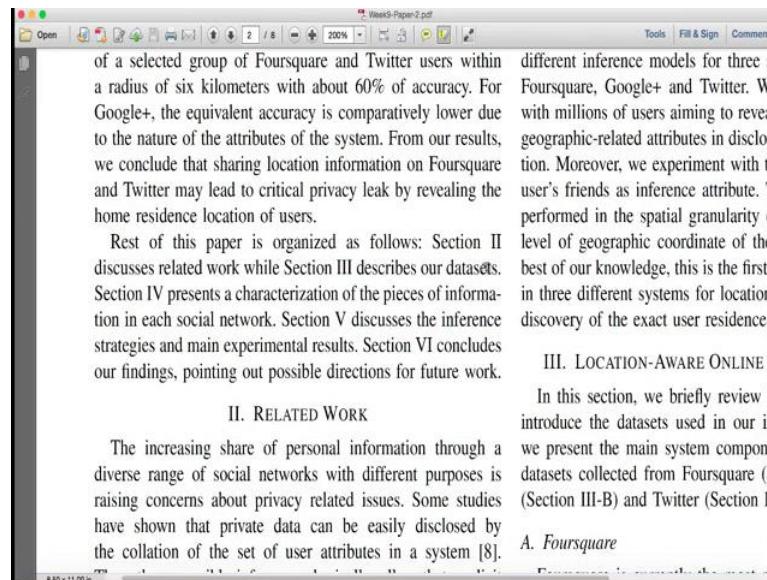
So, now let us look at the paper in terms of the same structure as we saw last time, introduction, talking about what a location based social networks are.

(Refer Slide Time: 03:04)



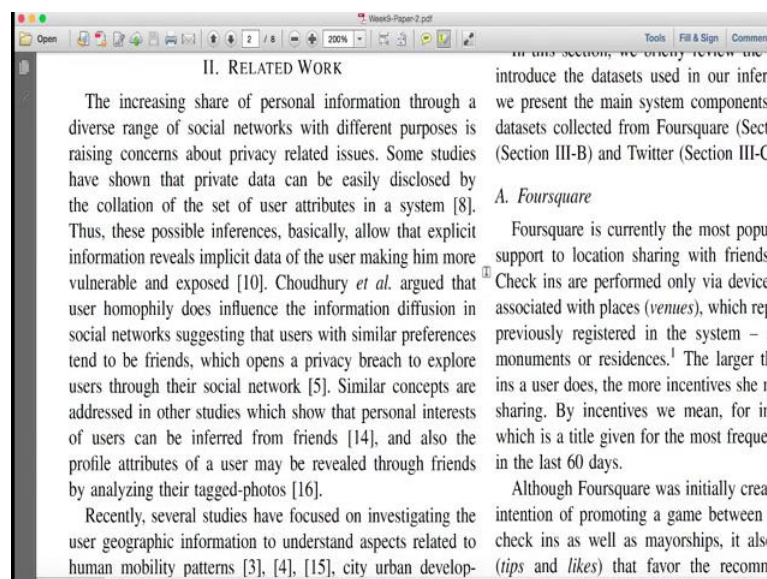
Talking about collecting data from a different social media services like Google plus Foursquare and Twitter and different research that are done in the context of Foursquare Google plus and Twitter. And I am talking about what information was collected, and a little bit of conclusions of the paper itself, and then talking about how the paper is organized.

(Refer Slide Time: 03:32)



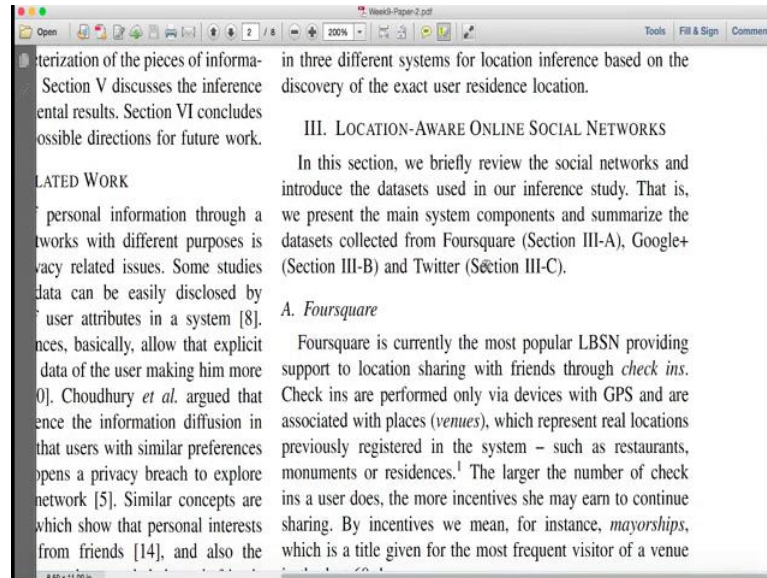
So, this is generally the structure that we saw even the last time, meaning almost all papers appear to see the structure would be the same a paragraph about the 30,000 feet high view of the problem. Then the paragraph about the current problem and what is missing, then the paragraph about what is, what was done in this paper and then some kind of a contribution from this paper. Related work again I am not going to detail in this particular related work.

(Refer Slide Time: 04:02)



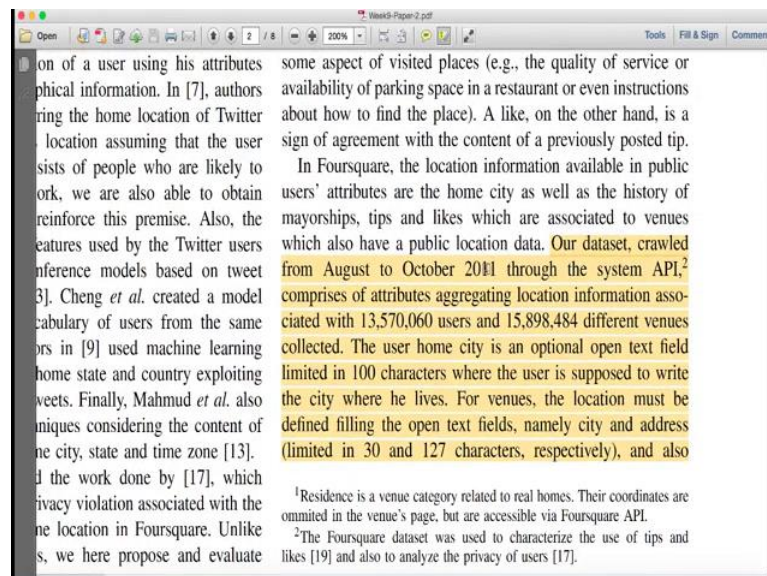
But related work generally talks about these kind of privacy leakages from location based services and work done on collecting data from these three social networks and inferences that were done.

(Refer Slide Time: 04:19)



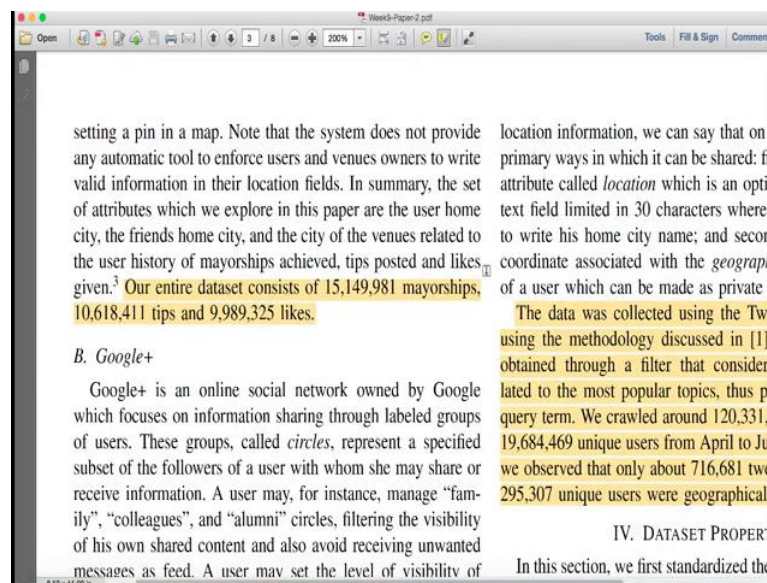
Then many a times researchers actually tend to write details about the social network that is being discussed in terms of just introducing the terminologies which we saw in the last paper also. Here it is talking about Foursquare then there would be about Google plus and then Twitter.

(Refer Slide Time: 04:42)



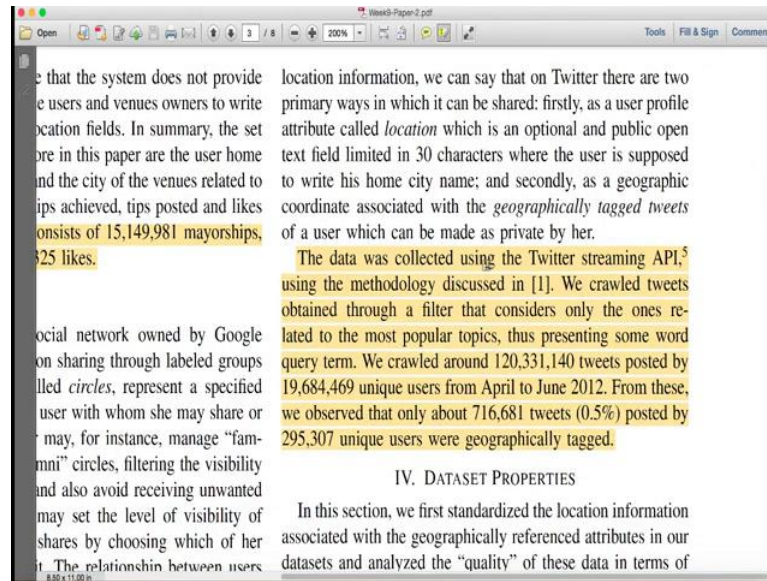
So, here if you look at it, the dataset that was used in the study is the same as the last paper. Dataset crawled between October 2011, through the system and it comprises of 13 million users and about close to 16 million different venues. And the user home city is an optional open text field limited to about 100 characters. For venue, is the location must be defined filling the open text fields, namely city and address limited to 30 and 127 characters respectively. That is the kind of data and that is the kind of information that is available when you collect these data for venue, tips and done.

(Refer Slide Time: 05:42)



So, the entire dataset is about 15 million mayorships, close to 11 million tips, and close to ten million likes. All right? So, likes is basically done in terms of Foursquare terminology.

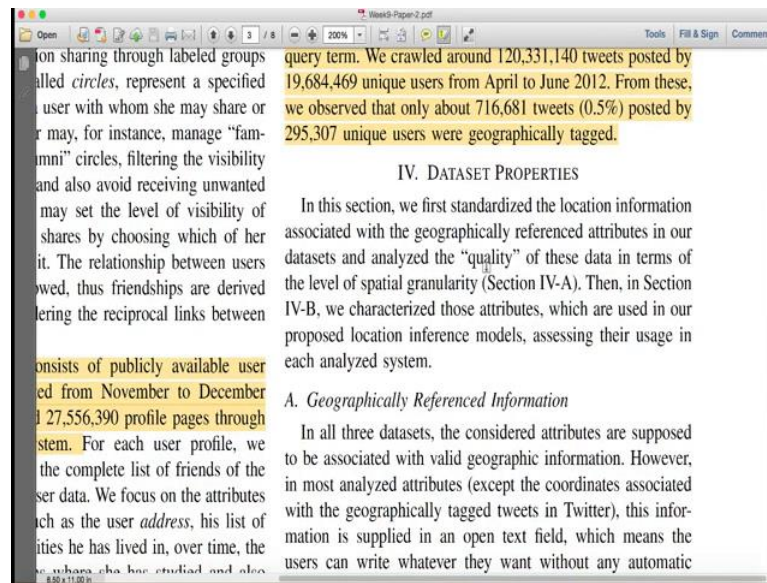
(Refer Slide Time: 07:34)



Now, let us look at the data from Twitter. So, the data the data from Twitter was collected using Streaming API, which all of you are aware of. And the crawl was done for 120 million tweets posted by about close to 20 million unique users from April to June 2012. 0.5 percent of the posts posted by unique users were geographically tagged.

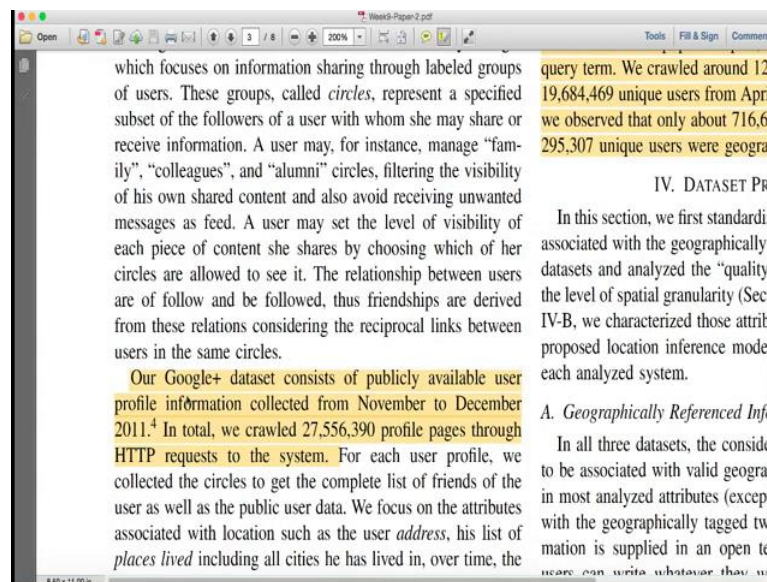
So, what does this mean, this means that there are only 0.5 percent of the total posts that were collected where there is geo tagged information for the post which is geo tagged information for the users also. There were about 700,000 tweets and about 300,000 unique users. That is the exact location, which we have discussed in the past, which is latitude, longitude of the post from where the post is coming.

(Refer Slide Time: 08:47)



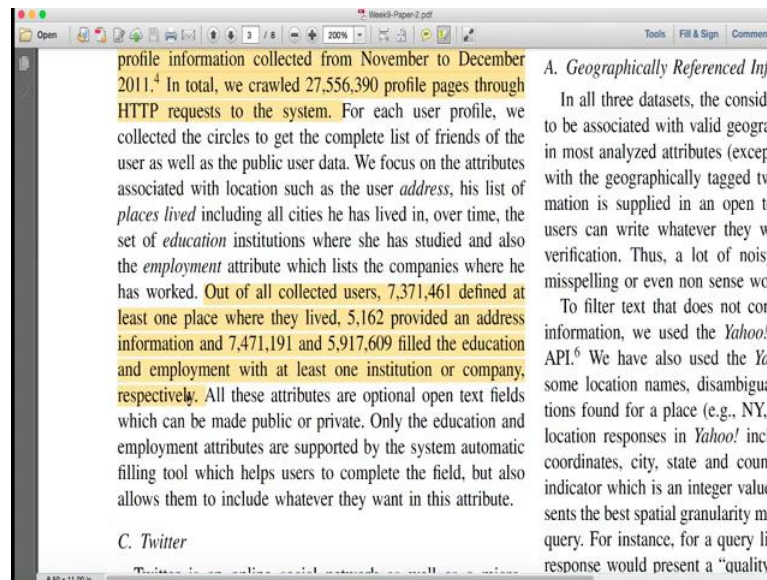
So, that is the background about the dataset. Essentially all of them are talking about in millions in Twitter it's about 0.5 million geographically tagged tweets geo tag tweets.

(Refer Slide Time: 09:00)



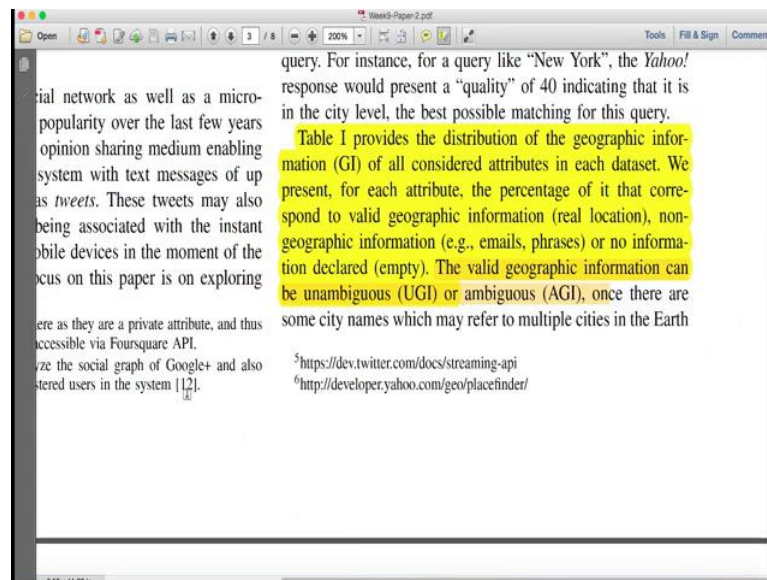
In Google plus that are about 27 million profiles that were crawled and about 7 million education and 6 million employment.

(Refer Slide Time: 09:07)



And Foursquare has details of about 16 million mayorship; 11 million tips and close to 10 million likes. That is the dataset we are going to play around with to do the analysis, to find inferences about the home location of the person.

(Refer Slide Time: 09:38)



So, in any data set, when you analyze, first you know you want to actually provide exploratory data analysis, and you want to provide what the data set looks like, because this will help reproducibility of that research. This will help others to actually collect data, if they were to, the point here is that if others want to collect the data which is very

similar to what you collected; and if others want to do the same analysis that you did the results should be the same. That is the idea for reproducibility of the research. So, explaining how you did collected the data, explaining what the data looks like is extremely important in terms of actually writing these research papers.

(Refer Slide Time: 10:45)

TABLE I
AVAILABILITY OF GEOGRAPHIC INFORMATION (GI) IN VARIOUS ATTRIBUTES IN OUR DATASETS. ALL VALUES PRESENTED ARE IN PERCENTAGE (%).

Statistics	Foursquare			Google+			Twitter	
	User Home City	Venue City	Places Lived	Address	Education	Employment	User Location	Geo-tagged Tweet
% valid UGI	95.35	55.45	61.85	0.01	52.95	34.52	73.28	100.00
% valid AGI	2.65	18.04	6.66	0.002	11.01	14.67	9.70	0.00
% non-GI	1.80	26.51	31.48	0.01	36.04	50.81	11.90	0.00
% empty	0.20	0.00	0.00	99.98	0.00	0.00	5.12	0.00

indistinctly, being *Yahoo!* unable to decide which is correct – e.g., “Springfield” is the name of ten different cities, only in the United States. For Foursquare, due to space constraints in the paper, we group tips, likes and mayorships as venue attributes, while users attributes correspond only to the home city field. Note that the vast majority of Foursquare users (98% of 13,570,060) provided valid home city locations, with only a tiny fraction leaving it blank (0.2%) or filling it with non-geographic information (1.8%). Moreover, 11.6 million venues have valid locations associated, although a substantial fraction of all venues have non-valid locations (around 26%) or valid but ambiguous location (18%). This large fraction of non-valid or ambiguous venue locations comes as a surprise, particularly considering that, unlike the user home city field, the venue location information is a mandatory attribute.

In comparison with Foursquare, the fraction of valid locations in our Google+ dataset is much lower for all considered education, employment and address attributes are more often provided at finer granularities, i.e., street level for employment and address, and Point Of Interest (POI) for education. Finally, the “quality” of the location provided in users’ tweets is either at the street (18.05%) or at the geographic coordinate (81.95%) levels. The availability of public finer-grained location information opens an opportunity for more specific inferences regarding user home location, such as user residence, as discussed in Section V-C.

B. Attribute Characterization

In the previous section, we analyzed the availability of valid and unambiguous geographic information as well as the “quality” of this information across all analyzed attributes. Now, we focus on the usage of these attributes and analyze their distributions across users in each dataset. We aim at assessing the potential of exploiting these attributes for inference

So, table one provides the distribution of geographic information of all considered attributes in each dataset in each dataset. We present for each attribute the percentage of it that corresponds to the valid geographic. Let us look at the tables. So, this table is the one that is referred. This table talks about availability of geographic information in various attributes in the datasets. So, if you look at the second column, which is Foursquare. So, the columns are referred three different networks Foursquare, Google plus and Twitter. And if you look at the statistics which are in the rows, it is valid UGI which is user geographic information, valid AGI, valid geographic information and that is empty.

So, this basically would help you to find out, what is the amount of data that is available which is valid geographic information, valid and ambiguous geographic information, valid ambiguous geographic information and valid non geographic information and empty. What is this all mean I will I will try to explain this. Valid and ambiguous it is actually latitude or longitudinal, it is actually New Delhi; there is no ambiguity in it. Valid ambiguity it is not clear, so it says near Taj Mahal or near Govindpuri metro

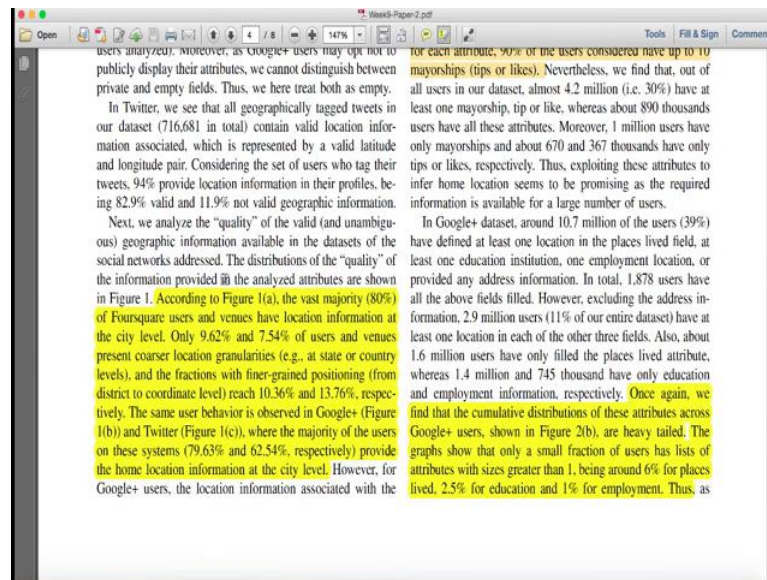
station, so these things are ambiguous. And non-GI – non-geographic information which could be I think as I said before, it could be somebody's heart, h e a r t.

And information like that is actually it is not geographic information at all, and sometimes it could be actually empty. So, essentially that is what is been given in the values. They are all percentages which says user home city is about 95 percent, ambiguous is about 2.6 percent, and non-GI is 1.8 percent, and empty is about 0.2 percent.

So, in Foursquare, it is user home city and venue city. In Google plus, it is places lived address and education and employment. In Twitter, it is the user location geo tagged tweet right. So, this basically tells you different types of information are collected from different networks; I mean that is a whole body of research in terms of actually using these different sets of information from different social networks.

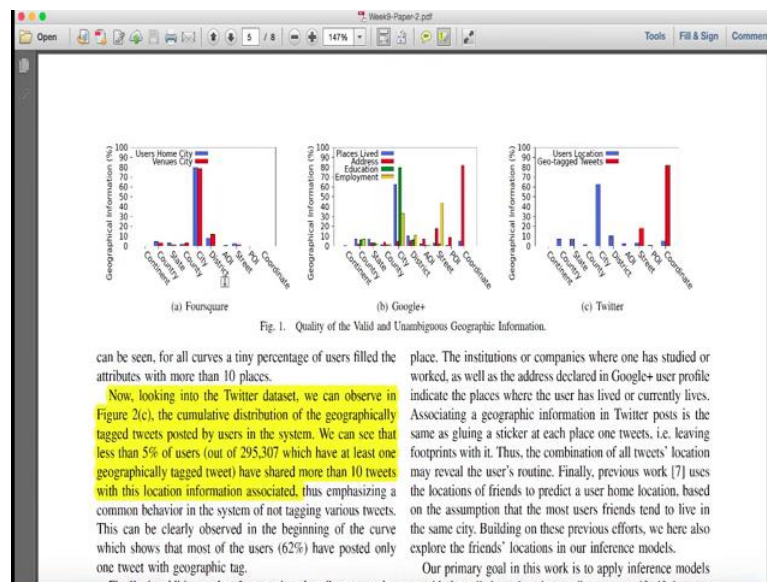
Then Foursquare it is user home city venue and venue. In Google plus, it is places lived address education and employment; in Twitters, it is user location and geo tagged tweet. So, if you look at the unambiguous geographic information for geo tagged tweet from Twitter it is about 100 percent. It is because all the tweets that where collected where actually at the 0.5 percent had geo tagged information in it, so that basically gives you a sense of what kind of data is collected. In terms of Google plus, 53 percent has an unambiguous education, so I studied that Carnegie Mellon University, so that is very very precise, there is unambiguity, there is no problem and actually recoding it or decoding it to a specific university.

(Refer Slide Time: 14:57)



So, let us look at different figures, different analysis that is been done using this data. Figure 1 the vast majority 80 percent of Foursquare users and venues have location information at the city level. 9.6 percent and 7.4 percent of users and venues present coarser location granularities at a state or the country levels.

(Refer Slide Time: 15:27)

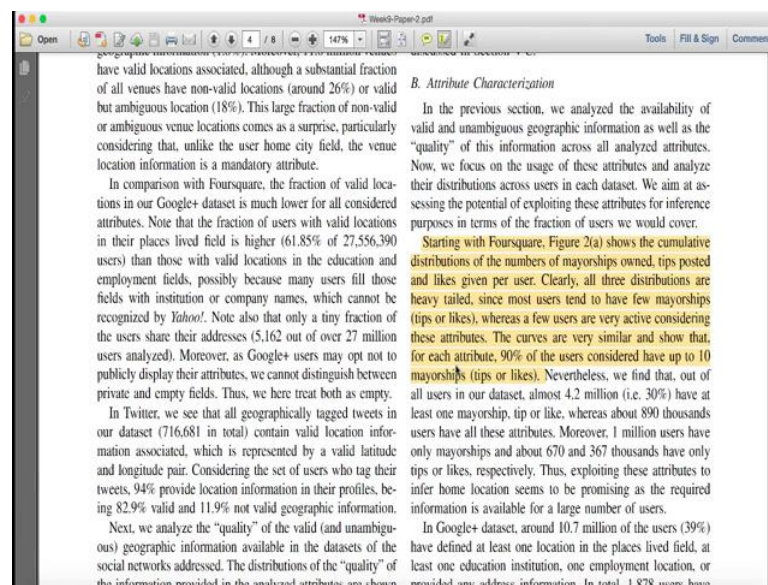


So, we look at the figures. So, this is figure 1(a), 1(b) and 1(c). So, if you look at here quality of valid and unambiguous geographic information. Foursquare, if you look at city, city gives you the users' home city and venue city; it is about close to eighty percent.

So, that is what is written here this says about vast majority 80 percent of Foursquare users and venues have location information at the city level. Some have at the country level, some have at the state level, some have at the street level, so that is the different level of details that the geographic information is available for the location from Foursquare.

So, if you look at Google plus, the information is maximum available for example, it is education that is available at a city level about 70 percent or 70 plus percentage, so that is what is its written here. The same user behavior is observed in Google plus figure 1(b) and figure 1(c) where the majority of the users of the system 79.63, 62.54 respectively provide the home location at the city level. City is highest in terms of places lived, city is highest in terms of user location also, so that basically says that we should be able to actually get the city level of information without any problem, because large amount of data for the information about the users is available at the city level.

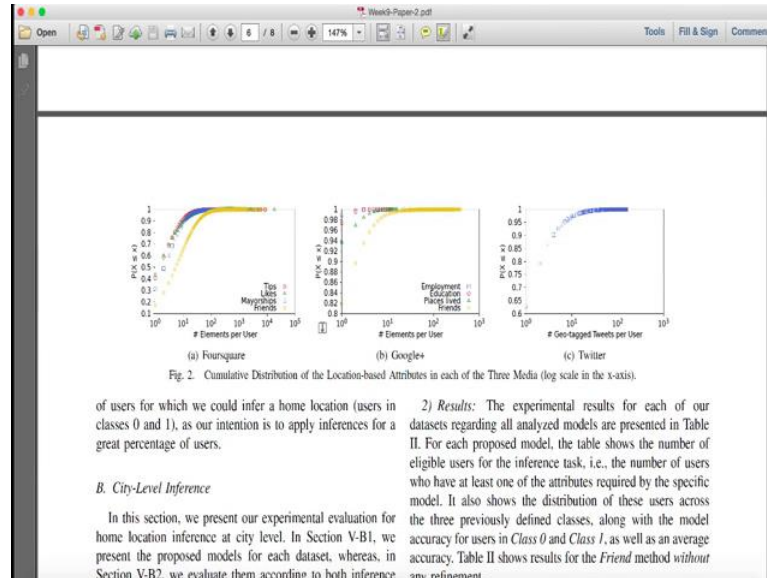
(Refer Slide Time: 17:36)



Now, let us look at more analysis with this data. So, now, figure 2(a) shows the cumulative distributions of the numbers of mayorships owned, tips, posted and likes. If you remember even in the last paper, we saw this kind of graphs, which is to show the cumulative distribution of the number of the mayors, tips and done right. So, if you remember the graph there were small set of people who had a lot of mayorships, and a

large set of people who had less number of mayorships, so that is the kind of general social media behavior also.

(Refer Slide Time: 18:40)

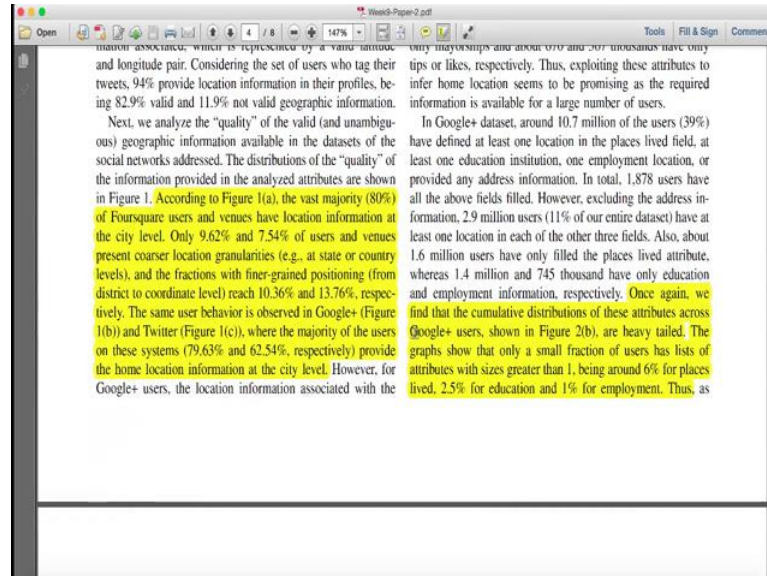


So, let us look at figure 2(a), which is giving you the distribution of the number of mayorships, tips and done. So, if you see here, the figure 2(a) is giving you the cumulative distribution of location based attributes in each of the three media; a is for Foursquare, b is for Google plus and c is for Twitter. Given that Twitter has all of them as geo tagged you can clearly see there is only one line, whereas in the other one there is, friends, mayorships, likes and tips that is in the Foursquare; in Google plus – employment, education, places lived and friends right. So, this is the graph and you can clearly see the graph is very similar to what we have seen in the past in terms of social media data.

So, clearly all three distributions are **heavy-tailed** (Refer Time: 19:31) which is what I just now said which is social media looking data, since most users tend to have few mayorships whereas a few users have very active considering these attributes. The **curves** are very similar and shows that each attributes 90 percent of the users considered have up to 10 mayorships right. So, it is the same principle Pareto principle that we talked about a power law that we talked about in the course all of that is playing into this data also. And this is very, very important to show because the reviewers and the readers can actually

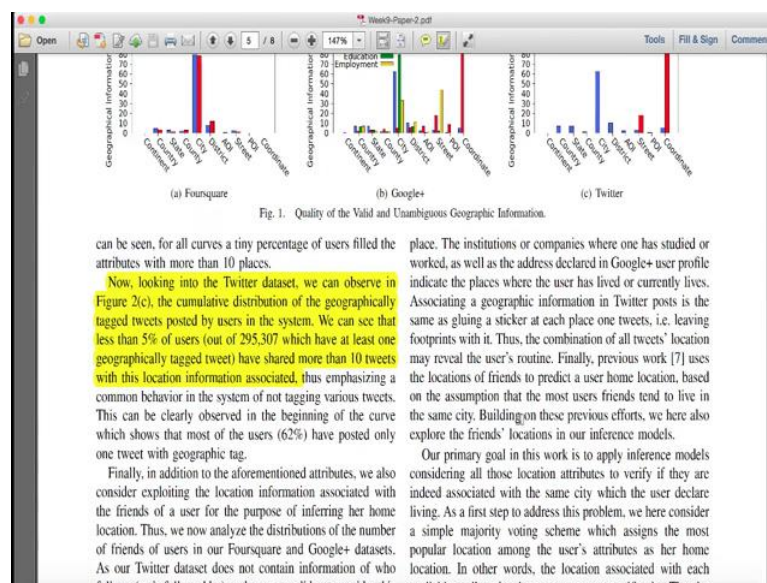
believe that this data is actually representative of other social media research that has been done and analysis that has been done.

(Refer Slide Time: 20:25)



Also if you see 2(b), again it's describing all the different social networks; figure 2(b) is showing you for Google plus the graph shows that only a small fraction of users has list of attributes with sizes greater than one being around 6 percent of places lived 2.5 percentage of education and 1 percent of employment.

(Refer Slide Time: 20:57)

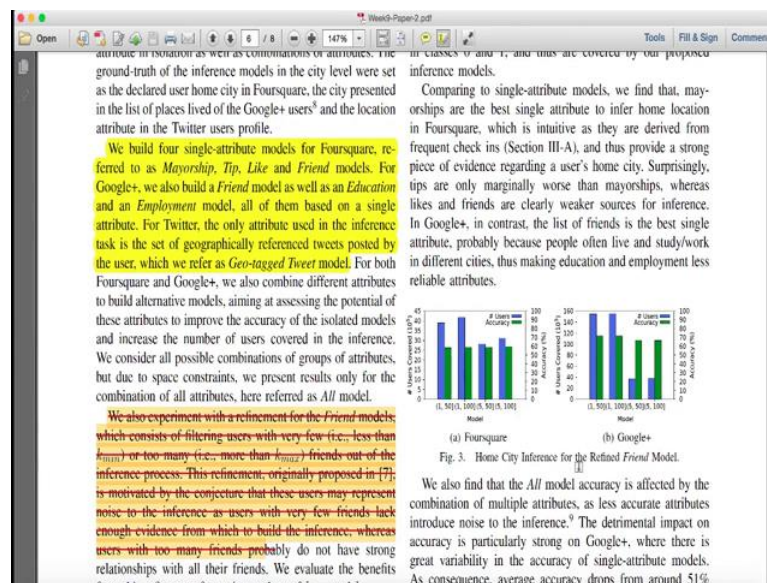


If you look at the Twitter data, we can observe that figure 2(c), the cumulative distribution of geographically tagged tweets posted by users in the system. We can see that less than 5 percent of the users have shared more than 10 tweets with this location information associated, which is again small percentage of people doing location information sharing, more than 10 tweets with the location associated with **them**.

Also now let us get into inferring location. The methodology that this paper uses is very same to the last methodology, which is written in this paragraph. We group users into three classes, class 0 consists of users who have only one vote that is only one location information that is predominant, and that is only one. Thus, allowing only a unique option to be assigned for the user's home city. Class 1 contains the user who has multiple votes with the predominant location across them.

And the class 2 as we have seen in the last paper also consists of users with multiple votes in which there is no single location that stands out. So, three categories of classes three classes that they are made 0, 1 and 2. We will see the table with 0, 1 and 2 that lets this locate the how the data was collected, how the analysis was done, how the bucketing was done. The results of our experimental evaluation are assessed using two metrics which measure the effectiveness of the proposed model. Accuracy is the fraction of correct inferences of users of class 0, or class 1, right yet again there are (Refer Time: 22:58) the current thing with class 2 will not work.

(Refer Slide Time: 23:04)



So, the model that was built was four single-attribute models for Foursquare, referred to as mayorship, tip, like and friends. For Google plus, friend model and education and employment model all of them based on single-attribute. For Twitter, the only attribute used in the inferences task is the geo tagged location right. So, basically this explains what details were used in collect and making the inference about the location.

(Refer Slide Time: 23:44)

We build four single-attribute models for Foursquare, referred to as *Mayorship*, *Tip*, *Like* and *Friend* models. For Google+, we also build a *Friend* model as well as an *Education* and an *Employment* model, all of them based on a single attribute. For Twitter, the only attribute used in the inference task is the set of geographically referenced tweets posted by the user, which we refer as *Geo-tagged Tweet* model. For both Foursquare and Google+, we also combine different attributes to build alternative models, aiming at assessing the potential of these attributes to improve the accuracy of the isolated models and increase the number of users covered in the inference. We consider all possible combinations of groups of attributes, but due to space constraints, we present results only for the combination of all attributes, here referred as *All* model.

We also experiment with a refinement for the *Friend* models, which consists of filtering users with very few (i.e., less than k_{min}) or too many (i.e., more than k_{max}) friends out of the inference process. This refinement, originally proposed in [7], is motivated by the conjecture that these users may represent noise to the inference as users with very few friends lack enough evidence from which to build the inference, whereas users with too many friends probably do not have strong relationships with all their friends. We evaluate the benefits from this refinement for various values of k_{min} and k_{max} .

³The Google+ users with more than one location in the places lived attribute were disregarded as the model ground-truth can not be dubious.

⁴In Foursquare this is not observed as the model that takes mayorships and tips into account has an average accuracy of 60.31%, while covering at least 24% more users in comparison with the models with attributes in isolation.

Figure 3 consists of two bar charts, (a) Foursquare and (b) Google+. Each chart compares the 'All' model (blue bars) with four single-attribute models: #Users (green), #Tips (red), #Likes (purple), and #Friends (orange). The left y-axis represents '# Users Covered (k)' and the right y-axis represents 'Accuracy (%)'. In chart (a), the 'All' model covers the most users and has the highest accuracy. In chart (b), the 'All' model also covers the most users, but its accuracy is significantly lower than in Foursquare, and it shows more variability across the single-attribute models.

Also this is also information we also experiment with the refinement of the friends' model which consists of filtering users. So, another way just think about the another way of looking at the friends model is the use the friends from that location to make the decision, so that is filtering with a very few, less than k kilometers, or too many that is more than k max friends out of the inference process.

The refinement originally proposed is motivated by the conjecture that these users may represent noise to the inference as users with few friends lack enough evidence for which to build the inference, whereas users with too many friends probably do not have strong relationships with all their friends. It is basically saying that we build a model where we take the users with very few, less than few kilometers; because they are not going to be connected. A lot of friends who may be connected from that location also will not lead a lot of information.

(Refer Slide Time: 25:03)

TABLE II
SUMMARY OF THE RESULTS OBTAINED FOR THE INFERENCE MODELS FOR HOME CITY INFERENCE.

Dataset	Inference Models	# Eligible Users	Classes Distribution			Accuracy		
			Class 0	Class 1	Class 2	Class 0	Class 1	Total
Foursquare	Mayorship	1,814,184	40.08%	46.74%	13.18%	51.61%	67.41%	60.12%
	Tip	1,589,430	43.62%	42.25%	12.13%	51.52%	67.29%	59.11%
	Like	1,194,907	45.76%	45.34%	8.90%	50.09%	61.74%	55.89%
	Friend (no refinement)	6,973,727	17.27%	61.56%	21.17%	33.03%	59.26%	53.51%
	All	7,153,078	16.69%	64.15%	19.16%	35.28%	61.03%	55.72%
Google+	Education	1,171,456	88.27%	1.30%	10.43%	21.17%	48.80%	21.57%
	Employment	639,265	92.41%	0.46%	7.13%	7.50%	22.29%	7.64%
	Friend (no refinement)	599,649	23.00%	25.97%	22.03%	40.43%	71.82%	50.89%
All	1,538,227	46.71%	16.02%	37.27%	17.13%	67.71%	30.22%	
Twitter	Geo-tagged Tweet	196,653	89.66%	10.34%	0.00%	82.50%	79.17%	82.16%

to only 30%. Nevertheless, these models achieve the largest user coverage, with about 1.5 and 7.1 million eligible users in Google+ and Foursquare, respectively. Thus, there is a clear trade-off between both metrics. Indeed, note that, despite a somewhat lower accuracy, these combined models make correct inferences for a much larger user population: about 3.2 million users in Foursquare and 291 thousand in Google+.

Similarly, we find that, in terms of accuracy, only the results for Twitter are far better than the best results for Foursquare, which in turn exceed those for Google+. However, the fraction of all users collected from Twitter that are eligible for inference (1%) is much smaller than the fractions in Foursquare (52.7%) and Google+ (5.5%).

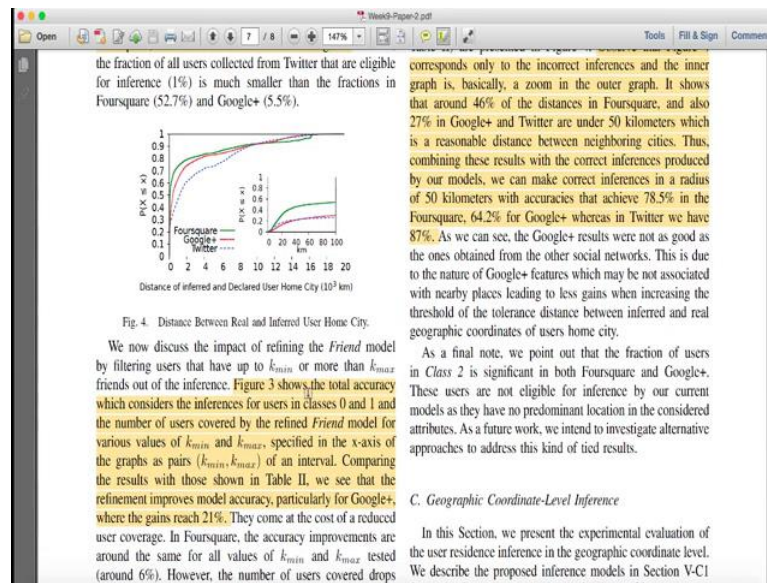
results for Google+ are quantitatively similar, although the reduction in user coverage is more significant (up to 88.2%) as we increase k_{min} to 5.

To better understand the errors in the models which led us to make erroneous inferences for users, we computed for each incorrect inference, the spatial distance between the inferred city and the one set in the ground-truth. The cumulative distribution of these distances for our most accurate models to each dataset (which their total accuracies are in bold in Table II) are presented in Figure 4. Observe that Figure 4 corresponds only to the incorrect inferences and the inner graph is, basically, a zoom in the outer graph. It shows that around 46% of the distances in Foursquare, and also 27% in Google+ and Twitter are under 50 kilometers which

So, this table is the most important analysis or inference from this paper which is to see the summary of results obtained for inference models for a home and home city inference. Remember, we did for three networks - Foursquare, Google plus and Twitter, the inference models that was used for mayorships, tip, like, friend all, education, employment, friend all geo tagged tweets.

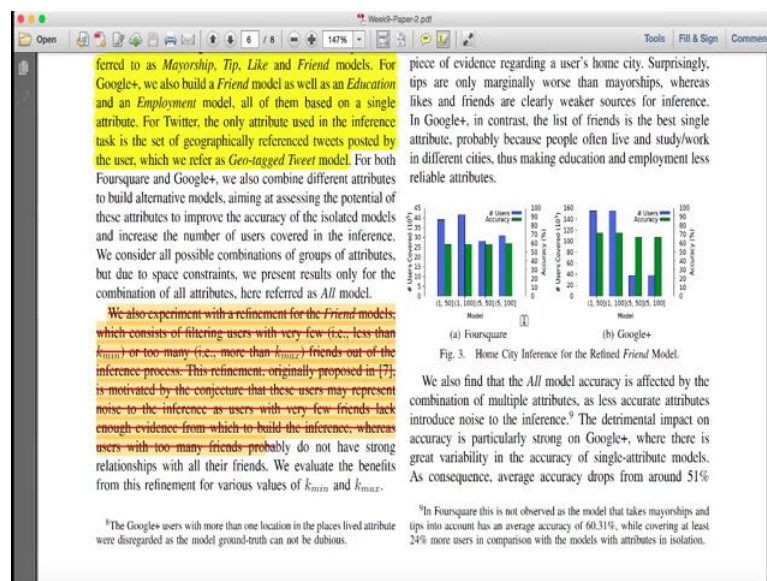
Classes distributed 0, 1 and 2; classes 0 and 1 are the only two things that can be done with this data (Refer Time: 25:45), so it is done 51. So, the way to read it is that just using mayorship, in the class 0, 51.61 percent you can identify the home city for that particular user in the category who has Foursquare account and mayorship data. Class 1, 67 percent; class 1 is basically there are multiple locations, one being predominant. Google plus, the highest seems to be with friend, no refinement; and in tweets, it is since the geo located the accuracy is also being more than anything else.

(Refer Slide Time: 26:43)



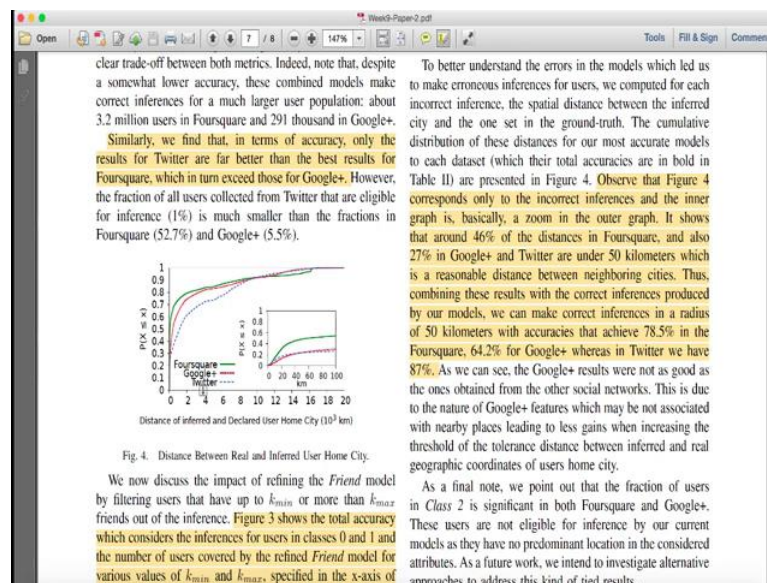
So, you can that is what I have said the accuracy for the Twitter seems to be higher than the rest of it. Let us look at figure 3, and then we will go back to the description of it. So, if you look at figure 3, figure 3 shows the total accuracy which considers the inferences for users in class 0, and 1. And the number of users covered by the refined friend model, for the various values of k min and k max specified in the x-axis of the graph. Often comparing the results with those in table 2, we see that the refinement improves model accuracy particularly for Google plus where the gain is about 21 percent.

(Refer Slide Time: 27:30)



So, essentially this is x-axis is the model that was used within the kilometers and y-axis is the percentage of users that were covered. So, if you see Google plus there is if you will infer the home city for a refined friend model, which is what we said where the min and the max were removed Google plus seems to be doing much better than the added advantage of removing these friends is higher for Google plus compare to Foursquare. That is the inference there.

(Refer Slide Time: 28:19)



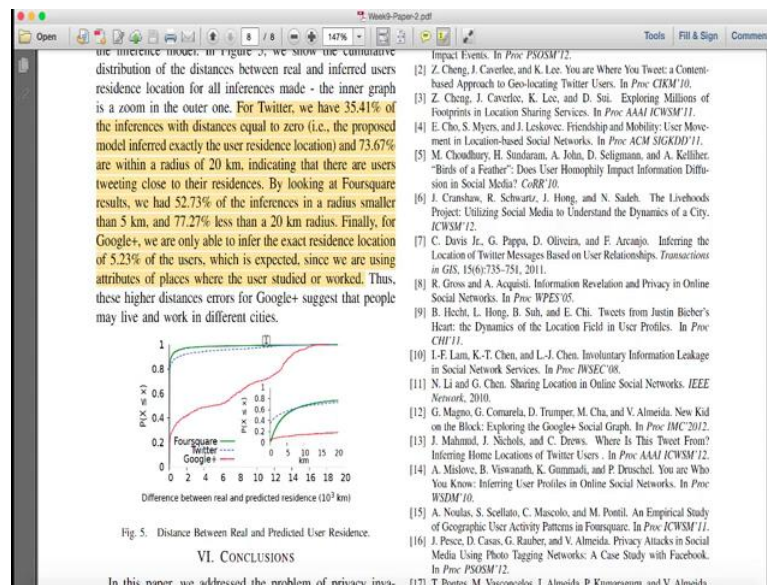
See there is another interesting inference, similar graph we saw on the last paper also. There we saw only for Foursquare; here we are seeing it for Foursquare, Google plus and Twitter. Figure 4, corresponds only to the incorrect inferences and the inner graph is basically zoomed into the outer graph. It shows that 46 percent of the distances in Foursquare, and also 27 percent in Google plus, and Twitter are under 50 kilometers. So, that is what is this here 50 kilometers is this part of the inside graph. 50 kilometers is reasonable distance between neighboring cities.

Thus combining these results with the correct inference produced we can make correct inferences in a radius of 50 kilometers with the accuracies that achieves 78.5 percent in Foursquare, 64 in Google plus and 87 in Twitter, which are the things that we saw in the table earlier. This is the representation in the graph which is x-axis is the distance of inferred and declared home city which is something that I declared. And something we were able to my account PK ponguru account has a location and I inferred through the

process the location what is the distance between these two, the lower the difference the better.

The inside graph is just showing you assumed immersion which shows that about 50 kilometers we were able to get about 46 percent, where if you see here, 50 percent and if this is about 46 percent. So, they just shows you that we are able to actually identify 46 percent of the distance in Foursquare is actually less than error of 50 kilometers, which is just neighboring cities, neighboring places, or sometimes it could be just in the same city.

(Refer Slide Time: 30:46)



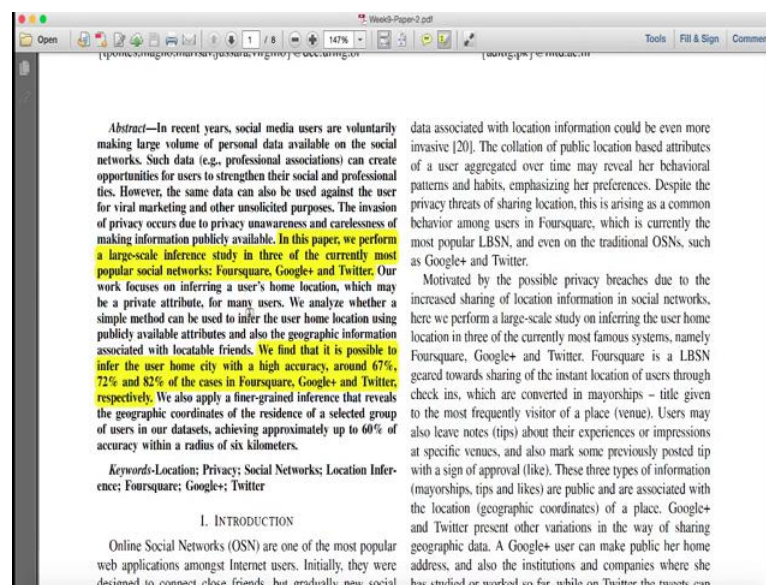
So, now the same thing you can actually do it for the residence right. Now what we did in this graph is basically showing you only the home city, whereas this graph is actually showing you the home residence. So, here is the graph for residence; red is Google plus; blue is Twitter; and green is Foursquare. You can see the inside the graph also here which is from 0 to 20 kilometers, whereas this is 0 to 20, but they are all ten to the power of 1000 kilometers is the distance here.

So, you can see that for Twitter we have 35 percent, where is Twitter, so Twitter is blue line, blue line is here, 35 percent of the inferences with distance equal to 0. So, that is the starting point here if you see, that is the proposed model inferred exactly the user residence. And the reason why this is so high and this is so accurate is because, we are, tweets were collected which were actually geo tagged right. And 73.67 percent are within the 20 kilometers radius.

So, if we see here this is 20 kilometers and if we see the blue line it is here that is about 76 percent, 73 percent, which is within the 20 kilometer difference, we were able to find out where the home is which is pretty good. Indicating that that **there are** users tweeting close to their residences, because I could be living in **Okhla**, I could be tweeting from somewhere near Jawaharlal Nehru Stadium in Delhi which is less than 20 kilometers, I went to watch a match and I actually posted tweet which is also **geo tagged**, so that is the kind of 20 kilometers that we can get.

By looking at Foursquare results, we find that Foursquare is green. We find that 52 percent of the inferences in the radius smaller than 5 kilometers. So, if you see here green one, if you go at that point it is about 52 percent, 52.73 percent less than 5 kilometers; 77 percent less than 20 kilometers, that is here, 77 percent. Finally, for Google plus, we are only able to infer the exact residence of 5.23 percent of the users, which is expected since we are using attributes of places where the user studied or worked, because here we are only using their employment and education details right.

(Refer Slide Time: 33:55)



So, that is how this paper ends, which is to show that let us go to the abstract again, which is to show that they used they used data from Foursquare, Google plus and Twitter. They used this data to infer the home location. This is an extension or the next step for the last paper that we saw which was done only on Foursquare. And they were able to actually show that about 67, 72 and 82 percent with that accuracy they were able

to find out the home city, and home location, for, with a high accuracy in terms of Twitter and then Foursquare, but with less data in a Google plus.

With that, I will stop this particular paper. I will see you soon.