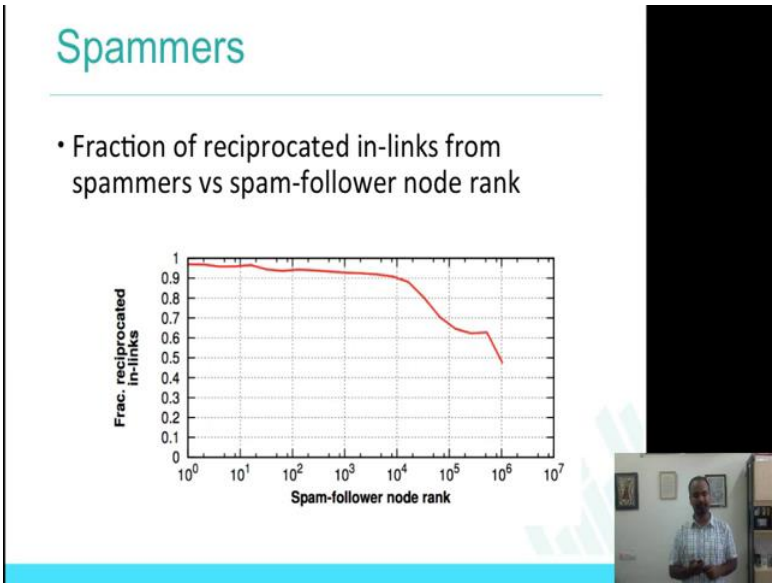
Privacy and Security in Online Social Networks
Prof. Ponnurangam Kumaraguru ("PK")
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Week - 7.1
Lecture - 22
Link Farming in Online Social Media

Welcome back to the course Privacy and Security in Online Social Media. So, this is week 7, I hope you got a chance to look at the content in week 6 where we looked at link farming spam in Twitter and some kind of work which was able to find out what link farmers are what is the characteristic of link farmers. So, today we will continue, this week we will continue a little bit about the same topic, we will finish it and move onto something else.

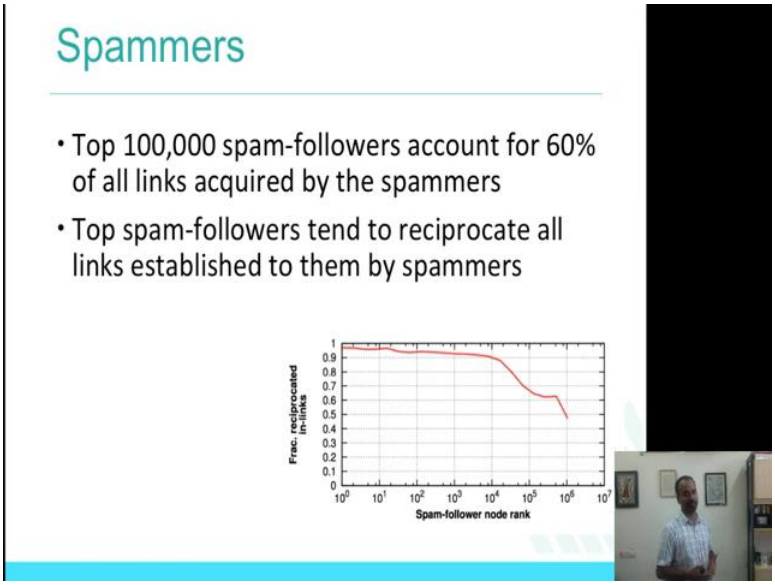
(Refer Slide Time: 00:44)



If you remember last week, I showed you about what is link farming and some data about link farming. So, here is a graph which has on x axis spam follower node rank which is what is the probability that spam followers are the accounts which actually follow spam. If you remember the graph that I showed you a, b, c, d, e, f where there was something as spam follower and then something that was spam followings.

So, x axis is spam follower node rank which is the rank of the account, which is being a spam follower and then on the y axis is fraction reciprocated in links, which is I think we had briefly mentioned this last week also, the probability of you following me when I actually follow you, reciprocity so to say. If I follow you what is the probability that you will follow me back that is what is put on the y axis. So, fraction of reciprocated in links from spammers versus spam follower node ranks.

(Refer Slide Time: 01:56)



That is if you look at the top 100,000 spam follower accounts for 60 percent of all links acquired by the spammers. So, what does this mean? This means top 100,000 spam follower accounts for 60 percent of all links acquired by the spammers. So, if there were 100 links that were created for the spammers, 60 percent of them are coming from the spam followers, which is an account which actually follows back the spammers. Top spam followers tend to reciprocate all links established to them by spammers.

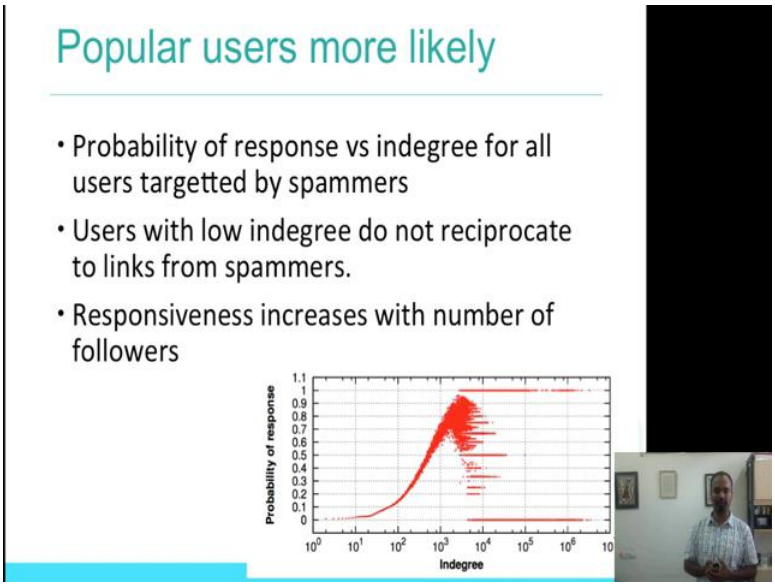
So, if you look at this graph the way to read this graph is on the x axis is the log scale, which is 1, 10, 100, 1000, 10000, 100000 and 1 million. So, that is how it is written on the x axis. y axis is the probability of you following me back, if I follow you. So, if you look at the first let us take a look at the first value 1 or 1 to 10. So, here the probabilities is almost close to one which is the top ranked spam followers which is the accounts

which are having the chances of following you back is very high. If you arrange the spam followers in the node rank which is the number of followers that they may have is actually very high.

So, what does this mean? This basically means that the probability of a spam follower following a spammer is very high, which is what spammers actually make use of, which is if there is a probability, if there is a chance that you will follow me back, the spammers will keep following people like you and there is a high chance that you will follow me back and therefore, spammers increase their in-links, which means on the topic that we discussed about link farming which means that my node rank is increasing, which is, spammers' node rank is increasing which is what they want because of that their content will show up on search, their content becomes more popular and therefore, they will probably benefit from it, I hope that connects the dots.

So, I am going to use the same data, to actually emphasize on what is the behavior of these spammer? And what the link framers do? What do the spam followers do?

(Refer Slide Time: 04:38)



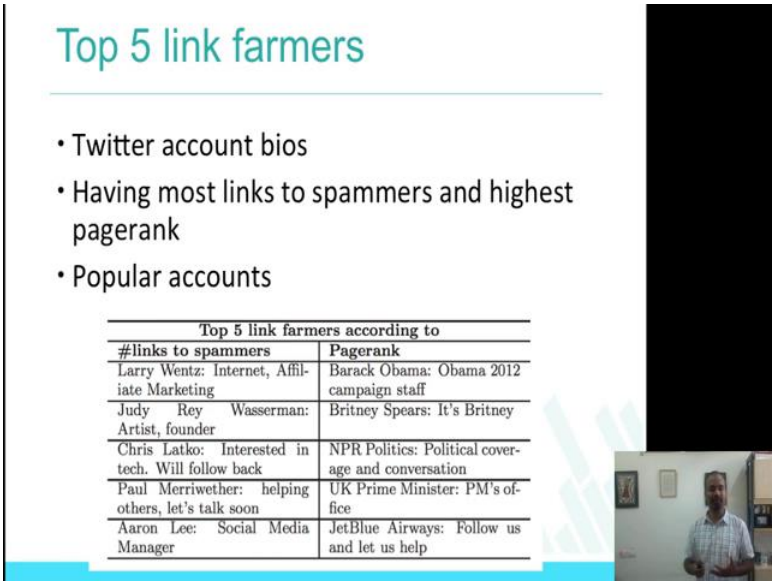
Let us look at another graph. This graph is showing the probability of response which is in terms of just responding to a request with the in degree, in degree is number of links

that they have. So, this is showing you probability of response was in-degree for all users targeted by spammers. If there are maybe, 10 to the power 7 users were targeted by the spammers, what is the probability that they are going to actually respond?

Users with low integrity do not reciprocate to links from spammers. If you look at the graph let us take less than 100, less than 1000 which is if I have followers which are less than 1000,, there is very less probability that I will actually follow the spammer back. Let us look at in the graph again, 10 to the power 3. Let us look at the value 10 to the power of 3, the probability of that users even then 10, 100 to 1000 if you see, the probability of somebody following back - the spammer is actually about 40 percent, 50 percent around 60 percent.

If you look at the later part of the graph on the x axis and in the y axis which is about 0.7, responsiveness increases with number of followers, in-degree is the number of followers. As the followers increase, the chances of somebody following you back when you send a request is high. I hope that is sinking in, again let me reiterate the point which is, on the x axis in this graph we are seeing the in-degree which is the number of followers, y axis is the probability of response when a request is sent for following or when somebody follows you, chances of you following me back. Users of less in-degree do not reciprocate to spammers; whereas, users with larger in-degree which is larger number of followers tend to follow back more.

(Refer Slide Time: 06:52)



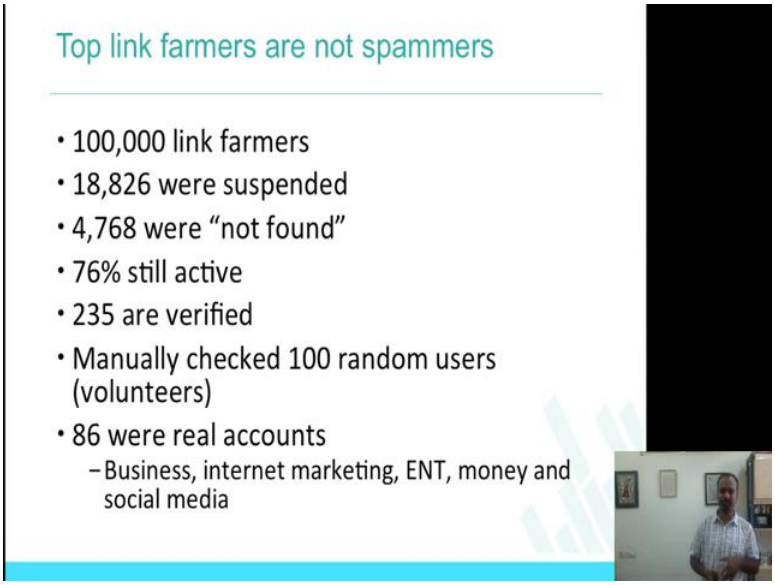
So, what they did was after looking at these two things, which is the probability of somebody following spammer part is high, they actually looked at the top five link farmers and looked at the bios, what are the accounts and here is a sense of what these accounts are Larry Wentz, Internet affiliate, Marketing; Judy Rey Wasserman, artist, founder.

So, these are all the accounts which had the links to spammers; top five link farmers according to the links to spammers according to the Pagerank; the word Pagerank is nothing, but the links that you have to be out which is the in degree and the out degree that is what is PageRank is. Chris Latko, interested in Tech, will follow back and Paul Merriwether, helping others, let us talk soon; Aaron lee, social media manager. So, it just basically shows you what kind of users of the top five link farmers which is creating these links to others and then getting others to follow you back. Internet, social media manager will follow back; these are the kinds of accounts.

If you look at the pagerank which is, higher the links of how you are linked to others also this Barack Obama, Obama 2012 campaign staff; Britney Spears; NPR politics; UK prime minister; JetBlue Airways. So, this is also showing that it is not just the real spammers, but malicious intention, they actually doing link farming; even legitimate

accounts even more so popular accounts are actually part of the link farming ecosystem and they increase their followers. That is a kind of revolution that they wanted to actually get which is to look at the bio, the accounts of link farmers and have some understanding of what kind of users these are.

(Refer Slide Time: 08:57)



Interestingly top link farmers are not the spammers. They looked at top 100,000 link farmers at the point of analysis of which of course, 18826 were suspended. Twitter did something, figured that all these are actually spammers, these are actually malicious accounts. Therefore, they suspended it. 4768 accounts were not found which is that they probably deactivated the account the account does not exist, whereas, 76 percent of the account of 100,000 link farmers which is, how do they get this hundred thousand link farmers; they do the graph of node rank, they do the graph of what, who were the most popular link farmers and they got this hundred thousand link farmers and of this 76 percent were still alive.

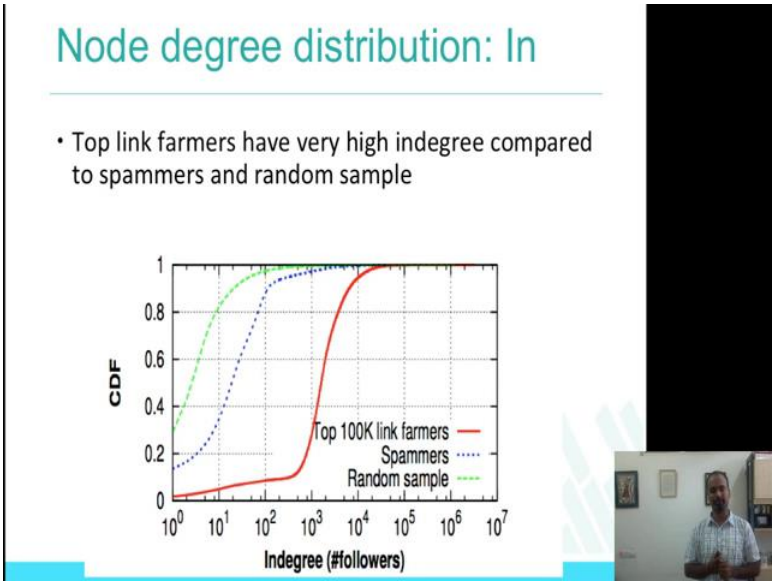
Interestingly of that, 235 were verified accounts, if you remember verified accounts are the accounts which has a blue tick next to the account and these are the accounts which are legitimate that is they know, they show that they are the real people which is Amitabh Bachchan has the real account, verified account; Obama has a verified account which are

the real people who they say they are.

They manually checked 100 random users of 235, but volunteers of course, they got some user volunteers to verify whether to look at these 100 random users and said something about the users. They found that 86 were real accounts, they actually got more than one people to look at it, therefore if more than one person says that it is a real account, there's a high probability that its a real account. They actually found that of 86 real users, people were like had the account as business, internet marketing, entrepreneurship, money and social media. These are the topics that the 86 real accounts are talking about. It just gives you the sense of, it also connects very well to the Twitter account bios that we saw in the slide which is top five link farmers.

So, this shows that the top link farmers are not really the ones who are standing in real world, but they could be actually, they are actually real accounts, they are actually verified accounts

(Refer Slide Time: 11:27)



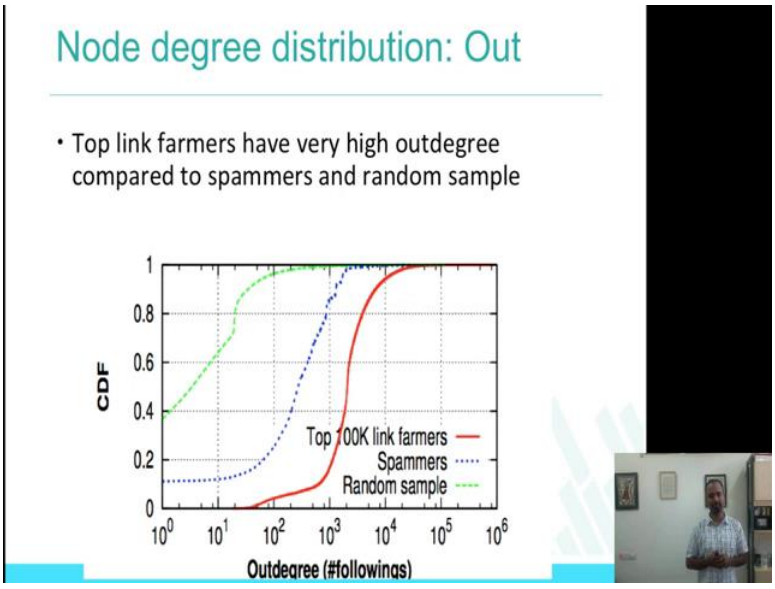
Let us delve little bit more into the node degree distribution which is how the in-degree followers are for top 100,000 link farmers, for spammers and for random sample. Well what is the goal here? The goal here is to try and compare the hundred thousand link

farmers they found with spammers, which are real spammers; their accounts that are recorded and random sample of users, if they compare these three types of users, the observation can be very helpful to understand what is the property of these 100,000 link farmers..

If you look at this graph, this graph basically shows that top link farmers have very high in-degree compared to spammers and random sample. So, let us go through again the graph in detail; x axis is in-degree, which is again log scale which is 10 to the power of1 to 10 to the power of 7, that is the number of followers, cumulative distribution frequency CDF is on the y axis. The way you look at this is that the more the value on the red graph, red line is which is if you look at the 1000 users which has 1000 followers which is in degree which is very high. If you look at that, the CDF is about 0.1.

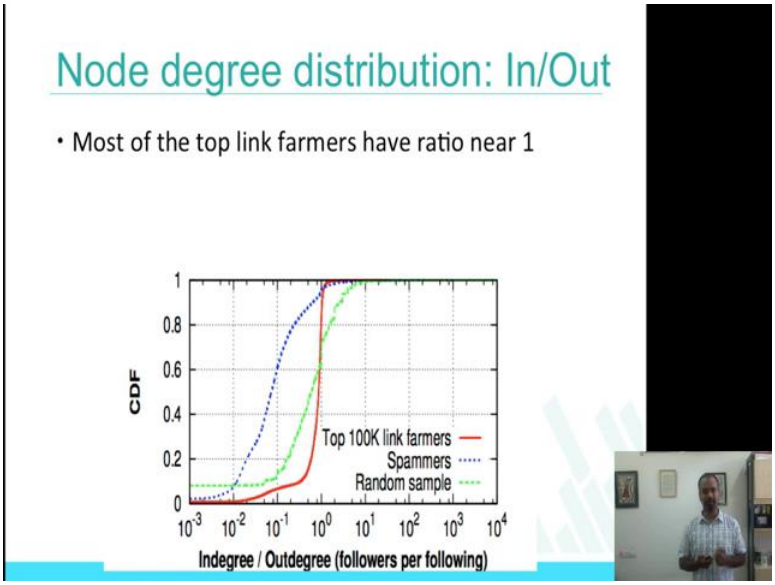
So, top link farmers which is, if you arranged the in-degree in particular order, the top link farmers have very high in-degree compared to spammers which is the graph for the spammers and the random samples is very different and also the CDF is actually very low within the in degree, which is 10 to the power of 2, 10 to the power of 3. Top link farmers have very high in degree compared to spammers and random sample.

(Refer Slide Time: 13:22)



Interestingly, they also found that for the out-degree also, the graph looks very similar which is, the number of followings that I have and the top link farmers have very high out degree compared to the spammers and random sample and slightly very different from, slightly different from the in degree graph, but still again the 100,000 link farmers graph is much higher in terms of out degree compared to the spammers and the random sample.

(Refer Slide Time: 13:57)



They also did an interesting analysis on finding the ratio between in degree and the out degree, this ratio is actually very useful because if you look at really large followers account like Amitabh Bachchan or Obama, the number of accounts that follows them will be very high versus the number of followings that they have is actually very low. That is one parameter, one way to look at the legitimate accounts. If you look at again legitimate accounts like mine, probably the number of followers and the number of followings\ are actually very close to each other.

So, this is what they want to find out which is, if you find the ratio of in verses out, how is this; what is the pattern of the users; most of the top link farmers have a ratio near to 1, if you look at the graph 10 to the 0 is 1 and interestingly, the value for the top link farmers, why is the ratio between followers and the followings is equal to 1, which is the

pattern that I was telling you for real users like mine. So, therefore, link farmers also have this similar behavior. So, the other point to take away from it is given all this, it is going to be hard to find out who is the link farmer that is the kind of intuition that is build behind all this analysis.

Again, if you see this graph x axis is so, 10 to the power of minus 1, minus 2, minus 3 is where the in-degree verses out-degree ,when the out is much larger than the n would be the one that are less than 1. So, you can clearly see that the top link farmers which is the red color has the ratio of one, whereas, if you look at spammers and the random sample they are not really one, there is some difference with the examples that I took like Amitabh Bachchan, Obama and myself. So, I hope that makes sense in terms of what distribution is in-degree, out-degree distribution? What is the ratio of in degree versus out degree? It give you a sense, go through the slides, go through the materials and if there is any confusion or any more clarifications needed, feel free to post it on forum, I will be actually happy to help in understanding these content also.

(Refer Slide Time: 16:16)

Now, if you look at the bio of top 100,000 link farmers and random sample, just to get an comparison of what are the people, what are the accounts which are actually link farmers and what are the accounts which are random sample talking about? The left one is

actually the link farmers, the right one is actually random sample. You can clearly see here that the left one is talking more about market, online, internet, social, love which probably is in random sample also - life, music, live, love, right.

So, the one some conclusion that they do from this analysis is that on the left you see LF, you can see that promoting their own business or content or trends in a domain, links to legitimate external sources. Of course, they are basically talking about some business, talking about some links that are outside twitter, outside the network that they are promoting this content. That is is right, do not tweet to external sources which is, there is not a lot of links to other sources.

Again this will be a pattern that this research formed, but the pattern if you like to study this today, it may be very different also. I am kind of looking at some of these classical worlds which looks at some of the question that we should be asking in privacy and security in online social media topic.

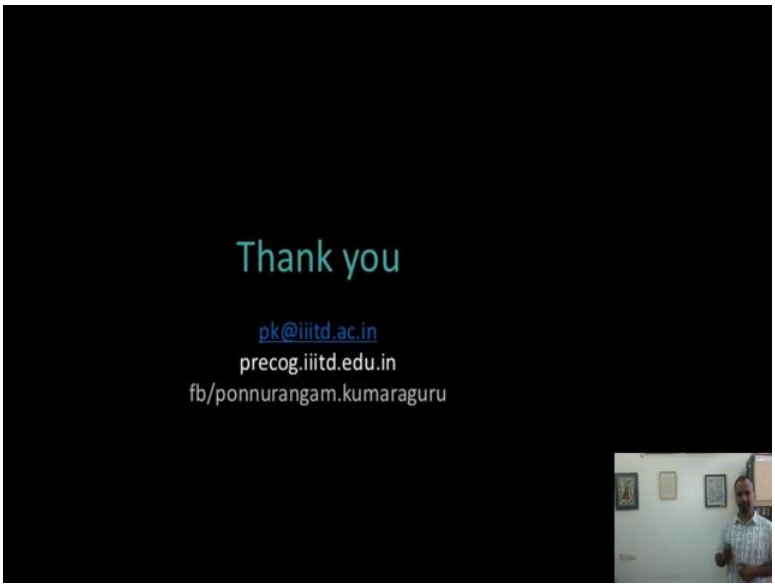
(Refer Slide Time: 17:43)

So, the final conclusion for this part of the work is, this part of the course / lecture is characteristics of link farmers we found the in degree verses out degree ratio is actually pretty high. The in-degree is very high compared to spammers and random sample. The

out-degree is also very high; the probability of spam followers is also very high in terms of requests sent to them or if I follow you, there is a high probability that you will follow me if you are a spam follower. Surprisingly, legitimate popular and highly active users such as bloggers and experts, most likely engage in link farming, these are accounts like Britney Spears, Obama all of these accounts actually have link farming behavior.

So, the problem is that this increases the social capital and the influence because if the link farming engage, if the concept, I will go back to the first slide again on this topic. If there is high probability of, if there are chances that you let your social reputation, the links between the users are higher, their social recognition are I mean today social reputation, influence is all measured by number of followers you have and the kind propagation of of the content that you have and therefore, link farming can be pretty effective in terms of increasing your social capital and influence.

(Refer Slide Time: 19:10)



With that I will actually stop this particular part of the lecture, which is week 7.1.