Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 06
Lecture - 01

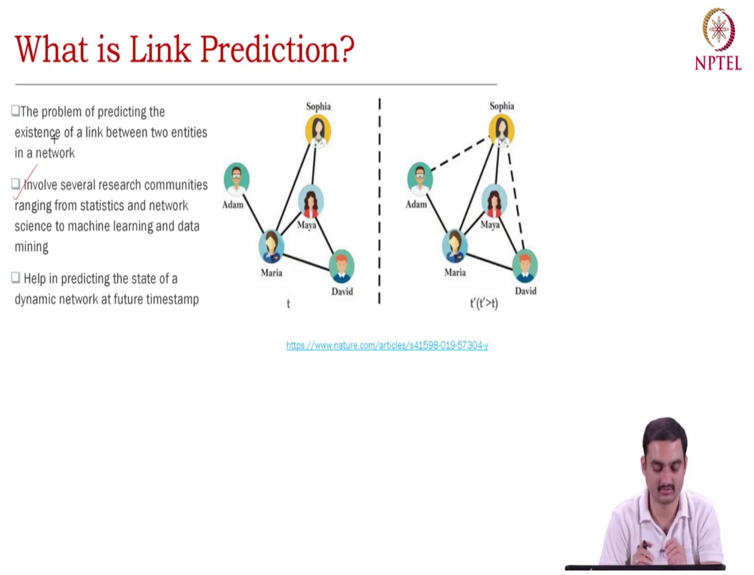
So, welcome to the 6th chapter of the Social Network Analysis course and to before starting this chapter I would like to congratulate you all to you know finish the half of this course. So, we have actually finished 5 chapters right and this is going to be the 6th chapter. So you know by now we have understood concepts like link analysis, community detection right, network growth model and you know different types of measures that we generally use for, link measurement node measurement prestige importance right and so on and so forth.

So, this is basically the beginning of new category of chapters that we are going to discuss more of an application kind of you know aspects of social network analysis. And the first application that I mean everybody talks about when it comes to graph analysis graph mining is link prediction ok. So, we will discuss in this chapter we will discuss what do you mean by link prediction.

We will formally define the link prediction problem and this is a very very well studied problem, so if you look at the literature there are thousands of papers on link prediction. We will try to understand you know applications where link prediction algorithms can be used and then we will look at algorithms for link prediction, simple algorithms most of the heuristic based algorithms we will discuss.

And then we will go into the deeper of you know deeper discussion of link prediction algorithms we will discuss you know cases where we use machine learning algorithms, a bit heuristics plus machine learning models in a kind of an hybrid setting and so on ok.

(Refer Slide Time: 02:16)



What is Link Prediction?

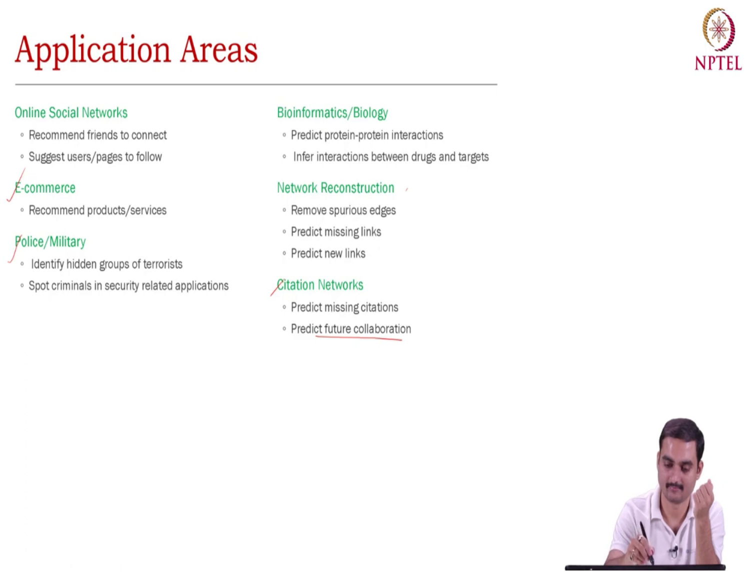
- ❑The problem of predicting the existence of a link between two entities in a network
- ❑ Involve several research communities ranging from statistics and network science to machine learning and data mining
- ❑ Help in predicting the state of a dynamic network at future timestamp

https://www.nature.com/articles/s41598-019-57304-y

So, let us gets started now what is link prediction as the name suggests, in fact I use the term link prediction I think multiple times in the last few chapters. So, you must have already understood the essence of link prediction, but nevertheless let us formally define what is link prediction right. So, link prediction is a problem of predicting the existence of a link between two entities in a network right. We basically try to predict whether these two nodes a given set of nodes a given pair of nodes are connected or not through links through a particular link right.

So, it basically involves research you know research communities ranging from statistics and network science and machine learning and data mining you know domains to work together to come up with solutions for link prediction.

(Refer Slide Time: 03:11)



Application Areas

Online Social Networks

- Recommend friends to connect
- Suggest users/pages to follow

E-commerce

- Recommend products/services

Police/Military

- Identify hidden groups of terrorists
- Spot criminals in security related applications

Bioinformatics/Biology

- Predict protein-protein interactions
- Infer interactions between drugs and targets

Network Reconstruction

- Remove spurious edges
- Predict missing links
- Predict new links

Citation Networks

- Predict missing citations
- Predict future collaboration

And the applications right. So, broadly if you look at applications it can be used in social network in online social network analysis, in general it can be used in e-commerce you know platforms, some sort of law enforcement and surveillance prospective as well bioscience biological networks ah, network reconstruction citation network and so on.

So, in online social network link prediction can be used for frame recommendation right for example, in case of Facebook essentially when you recommend a friend right when Facebook recommends you a friend, Facebook basically predicts that you and that friend; so you and that friend are going to be connected in future right. And Facebook always tries to recommend you with friends which are highly likely to be I mean recommends you with users who are highly likely to be your friend right.

So, false positive I mean Facebook always tries to increase false positive right. In on in again in online social network you can see again cases like twitter where you know all your followers are recommended. So, you may follow this user you may follow that user and so on and so forth this is also a link prediction problem, but in a directed network. It oftentimes also suggest pages right say for example, a story or a news article that you might be interested in right.

You may see in your social media social network feed that oftentimes some news article links pop up and you are absolutely uncertain that why suddenly this has arrived right and sometimes it also happens at least in my case is that you know things that are recommended are completely unusual right. Say for example, you know some days back you know I was talking about some sort of you know Bengali sweets right.

So, I was just chatting with my wife and suddenly when I opened the Facebook feed I saw the advertisement of the same sweet right. Now I do not know whether things are being recorded or not, but in general you know this social media pages always try to recommend you with items or products or news feeds which are relevant to you right.

This is also link prediction because if you think of web as a whole, social network is a part of the web then you have news articles right, blogs, other online advertisements and ultimately the task is to link between user and that item right. User and advertisement user and news feed right news article blogs ok, but in an heterogeneous network setting.

E-commerce of course, when you look at platforms like say Amazon or Flipkart products are being recommended to you right. And we have discussed already briefly that how this recommendation engine works right. Once you know the community structure clusters and once you know that within a community these the other nodes have already showed interest to on these products it is highly likely that you will also be interested to buy this product right.

So, recommendation system is essentially a link prediction link prediction problem right of course, in case of law enforcement agency police military identifying hidden groups hidden identities hidden links right spot criminals these also come under the broad purview of link prediction.

Sometimes what criminals do criminals try to you know hide their identities, they do not interact with each other quite frequently. So, that I mean if they interact with each other frequently their call may be trapped their interactions may be trapped. So, therefore, they tend to hide themselves right they tend to avoid in introductions in an in a in a frequent manner right. Now how do we predict such interactions when there is less history right, so you can use link prediction for that.

In biological network particularly protein interaction network can we predict that these two types of proteins will be will interact in future. So, in history if you look at the history there is no such evidence or less evidence about the interactions of proteins and no one knows like in the future possibly some new proteins or now new entities might be might start interacting ok.

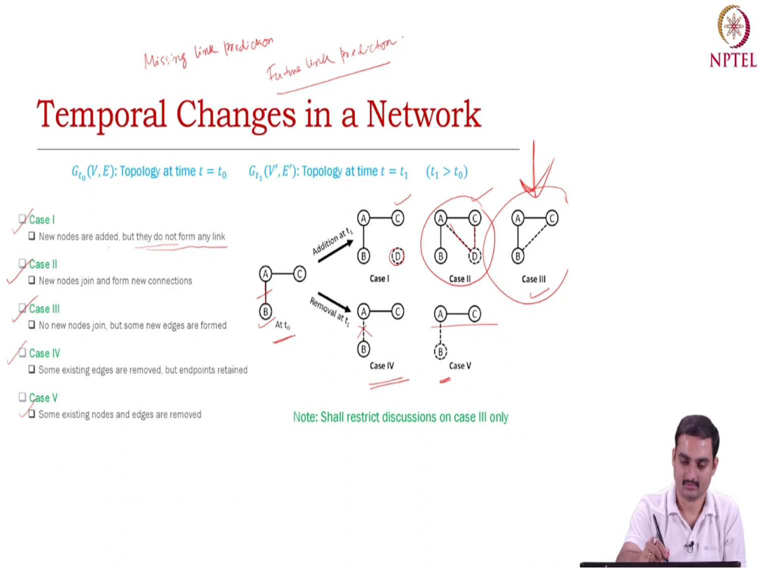
In case of network reconstruction right, say when we scrapes data from social network what happens is that we generally; you know we generally are unable to scrape all sorts of data sets right. Say for example, when you scrape Twitter only allows you to scrape say 1 percent or 0.5 percent of the data right. Now your data will be incomplete right now, it may not be recommended to design an algorithm on an incomplete data set.

So, then what you would do you try to complete it you try to complete the network as much as possible. So, you know before running your algorithm you want to run a link prediction algorithm and want to predict the links which might have been missed right while scraping the data set right.

So, you may want to run link prediction algorithms try to make the network complete as much as you can and then you run your favorite algorithm say clustering algorithm or other say anomaly detection algorithms and so on. Citation network particularly in collaboration network it is helpful because say for example, there is a researcher collaboration right.

If you predict that ok these two researchers are going to collaborate in future you can also suggest you can also recommend, the user the researcher we set up other researchers that he or she might be interested in collaborating ok. So, in future research collaboration link perdition can be used you know massively.

(Refer Slide Time: 10:10)



Missing link prediction

Future link prediction.

Temporal Changes in a Network

$G_{t_0}(V,E)$: Topology at time $t = t_0$ $G_{t_1}(V',E')$: Topology at time $t = t_1$ $(t_1 > t_0)$

- □ Case I
  - □ New nodes are added, but they do not form any link
- □ Case II
  - □ New nodes join and form new connections
- □ Case III
  - □ No new nodes join, but some new edges are formed
- □ Case IV
  - □ Some existing edges are removed, but endpoints retained
- □ Case V
  - □ Some existing nodes and edges are removed

Note: Shall restrict discussions on case III only

Now, let us look at the see let us try to understand these two problems. So, one is called link prediction one is called missing link prediction and other is called future link prediction ok, so missing link prediction and other is future link prediction. Apparently they may sound alike right, but there is a very small you know very tiny difference between these two problems ok.

Missing link predictions basically it says that you are given a static network, you are given a static network ok and in the static network due to crawling problem due to privacy issue you may not be you know you may have missed you may have missed out some of the edges ok in the static network. Can we can we try to complete the network by predicting the missing edges it mostly it is related to static network given a network you predict the edges which should have come, but due to some problems crawling problem or network construction problem.

They have not been you know added future link prediction deals with temporal network ok. So, the idea is that you have a graph till time t ok can we predict whether a pair of nodes x y are going to be connected at t plus delta t time period ok. This delta t time period this slack right this window you do not have any idea about ok, you only you are only given a network till time t you may have snapshot you may have also been given snapshots of networks.

Say at time 0, time 1, time 2, time 3, dot dot dot time t and you predict your prediction would be what would be the you know which pairs are going to be connected at t plus 1 right. This is future link prediction we need to look at the time temporal dimension very carefully. Oftentimes when we do research on link prediction we kind of you know mix and match these two things together, but that should not that should not happen that that should be avoided ok.

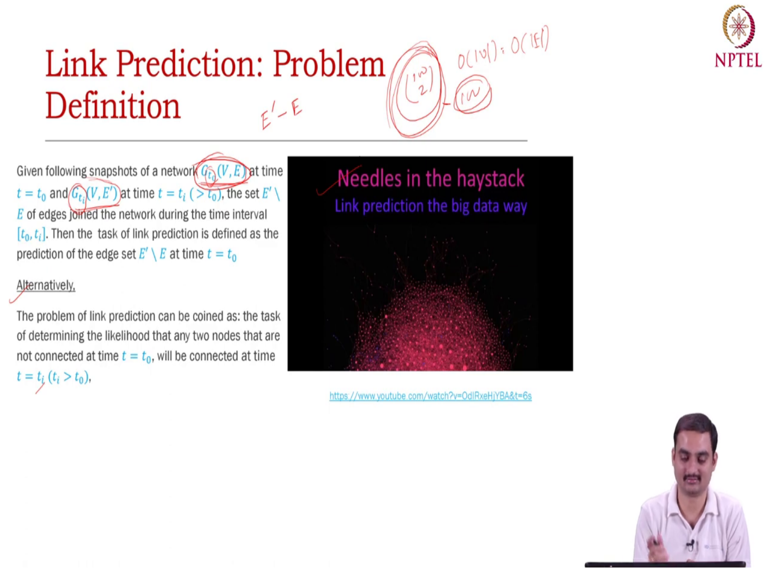
So now, let us look at the temporal changes in a network. So, there can be five possibilities when you look at the temporal evolution there can be five possibilities. So, you are given a network at time t 0. So, in case 1 new nodes are added, but they do not form any link just nodes are added this is case 1 ok.

So, at time t 1 this new node D is added and, but there is no new edge right this is case 1. What is case 2? New nodes join and they also form new connections this is case 2. D has joined and D has formed two connections with C and with A. Case 3 no new nodes join some new edges are formed this is case 3, no new node but B and C are now connected.

Case 4 some existing edges are removed this is case 4, earlier there was an edge between A and B now there is no edge between A and B ok, so edges are being removed. And case 5 some existing nodes and edges are also being removed. You see here N is being removed therefore, this edge is also being removed right; so you have only A and C.

These are five possible cases right. But when we talk about link prediction we will only focus on this part, in fact this part also right. So in fact, most of the times we actually focus on this part because we assume that the set of nodes is fixed ok we try to predict whether within the existing nodes which edges are being which edges will be formed in the future ok. So, in the remaining part of the chapters we will only focus on this scenario case 3.

(Refer Slide Time: 14:56)



Link Prediction: Problem Definition

$E' - E$

$\binom{|V|}{2} - |V|$

$O(|V|) = O(|E|)$

Given following snapshots of a network $G_{t_0}(V,E)$ at time $t = t_0$ and $G_{t_i}(V,E')$ at time $t = t_i\ (> t_0)$, the set $E' \setminus E$ of edges joined the network during the time interval $[t_0, t_i]$. Then the task of link prediction is defined as the prediction of the edge set $E' \setminus E$ at time $t = t_0$

Alternatively,

The problem of link prediction can be coined as: the task of determining the likelihood that any two nodes that are not connected at time $t = t_0$, will be connected at time $t = t_i\ (t_i > t_0)$,

Needles in the haystack

Link prediction the big data way

https://www.youtube.com/watch?v=OdIRxeHjYBA&t=6s

Now this is kind of a drawbacks, I mean you may wonder why we do not want to make it generic, why do not we also predict that node which node is, which node is going to going to be going to join that is not possible right. At time t standing at time t how do you know that which node is going to be going to be joining right. But of course, deletion link deletion prediction is an important problem which is not there in the literature, so that can be a good research topic right to study.

So, now let us define link prediction formally given a snapshots given a following snapshots of a network G t 0 V comma E, t 0 is at a particular time t particular time t 0. And we want to predict and you are also given another snapshot G t i V V E dash, look at the difference t 0 and t i t i is greater than t 0 meaning this is the later snapshot this is the earlier snapshot.

Set of vertices is same V and V, but set of edges they are different ok, this is E this is E dash. So, what is the task? The task is to predict E dash minus E edges ok. So, you are given this one you want to predict E dash minus E edges which are going to be generated right to create this snapshot G t i this is the task ok. So, alternatively the problem of link prediction can also be coined as the task of determining the likelihood.

That two nodes that are not connected at t 0 are going to be connected at t i essentially we are doing this thing. We are given a pair of nodes which are not connected at t 0 your task could be to predict whether these two whether this pair of nodes will be connected at t i ok. Now think of the nature of this problem right, say there are let us take a small network there are say 100 nodes right.

How many pairs of nodes are there are 100 c 2 pairs right and generally order of mod V is order of mod E in a large network. So, you can expect that approximately 100 edges are there in the network, so out of these many edges only 100 edges are present right at time t 0. So, the remaining node pairs are not connected at time t zero, but some of them will be connected at t i. So, from these many pairs you need to predict how many of them will be formed.

How many of them will be connected say only 1 percent or 0.5 percent will be connected in the future right. Now think of the sampling think of the negative sample and positive sample you have these many number of positive samples and you have these many number of negative samples, a huge imbalance problem right and out of so many negative samples right, so many disconnected pairs of nodes right. How many of them will be connected, in order to predict this? This is essentially you know finding needles in the haystack, is not it?

So, you have a huge population and from that your task is to identify small things which is really difficult ok. We will discuss about it later. So, in the next lecture let us stop here in the next lecture we will discuss the evaluation process right before moving into the algorithmic part, we will discuss the evaluation process and we will see the we will see you know the major challenges in link prediction you know problem in general ok.

Thank you.