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

Chapter - 08
Lecture - 01

So, today we will start a new chapter called anomaly detection in graphs. So, this is the 8th chapter of this course social network analysis and so far we have learned almost you know the fundamental all the fundamental concepts that are needed to understand this chapter. So, you know the prerequisite for this chapter is hopefully you have already learned the link analysis part, the link prediction part, the community detection, the clustering you know the clustering algorithm of a network.

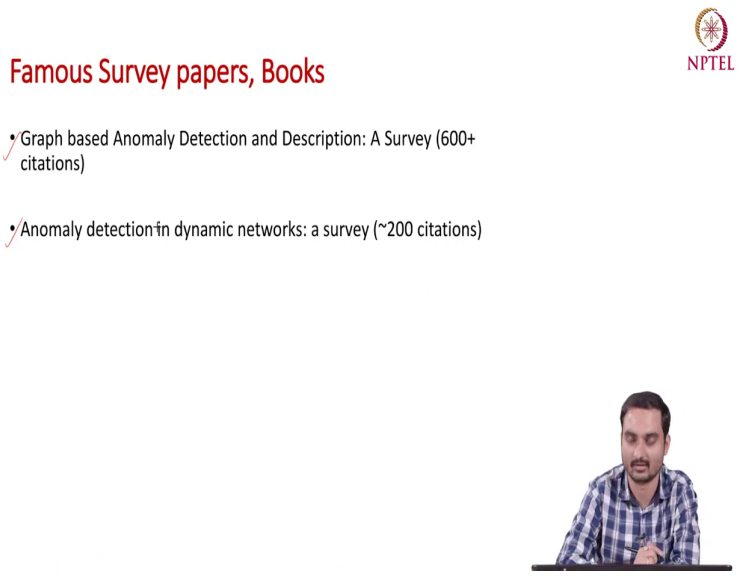
And you know a bit of you know centralities and all these stuffs. So, in the last chapter we have discussed one important application which was you know how information spreads on social network, how cascade grows on social network and what are the factors you know which lead to a growth of a cascade on online social network and that application was important in the concept of I mean in the context of say vitality prediction right.

So, you want to predict the virality of a of an of a particular information of a particular tweet. For example, so what would be you know what would be the approaches based on which you can predict the virality or a cascade growth right. Now you know sometimes when we talk about virality or you know information spread we also look at who is spreading the information right. Which user is responsible for spreading say a malicious information. Say fake news or misinformation or you know hate speech right.

And we look at we try to identify bad actors right in the context of social network and try to reason about you know how such bad actors you know start spreading information and how this spread is different from the normal spread right. So, in this chapter we look at how we identify nodes or users which are you know abnormal which are different from the rest of the network right with respect to some application.

This application can be fraud detection application can be you know say credit card transaction fraud transaction detection and so on and so forth ok alright.

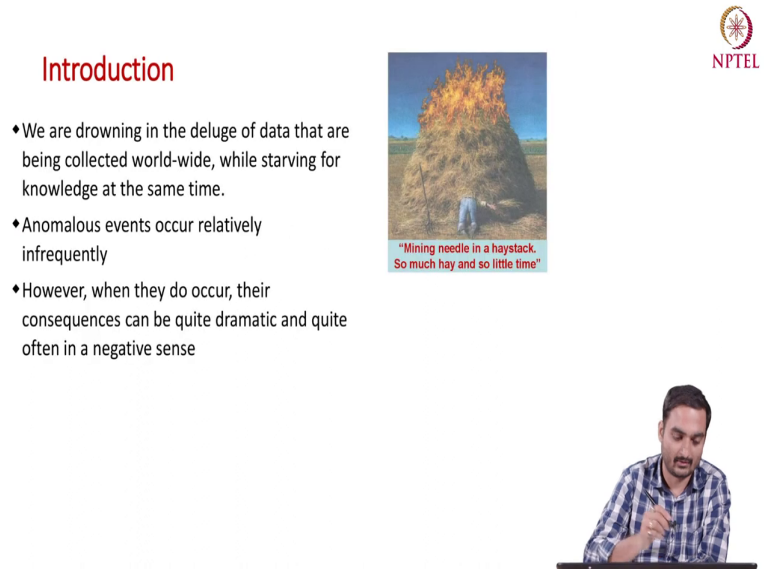
(Refer Slide Time: 02:55)



So, to begin with these are the two survey papers which we are going to follow and these are heavily cited survey papers on graph anomaly detection. So, you can you may want to look at you know this survey papers appropriately ok. And the slide the slides the you know today's slide is based on these two survey papers.

So, if will try to give you a brief of you know the methods that we generally follow for anomaly detection, but if you really want to go deeper into the methods each and every method you should look at the survey papers. I will try to cover you know some important algorithms which are generally used for graph anomaly detection. But of course, those two those few algorithms are not enough. You can look at these survey papers ok alright.

(Refer Slide Time: 03:45)



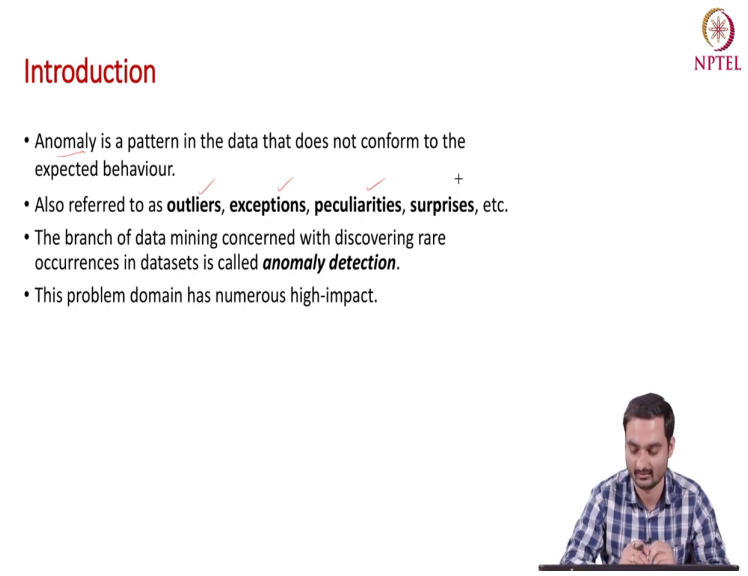
So, if you look at the anomaly detection problem in general right. It is basically about identifying rare events ok infrequent events right and when we talk about infrequent or rare events we if you look at the entire population like for example, if you look at Twitter or Facebook in general millions of users billions of relations right.

And from the from billions of relations or millions of you know nodes or users it would be extremely difficult to spot anomalous nodes or outlier activities right and if you look at the population its basically say 1 percent of the entire population. Anomalous behavior is 1 percent even less than 1 percent of the entire population. So, it is extremely difficult to identify to spot anomalous behavior in the context of social network.

And not in the context of social network, but in general if you look at any outlet detection you know applications right. Generally the proportion of outlier data points is much much lesser than you know the genuine data points. So, it is essentially you know mining needle in a haystack right. So, essentially you have huge population and from this massive population your task would be to spot bad users bad actors right inorganic activities right, fraud activities right in a systematic manner.

And there are multiple challenges associated with anomaly detection that we will discuss in the later part of the slides ok.

(Refer Slide Time: 05:26)



So, what is anomaly? Anomaly is a pattern in the data that does not confirm to the expected behavior. Now this is a very vague definition. In fact, we were we will you know talk about many such fake definitions right which may not make much sense without identifying the context right.

So, first we identify the context the application and then we define what do we mean by anomaly anomalous behavior or anomalous entities with respect to that context ok. So, this anomaly the term anomaly can also be you know can also be referred to as outliers or exceptions or peculiarities or you know surprises and etcetera. Remember when we talk about anomaly detection it does not always refer to bad activities or fraud activities or in genuine or inorganic activities.

Anomalous behavior can also be exceptionally good behavior ok. For example, if you think of a citation network or researchers network right or scientific articles and their relations right through citations. Anomalous nodes are those papers which are which have been exceptionally highly cited right. Anomalous researchers are those users those researchers who have heavily been cited. Who have received a lot of awards.

Who are say Nobel laureates for example, because their behavior is very different from the normal behavior right if you look at the population in general. So, Nobel laureates research activities or their publications their innovations are very different from the normal behavior. So, it is not the case that anomalous nodes are always bad actors right. Anomalous nodes or

anomalous users are those which are which behave very differently from the rest of the population ok. This is very important to note ok.

So, it has so the problem of anomaly detection has numerous high impact for example. For example, in case of say fraud detection right or in case of say fake news detection right, even if the amount of fake news that are being spread every day compared to the real news is much much lesser, but the impact that this fake news creates right or makes on the society in general is huge right.

We have seen cases like communal riots right, violence even death cases because of the fake news. So, particularly in the context of Covid-19 where this misinformation about health related you know tips right has been has spread and you know people basically adapted those advice and you know that again has led to even death right.

(Refer Slide Time: 08:39)

So, therefore, this is very important and our task could be identify anomalous behavior from social network. We will particularly look at social network. But, the techniques that we will discuss might also be useful for other kind of networks as well ok. So, now let us look at some of the real world applications right, credit card frauds right.

If you look at the abnormally high you know purchase made through a credit card right. You can say that this is kind of anomalous behavior. Similarly in the context of cyber intrusion if you look at some sort of involved web server involved in say ftp traffic. This can also be a

potential threat right of course, fake followers, fake retweeters, fake reviews, these all come under the broad umbrella of anomaly detection ok.

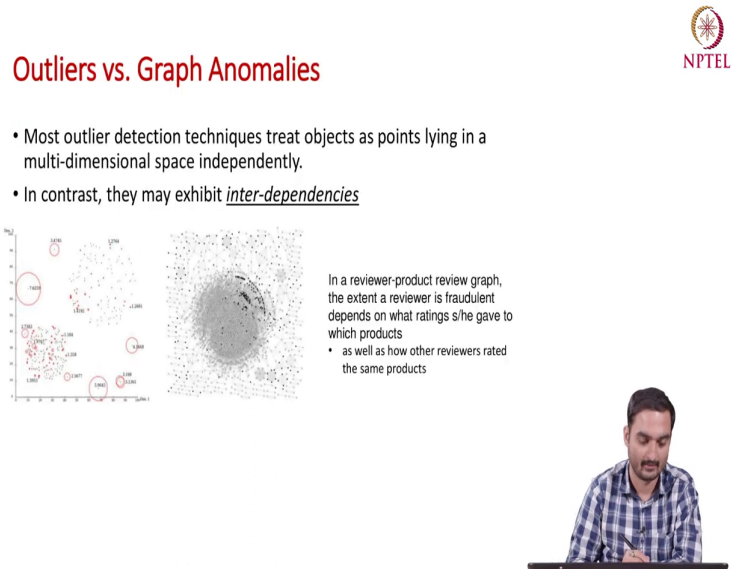
(Refer Slide Time: 09:37)

Apart from this we have a series of applications. For example, you know calling card and telecommunication fraud auto insurance fraud Email and web spam. This spam detection methods right opinion deception and review spam. If you look at the say E-commerce services like Flipkart Amazon, there the kind of reviews that people generally write.

Many of them are frauds right fraud reviews even people do not buy products, but they write reviews right. Auction fraud, tax evasion, you know customer activity monitoring and user profiling fraud clicks right. Clickbait is something which is again is under the broad umbrella of outlier detection.

Then malware and spyware detection, false advertisement and image video surveillance, all these applications come under you know under the broad umbrella of anomaly detection or outlier detection ok alright.

(Refer Slide Time: 10:41)



So, now, let us look at the you know difference between outlet detection and graph anomalies. So, outlet detection is generally studied in data mining and you must have heard about techniques where we right given an embedding given a you know embedded space or any Euclidean space.

Data points are mapped and you try to identify those data points which are say farther apart from the other data points right. This is basically out layer detection in the context of data mining right, but here we will mostly talk about graph centric anomaly detection ok. Now why graph centric detection is useful? Why graph is useful in the case of anomaly detection? So, the reason is as follows let us take an example of say reviewer fraud detection right.

You have a product and you look at the reviews of the product and ratings of the product are given by different users and your task is to identify those reviews which are fraud right or fake right. Now if you look at reviews individually right individually in the sense like it is not the case that you look at one review and you analyze the second review with respect to the first review. You are looking at reviews independently ok as if they are sampled from an IID right.

So, if you look at reviews right a fraud review in general looks like a genuine review right. For example, it may you know it may talk about the some problems in the product right. Delivery related or features related problems right and therefore, automatically the rating is

low right one or say two out of five. Now how do you know that this review is actually a fraud review ok without looking at other reviews right.

Similarly, if you think of another setting where you look at all the reviews and try to understand their relations right which review is similar to which are the reviews right and you also look at the patterns when these reviews have been posted right their rating behavior right and so on and so forth. You might find some cases where you know reviews that the text the you know the semantics that takes the content of the reviews is almost same right.

They are posted at the same time and the corresponding ratings are also same. Their content is same, their rating is also almost same, they are also posted around the same time. Then you may wonder that how I mean how has it happened right because how could you know users write similar kind of reviews very similar kind of reviews with same ratings at the same time right.

So, you are looking at the interdependencies between reviews right with respect to the content with respect to the rating with respect to the time. So, there is a synchronicity in terms of the content rating and time and maybe some other features as well right and then you may wonder that possibly this post this set of reviews have been posted by the same set of users.

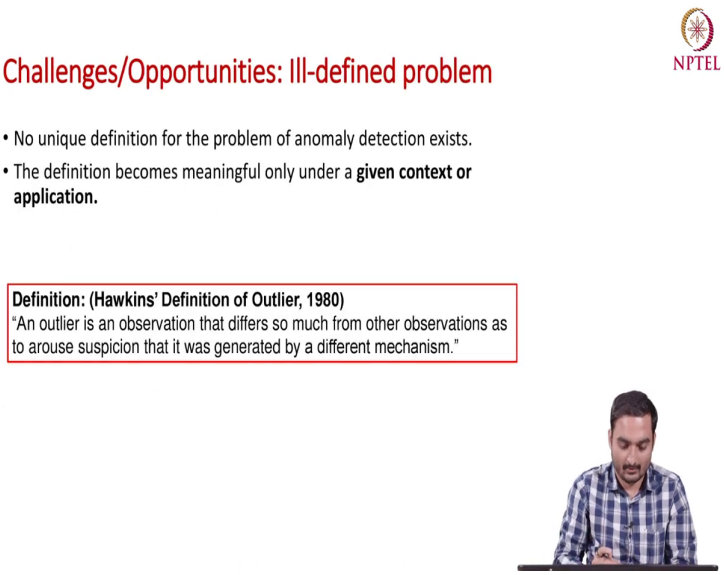
There might be a collusion between reviewers right between users they might have been paid to write fraud reviews about the product right. Again remember when we talk about anomaly detection right, it can also be genuine reviews right. In the sense like not a genuine reviews, but very high quality reviews on a particular product, but there might also be some collusion right behind this kind of. For example, you know a product the seller of the product the seller who wants to boost the rating of the product right.

He may hire, he or he may hire a set of reviewers right a set of users who would start writing you know good quality reviews about the product with heavy ratings right. Those are also anomalous ok. So, why graph is important? Because graph structure would give you the ideas about the dependencies of reviews right, the dependencies of users the dependencies of ratings right in some ways.

This interrelations has can easily be captured using graphs right in contrast to if you just look at the embedding space and data points in isolation you will not be able to identify the

dependencies. So, therefore, graph plays a very important role in modeling the anomaly anomalous behavior detection task right ok.

(Refer Slide Time: 15:53)



So, now, let us look at the challenges and opportunities right. The major challenge or the major opportunity is lies in terms of the definition itself right. I mentioned in the chapter of community detection that that the community detection problem itself is an ill defined problem.

It is not an well defined problem because what is a community is not very well defined it is not very clear right. I can define in my own way you can also define in your in your own way and that is challenging and that is also opportunity because you know you can come up with your own definition and if you convince others that this definition makes sense, you can you know write papers ok.

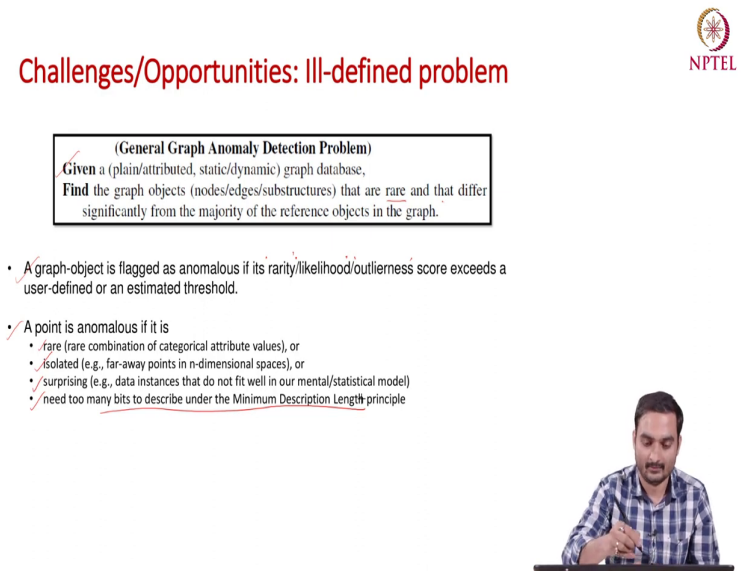
You can design methods which can capture communities of your own type. Here also anomaly detection is an ill defined problem. There is no unique definition of what do I mean by anomaly detection right in the context of a particular application right. So, what we generally do?

We take an application and we define you know meaningful definition of an anomaly and then we design methods to identify that kind of anomalous behavior. So, now let us look at the you know this Hawkins definition of outlier which was proposed in 1980. It basically says

that an outlier is an observation that differs so much from the other observations as to you know arose suspicion that it was you know generated by a different mechanism.

You know the definition is little complicated, but do not worry I mean we will try to make it simple ok.

(Refer Slide Time: 17:45)



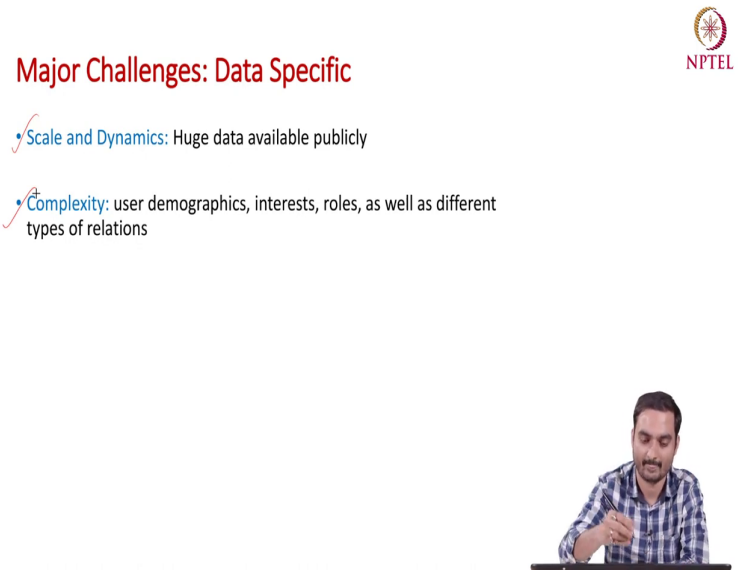
So, let us define the task of graph anomaly detection problem as follows. So, you are given a graph, the graph can be a plain graph or an attributed graph. It can be a static graph or a dynamic graph and what is the task? The task is to find graph objects nodes edges or sub graphs that are rare and that differs significantly from the majority of the reference objects in the graph ok.

So, a graph object is flagged as anomalous. If its rarity likelihood outlierness score exceeds a user defined or an estimated threshold ok. So, so we will define what do we mean by reality, what do we mean by likelihood, what do we mean by outlierness right. Again with respect to an application ok.

So, point is anomalous if it is rare right a rare combination of categorical attribute values right. It is isolated, it is far apart points right and they are far apart points in a n dimensional space and there is some sort of surprising factor right and of course, a data point needs too many bits to describe under a minimum description length principle.

Now this is something that we will avoid in this particular chapter because it requires lot of understanding of information theory which we will not you know discuss, but one can also look at entropy information gain and so on and so forth to encode data points using bits and then we can you know detect anomalous behavior or outlier behavior ok right.

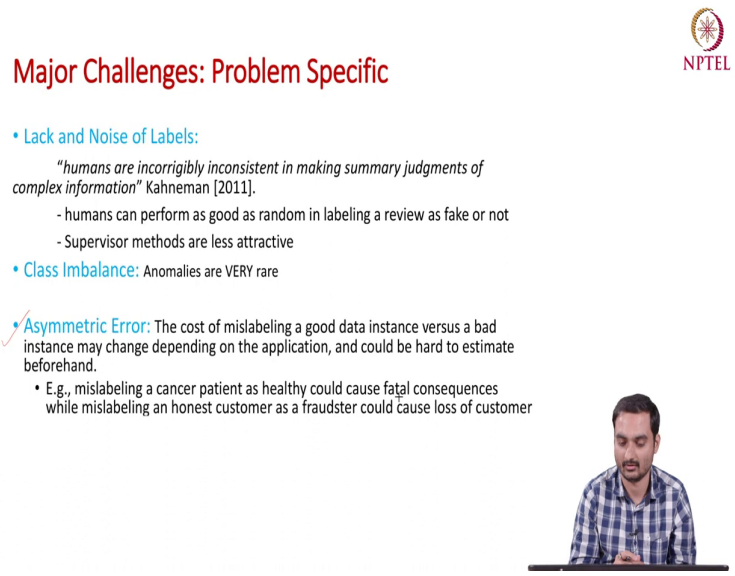
(Refer Slide Time: 20:03)



So, now let us look at some of the major challenges. So, the first challenge lies in the data specific aspect of the problem. So, the that the scale or the dynamics of the data point itself is a problem. We have huge data point right which are publicly available. Now from this huge data point how can we identify data points which are very rare right.

This is a problem and the major and the other problem is the complexity. If you look at the user demographics interest role as well as different relations right, we need to consider all these aspects together right. So, the complexity increases as we incorporate more and more features more and more attributes in the outlet detection problem right ok.

(Refer Slide Time: 21:01)



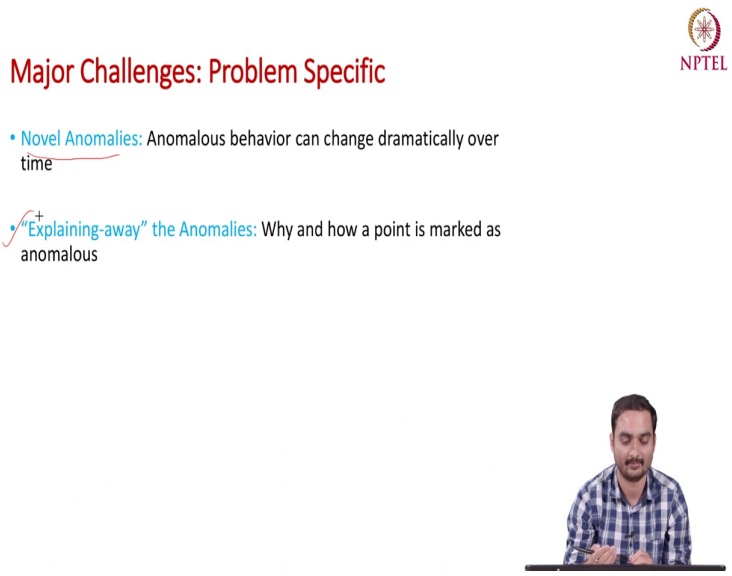
Now, let us look at the other challenges. The second challenge lies in the problem itself right. If you look at the problem the outlet duration problem it is basically an classification problem a classification problem, but it is and you know it has huge class imbalance. You have 0.1 percent or 1 percent population which is outlier and remaining 99.5 percent 99 percentage population is non outlier right. This is a problem. So, now, how do we tackle it?

We will discuss I mean we can use say class imbalance problem specific task right. If you we take say classifier which handles the class imbalance problems in an in a intrinsic way and that may solve the problem, but when it comes to graphs right those methods may not be may not be enough right. The second aspect is the asymmetric error. What do you mean by this? What is asymmetric error ok?

So, if you look at the cost right when we you know identify or when we manually identify anomalous behavior right. If we wrongly identify an anomalous node right the cost may be detrimental. For example, if we misclassify a cancer patient as a healthy patient right.

The effect can be a fatal consequence right the it can have a fatal consequence. If we misclassify an honest customer as a fraudulent customer you can lose out you may lose out you know the customers loyalty or customers trust on the platform itself right. So, there is no luxury of annotated data points in a wrong manner right in a in an erroneous manner which may lead to some you know some detrimental consequence right ok.

(Refer Slide Time: 23:24)



So, the other problem specific challenges include the anomaly detection problem the anomalous behavior itself is novel why? Because this fraudulent customers or fraudulent users they keep on changing their behaviors over time. So, let us say today if you come up with a method that can detect fraud activities right.

That method may not be enough tomorrow to detect fraud activities because meanwhile it may happen that the fraud activity has completely changed. The way people behave the way fraudulent users behave that may change drastically. In fact, our lab you know. So, we have been working on this fraud detection problems since last 3 4 years.

And then and one of the questions that people generally ask is that lets say a fraudulent users will read our paper and we will understand how we detect fraud behavior and they change their behavior tomorrow right. How would your method which has been designed yesterday that method would be able to identify you know the evolved behavior right and that is [Laughter] very difficult to answer.

Because I mean I generally answer in the following way. I can say that look you know remember fraudulent users they do not have enough resources right, they do not have enough say computing facilities or enough you know enough scientific knowledge to understand you know the behavior well. So, meanwhile when they change their behavior since we keep on collecting data points online right.

Meanwhile it may happen that we will also adopt you know we will also modify our algorithm and try to capture their behavior their evolve evolved behavior right ok. So, the second the second challenge is the explainability. So, when we pinpoint when we say that look this user or this behavior is a fraudulent behavior you may be challenged right. So, you have to explain why you are saying that this guy is a fraud guy.

Why this activity this activity is a fraud activity right. So, your model should be able to explain right the result in an appropriate manner. So, that people will understand you know the reason behind the class label that you predict ok. So, these are the major challenges these are the major opportunities I would say and therefore, this problem is extremely interesting in the context of data mining as well as social network analysis.

So, hopefully you I mean I was able to motivate you why outlet detection is an important problem, anomaly detection is an important problem. So, we stop here. In the next lecture we will start discussing you know about important algorithms we try to come up with the taxonomy of anomaly algorithms and then we start looking at important algorithms both in terms of like you know the way people approach this problem and how you can use modern techniques to design your own methods of you know detecting outlier outliers from a graph ok.

Thank you.