

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

Chapter - 05
Lecture - 01

Hello everyone welcome back. Today we will discuss we will start a new chapter on, something called Community Analysis in Networks. This is perhaps the longest chapter of this course, but this is very important. The fundamental property that we are going to discuss today. So, far we have learned you know different metrics to quantify a network, we have learned how to model a network, right.

We have also learned how to quantify links in different ways, right. And, we have also learned you know some of the properties that are often you know seen in real world networks. For example, scale free property, right, power law degree distribution, small world, high clustering coefficient and so on and so forth. So, similarly community structure is one of such properties which is quite prevalent across different real networks.

Be it social network or information network, citation network, right, biological network and so on, ok. So, let us get started.

(Refer Slide Time: 01:33)

'Homophily' in the Society



- Tendency of individuals to associate and bond with similar others
- Similar nodes tend to attract each other, and dissimilar nodes tend to get away from each other
- Causes formation of a **community structure** in a social network
- Homophily occurs against a number of categories:
 - Age
 - Sex and Gender
 - Class: Education, occupation, and Social
 - Religion, Race, and Ethnicity
 - Interests
 - Organizational role, etc.



So, let me ask a you know fundamental question that, you know how many times your decision on you know watching a web series you know got influenced by your friends? I am pretty sure most of you generally take suggestions from your friends to you know to watch a particular web series or to watch a particular movie, right.

Also, I am pretty sure that most of the times you when you purchase a product say from Amazon or from Flipkart for example, you often talk to your friends to understand the quality of the product, you look at the you know reviews and so on. So, you know in our real life our decision of you know taking something or watching something or going somewhere else or going to a particular you know restaurant and so on and so forth, all these things are dependent on our neighbors decision, right.

We generally talk to our neighbors, we generally talk to our friends, communities and then we take decision, right. And, we you may also see that you know when an unknown person say comes to a; comes to a society, right, he or she always you know starts looking at communities where he or she can best fit, right. For example, say I am a Bengali, right. And, when I go to when I you know go to any party or any society, I always start looking at you know friends who are also Bengali.

You know, so there is always a homophily. And this homophily property is something which attaches you know nodes having similar properties, right. You, I mean hopefully you remember we discussed something called assortativity, right, assortative mixing and so on in the 2nd chapter on network measure. And, you can also measure homophily based on the assortativity of a network.

So, the term homophily is very important which basically says that, nodes with similar properties they tend to interact with each other quite frequently. And nodes with dissimilar properties they do not interact that frequently, right. And because of the frequent interactions among nodes with similar properties, we often see that similar nodes often times come together and form a close community or close group, ok.

And this property the dense groups or the close groups you know in networks, these properties quite prevalent across different types of networks, right. For example, if you look at say Facebook, in Facebook we see that we can also create communities, we can create groups, right. So, remember I use the term community, group you write module the these terms quite interchangeably, ok.

So, in Facebook we often see that there are you know there are existing groups or you can also create your own group. Say for example, you create Sachin Tendulkar fan club, right or say Lata Mangeshkar fan club and anybody you know who is a fan of Sachin Tendulkar or Lata Mangeshkar so, they will basically join the group and then you start discussing you know about their past activities, the songs etcetera, right.


And you know. So, we so, this group basically defines certain attributes and if you also have the same kind of attributes, you basically join that group and be a part of the be an active part of that group, ok. So, these kind of communities can be formed naturally or artificially. For example, in case of Facebook you know you proactively create groups and then people start joining in.

Whereas if you think of say you know user-user interaction network or a friendship network, where you see that you know groups like a music club or a gym club or you know a department, right. These groups are basically created naturally. Say, if you are; if you are fond of a music, you go and join a music club, right.

So, homophily is the property for which we see community structure emerging on different social network, different social as well as other types of networks. So, we generally quantify homophily based on the features, right, it can be age, it can be gender, class, religion, race, ethnicity, interest, organization, organization role and so on and so forth, ok.

(Refer Slide Time: 06:16)

Communities in a Network



- Identifying communities gives an insight about the inherent network structure
- Community detection is an *ill-defined* problem
 - What we mean by a 'community' is often not concrete
 - Often hard to reliably define a ground-truth annotation for communities
 - No standard measure to assess the performance
- Diverse approaches to the problem depending on how we define a community structure in the network

<http://bit.ly/3U7a692>



So, what is community? What is cluster? So, as the name suggest you also understand that you know when I talk about cluster, it means that you know there are groups which are densely connected internally, right. So, community the definition of community is not sacrosanct. There is no concrete definition of what is a community, ok. It has a fuzzy definition and therefore, community analysis or community detection is an ill-defined problem it is not an well-defined problem, ok.

So, you can define community in your own way you know, anybody else can define community in his own ways and then you know you can start detecting communities in that manner. So, what is a community? Generally, a community is defined by groups of nodes such that within a group nodes are densely connected and across groups nodes are sparsely connected, not densely connected, ok.

So, now if I you know define community in this way, then the natural question would be you know how do we, what do we mean by densely connected? What do we mean by sparsely connected? Can we quantify that? Ok, after certain point. So how do we quantify density? We have already discussed a matric called edge density, which basically measures the fraction of edges, right. Its basically a fraction of actual number of edges and the total number of edges.


I mean the possible number of edges present in a particular group, right. So, once you measure edge density, then how do we quantify you know the meaning of highly dense or densely connected or sparsely connected. Is there any threshold, after you know after which we will say that, ok this is densely connected and below which we say that this is sparsely connected; there is no such definition.

Therefore, you know this is not an well-defined problem as such, ok. So, it is good as well as it is bad. Why it is good? It is good because you know you can define your own way you know you can define you can come up with your own definition of communities and you can write papers, ok. And, it is bad because since there is no standard definition, therefore it would also be difficult to convince others that the definition that I am proposing actually makes a sense, right.

So, in that sense this is difficult, ok. So, then how do people essentially deal with this problem? So, what people do is, they try to come up with the definition of a community and of course, it needs to be convincing enough and then you detect communities from the network in some ways, ok.


So, in the subsequent slides, we will define communities in different ways and we will see how you know based on those different definitions we basically extract communities from a large network, ok.

(Refer Slide Time: 09:12)



Community Detection in Networks: Applications

- ✓ Performance enhancement of the similarity-based [link prediction](#) algorithms
- Improving recommendation quality in [Recommender systems](#) by separating like-minded people
- ✓ Controlling [information diffusion](#) within a network by identifying community memberships
- ✓ Designing better [marketing strategy](#) by identifying position of the target group within the network
- ✓ Restricting [epidemic propagation](#) by suitably isolating and immunizing the vulnerable population
- ✓ Better [anomaly detection](#) in nodes, especially in evolving networks
- ✓ Studying [evolution of communities](#)
- Applications in [criminology and detecting terrorist groups](#)



So, what are the applications of community detection? It is useful for link prediction. So, in the next chapter we will discuss what is called link prediction. So, basically the idea of link prediction is that you know let us say you have 2 nodes which are not connected at timestamp t , what is the probability that at time t plus 1 these 2 nodes will be connected? Can we forecast; can we predict the likelihood the 2 nodes which are apparently not connected at current point in time we will be connected at next point.

Or say t plus 1 or t plus delta t time period this is called link prediction. And how community structure plays an important role here? Say let us say let us say let me take an example, right.

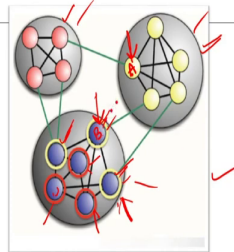
(Refer Slide Time: 10:00)

Community Structure



Theoretical reasons

- ✓ Organization
- ✓ Node features
- Node classification ✓
- Missing links ✓



Let us say this is the network, ok, this is the network and let us say you already know that this is the community, this is another community and this is another community; there are 3 communities, ok. And the question is can we predict that this node say node a and node b these 2 nodes. So, they are not connected at current time period, ok.

What is the probability or what is the likelihood that a and b will be connected later? Ok. So, if you know the community structure, you can say that look a and b, right, they are part of the same community therefore, it means that a and b possess exhibit same kind of properties or same kind of homo I mean same kind of structural functional properties, right.

Therefore, they are part of same community. So, this is highly likely that in the future a and b will also be connected. So, if you know the community structure and if you know that these nodes are part of a community and if those nodes are not connected at this current time period, it is highly likely that at $t + \Delta t$ time period they will be connected, ok.

So, community structure plays an important role for link prediction. We will discuss more in the next chapter. And, of course, link production you know one of the applications of link prediction is recommendation system, right. For example, product recommendation system or movie recommendation system, right.

It is also a link prediction task. So, of course, then community will also help in you know in recommending products to certain users or movies to certain users, ok. The third application

is information diffusion, ok. We will discuss in one of the chapters later, that if a information starts spreading from a particular node and if I know that this node belongs to a particular community. So, it is highly likely that the information will first spreads within the community.

So, the nodes which are part of the same community, they will be affected first they will be exposed to that information first and then that information will move from that community to the other community, ok. So, if you know that within a community these nodes are part of it is highly likely that those nodes will be affected you know by the information first compared to those nodes which are not part of this community.

It is useful for you know marketing strategies. For example, you know we discuss something called targeted advertisement, right. What is targeted advertisement? Let us say you have build a product for youth, right or you have built a product for a certain region of India, ok. And you want that part of; that part of the network will get exposed to that particular product, ok.

So, if you know the community structure, what would happen is that you target that community, where your potential customers belong to and you basically start probing them, ok. Start convincing them to buy your product, ok. So, it is useful for targeted marketing, ok. This is also useful for restricting the epidemic spread.

If you remember we discussed in one chapter in one of the previous chapters is that you know there are there are nodes which are called go gatekeepers, ok. And, gate keepers are generally those nodes which you know connects which basically connects 2 communities in a network 2 or multiple communities in a network.

So, if we know the communities it would be easier for us to understand to identify the gatekeepers or articulation points. And then, accordingly we will vaccinate. We will vaccinate those users so that, the epidemic will not spread from one community to the another community. This is also useful for anomaly detection. Again, in one of the chapters later we will discuss, what is anomaly?

So, anomalous nodes are those nodes which do not look like normal nodes in a network, ok. And again, when you define anomalous nodes, we generally define with respect to a particular community, ok. Let us say there is a community of students where all the students

are or the most of the students are Ph.D. students, right. But there is one student who is a master student, right. But, all of them belong to the same community.

So, all of them say belong to the same research group, right. So, with respect to that community the master student is an anonymous node, ok, but with respect to the overall network there are many other master students. So, overall that node is not an anomalous node not an outlier node, but with respect to that community its an outlier node. So, if we know the community structure, you can detect anomalous entities in a network in a better way.

You can also study the evolution of communities. what do you mean by evaluation of communities? Let us think of you know a scientific network, say citation network, ok. And, in the citation network generally communities are different research areas. Say, in computer science citation network you can think of AI, database, right, algorithm and theory, right say robotics. These are different communities, ok.

And then, you look at how you know over the times how does this network changes? And at the same time, how this community structures change? Meaning, that how new nodes are coming in since citation network is a growing network, it will never you know it will never the number of that the size of the network will never reduce, its a growing network.

So, we basically look at how nodes are coming in within communities and how interactions are happening across communities and within communities, right. And depending on that you can also decide which fields are going to be merged together to form a new interdisciplinary field in the future, right.



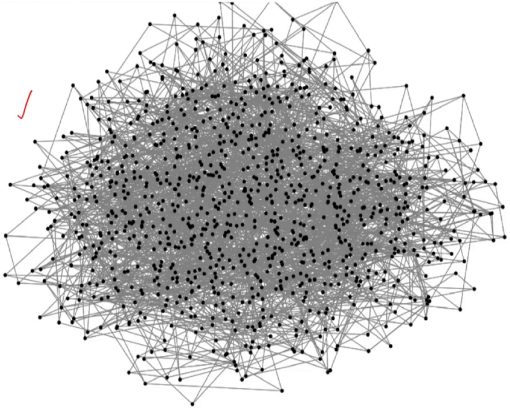
In fact, there are studies we also did a study where we showed that how you know computer science if you track the citation network of computer science how it fields like NLP or field fields like computational biology, these are all interdisciplinary fields, right. How these fields emerged you know from the cross fertilization of you know this communities like AI, database, theories and so on, ok.

Of course, this is useful for criminology to detect terrorist groups, because you know if we know that in a society you know generally these groups are legal groups or these groups are you know this group should be there in a network. You can identify those groups which seem

to be quite abnormal, anomalous and that can help you detect those fraud users or you know criminals from a network, ok.

(Refer Slide Time: 17:14)



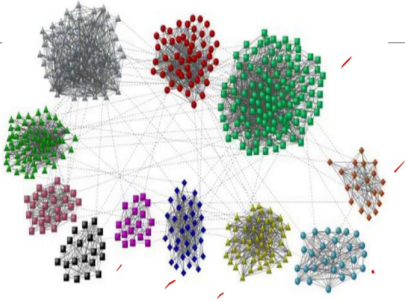
The Network



There are many applications. The community structure is something which is used almost every application in network science. So, let us now try to understand what is the task? What is the task that we are looking at here. So, we are given a network like this, ok. So, this is always a small network. But again, even if you look at this network manually, you would not be able to understand the clusters, right.

(Refer Slide Time: 17:41)

The Community Structure



And, what is the task? The task is to group nodes from this network, right, in such a way that you know these groups basically form different communities. So, community detection is nothing but, grouping nodes in a systematic way. So, that we see, right, this module module structure in the network properly, alright.

So, why do we care about you know detecting community structure? We have already discussed. But still, there are you know theoretical reasons there are practical reasons. So, let us look at the theoretical reasons. So, if we know the community structure, it would help you understand the you know organizational block organizational blocks of a system.

For example, if you think of user-user interactions in an institute, right. So, in an institute, you can think of these communities as different blocks as different organizational blocks. For example, one block can be an academic block, one block can correspond to one community can correspond to a sports block, one community can correspond to library block and so on.

Because, within a block, within a community users generally interact quite frequently compared to across blocks, ok. You can actually you know infer node features, if you do the community structure. How? Let us say, you know let us say in this particular network say let us take this community, right.

And, let us see let us assume that you know you know the node features, right, for example, the demography, the age, the gender and so on of this node, this node and this node all the all 3 yellow nodes, right. But you do not know the node features of these 3 red nodes, right. You can infer the node feature based on the known node features. Why? Because, since we know that the community is formed due to the homophily nature therefore, it is highly likely that this red nodes may also exhibit the same kind of property same kind of features that the yellow nodes have, ok.

You can also classify nodes because, essentially community detection is a clustering technique, right. Its a clustering problem. And clustering of a graph, clustering of a data set is nothing but an unsupervised come also weight classification task, right. So, you basically group nodes into different clusters. So, in that way you can classify nodes, ok. You can also predict missing links, ok or link prediction the one that I mentioned, right.

Say for example, you want to predict whether this node and these nodes are connected or not? What happens is that, when you scrape data points when you scrape data from social media, for example, Twitter or Facebook, what happens is that we sometimes may not be able to scrape the entire data, ok because of different reasons.

For example, there are problems in the crawling's process, there are problems in the in there are different restrictions, security restrictions. For example, some nodes are private, right, some edges you may not be able to look at, right. So, those information you may not be able to curate properly.

So, it always happens that when you scrape data sets some edges, some nodes will be missing, ok. So, if we want to make it a concrete data you need to predict those missing links, right. Remember, missing link and a future link, these are different problems. We will discuss in the next chapter. But missing links, basically you know missing link prediction you know is a task where at current time you want to understand or you want to detect whether link exist or not.

And future link prediction is basically says that, you know can we predict at time t whether link will come or a link will be formed at time t plus Δt . Its more of a forecasting, ok. So, let us say you want to predict whether node a and node b these 2 nodes are connected or not? Right. Is there a missing link between a and b ? ok.

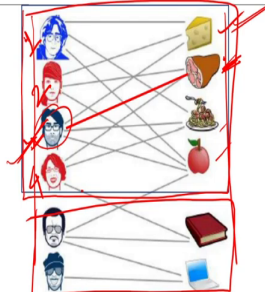
You can say that if you know the community structure, you can say that it is less likely that a or a and b are connected. Why? Because, a and b are part of different communities so, it is less likely that they will be connected or they are connected at the current time period, ok.

(Refer Slide Time: 22:17)

Community Detection



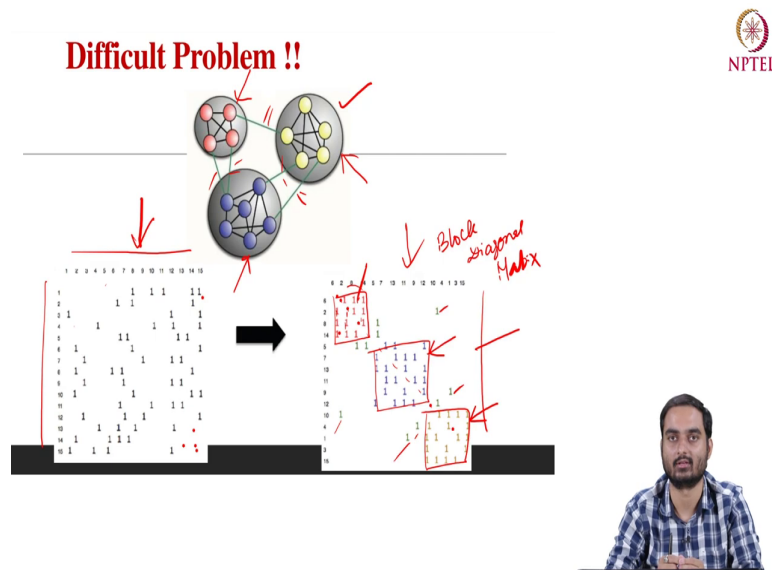
Practical Reasons: Recommendation Systems



Of course, community detection has practical applications. Recommendation system you can earn money, ok. For example, this product user product bipartite network. So, this is the bipartite network. In one partition you have users, another partition you have products and you basically want to predict whether you know whether this user will buy you know you also know that all these 4 users belong to a community and these 2 users belong to another community.

So, and in this particular community, the user 1, user 2 and user 4 all 3 users have already bought you know this bread, right and some other fruits and ham, right. It is also and say and this guy has not bought ham so far. Since this guy also belongs to the same community you would actually recommend ham, this product to this guy. Because, the other you know other users were part of the same community they have already bought ham. So, it is highly likely that this guy will also like to buy ham, ok.

(Refer Slide Time: 23:35)



So, why it is a difficult problem. Let us try to understand, why community detection is a difficult problem. So, if you think of it you know in a careful manner, what it is about? So, let us think of this network, right. There are 15 nodes you see here. So, you can create this adjacency matrix, 15 cross 15 adjacency matrix. And, you basically have ones corresponding to those pairs which are connected.

And this other you know places are 0s, ok. I have not added 0s because, if I mean if we just add 0 this would make this would be little clumsy, ok. So, this empty points, empty entries these are 0s, ok. This is sparse network. So, community detection is nothing but you know adjusting the adjacency matrix in a different manner so that you get these kind of structure, ok.

And, this is called block diagonal matrix this is called block diagonal matrix, ok. Why it is called so? If you look at the block so if you look at the diagonal element this is the diagonal elements. You see that there are blocks, right, blocks of 1s, ok. These blocks correspond to different communities. For example, this red block corresponds to this community.

The this one corresponds to this community and the yellow one correspond to this community, ok. And therefore, all these ones within a block, these are basically intra-cluster, intra-community edges, ok. And what about this green you know numbers, green ones? They correspond to intra-community inter-community edges, right. There are one 1, 2, 3, 4, 5

inter-community edges. So, you will basically see inter community edges in this modified adjacency matrix, ok.

So, its basic community detection is basically a switch to convert a raw adjacency matrix to a block diagonal adjacency matrix, ok. I mean one may argue that what is the problem here. I mean we can essentially you know swap rows and columns in a systematic way and we can easily get this kind of block diagonal structure, right. We do not need any algorithm.

But think of a large network, right. Million cross million adjacency matrix. How do we manually get this kind of block diagonal structure, right. So, we essentially need certain algorithm which basically convert an a raw agency matrix to this kind of required adjacency matrix, ok. And this is basically about community detection, ok.

So, we stop here. We in the next lecture we will continue with you know the type different types of the community structures and we will also discuss different algorithms that we generally use for you know community deduction.

Thank you.