Chapter - 05
Lecture - 06
Social Network Analysis

So, we have discussed what is modularity and how we can use modularity for community detection.

(Refer Slide Time: 00:33)

And this is this seminal paper that was published by Mark Newman in 2006. If you want to know more about modularity, you can have a look here ok.

(Refer Slide Time: 00:40)

Permanence and Community Detection

$$\frac{1}{3} \sum_{i=1}^{|c|} (e_{ii} - a_i^2)$$

$$P(v) +$$

- ❏ Modularity is a network-centric global metric
  - Considers the entire network structure during maximization process
  - Not suitable for large and evolving networks
- ❏ Requires a method that looks at the local neighborhood while detecting communities
- ❏ Chakraborty et al. proposed a metric, named Permanence, which is a local metric for community detection
- ❏ A vertex-centric metric
- ❏ Two communities A and B are neighbouring communities if $\exists u \in A, v \in B$, and there is an edge between u and v

So, the problem of modulating maximization is a is basically threefold right we have already discussed. The first one is resolution limit, the second one is the degeneracy of solution and the third one is you know asymptotic growth of the modularity value right. So, in order to deal with this problem along with some other problems people started defining other kind of you know metrics. So, we also defined way back 2014 we defined a metric called permanence ok.

I mentioned earlier that you know this is the beauty of this of this particular area that you can define you know your own way. I mean you can define this in your own way and you can quantify, but you also need to show that your method is better your metric makes more sense and so on. So, we define something called permanence is a metric like modularity, but it has certain advantages.

So, first problem of modularity is that modularity is a global metric is a graph centric metric meaning that given a graph and the entire community structure right you can measure modularity. If you remember the curve if you remember the formula of modularity we basically said that let us take all the communities i from i equals to 1 to mod c and then we have two quantities one is intra community edges which is denoted by e ii right.

So, the quantity was this right i equals to 1 to total number of communities, then e ii which is the fraction of intra community edges minus a i square which is basically a fraction of two things. One is the total number of edges 2 into m and then we also had a that sum of degrees of all the nodes in the community. So, given a community structure we can measure this. A

permanence on the other hand is basically a local metric ok. You can define permanence for a particular vertex, but you cannot define modularity for a particular vertex, you cannot do this ok.

So, permanence can be defined for a particular vertex, for a particular node right and the other good part and since this is a local metric. The other good part about local metric is that it can be used when a network changes over time right. Let us say you have a network you measure the community structure you detect the community structure using a metric some metric and then let us assume that some new nodes are coming in you do not need to you know you need to you do not need to detect the community structure again from the scratch, what you can do?

You take the old community structure and then you see where these nodes new nodes are getting added and since your metric is a local metric you try to now you try to optimize with respect to those portions of the graph which have been changed due to the addition of new nodes are edges. You do not need to optimize your metric on the entire network again right. So, that would also help you for detecting communities in the evolving graph.

(Refer Slide Time: 04:06)

Chakraborty et al., Nature Scientific Reports, 2013
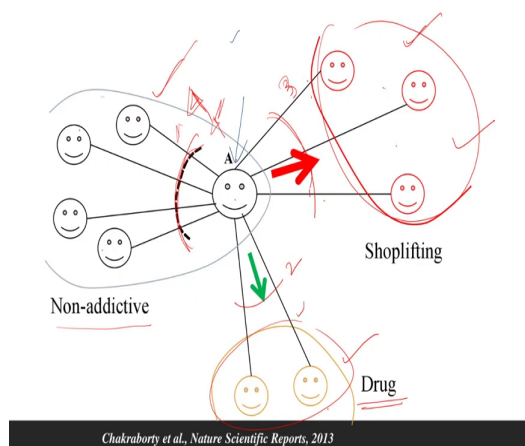
So, let us look at the formulation of modularity permanence ok. So, I will basically discuss the formulation using two stylized you know cartoon example. So, this is the first example ok. So, permanence basically is built on two heuristics. So, the first heuristics is as follows.

So, let us assume that and remember permanence is again a metric which was defined to quantify the quality of a community structure ok.

We assume that we already know the community structure and we see a whether the structure is better or not right, the quality is better or not. And later of course, we will use permanence for community detection the way we did for modularity ok. So, let us assume that you have this kind of network and you are A ok and you belong to a community where there are two there are 4 other nodes and this community is a non addictive community right. Your friends are not addicted to any bad activities or whatever right.

So, but you also have friends not in the same community, but in other community who are addicted to shoplifting and there are two other friends who are addicted to drug right, but you belong to a community which is non addictive and you have 4 non addictive friends ok. Since you have so many friends, so your external friends meaning your shop you your friend from this community and your friend from this community they are insisting me to join their communities ok.

So, in other words you are kind of experiencing a pool an external pool from this community and from this community separately ok. You are also experiencing a pool from your own community internal community. Let us assume that this pool is proportional to the number of a neighbors right. So, the pool is proportional to 3 here and the pool is proportional to 2 here, but here the pool is proportional to 4.

So, as long as you have more internal pool compared to the other external pools you are safe. You will not move to other communities you will remains remain in your own community. So, here you see that your internal pool is 4 proportional to 4 your external pool your maximum external pool is 3. So, 4 is greater than 3. So, you will not move to shoplifting community you will remain in your own community.

(Refer Slide Time: 06:50)

Total internal connections > **maximum external connections** to any one of the external communities

X Modularity, Conductance, Cut ratio
consider **total external connections**

So, the first heuristics is that your total internal connection should be greater than the maximum external connection that you are that you have with your neighbors with your you know neighbor neighborhood communities that is the first heuristics. We will capture this heuristics using some quantity ok.

(Refer Slide Time: 07:12)
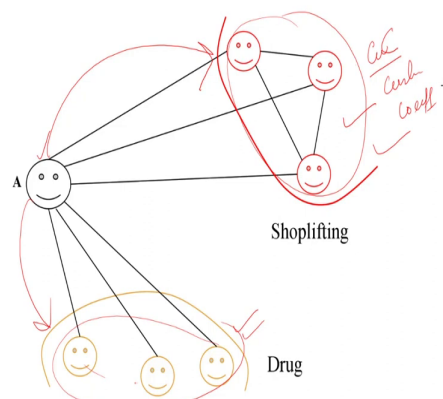
Chakraborty et al., Nature Scientific Reports, 2013

Let us say the say let us see the second heuristics. Here assume that you have 3 friends in the shoplifting group and 3 friends in the drug group, but you do not belong to any community again they are insisting me to join their community, but you are not interested. What they

would do? They would internally discuss which each other and probably try to come up with some interesting ideas to convince you to join their community.

So, if you look at the shoplifting groups nodes are highly connected meaning you know meaning they have high understanding whereas, in drug groups nodes are not connected at all meaning they do not have high understanding. So, the probability or the likelihood that shoplifting group will discuss among each other and try to come up with an interesting solution interesting proposal that may convince you to join this community is higher than the likelihood of joining you to this community right.

So, what we are trying to quantify? We are trying to quantify how my neighbors are connected right? How my neighbors are connected? Here my neighbors are connected strongly connected, but here my neighbors are not connected. So, how do we quantify this? We quantify this using something called clustering coefficient we have discussed earlier, clustering coefficient right. So, we have two heuristics.

(Refer Slide Time: 08:40)

Now, let us see so now what is the heuristics here the heuristics is that my internal neighbors should be highly connected. If my neighbors are highly connected I will not move to the other community. I will remain in my own community because my neighbors are connected. So, they have high understanding and I am also connected to my neighbors. So, therefore, I will be safe ok.

How we capture this? We capture this internal I mean the connective to connections within our internal neighbors using clustering coefficient. So, higher the clustering coefficient higher the chance that you will remain in the same community ok.

(Refer Slide Time: 09:17)

Permanence and Community Detection

□Hypothesis 1:
- The number of internal connections of node v should be greater than the number of external connections of node v with any external community

□Hypothesis 2:
- In a community, all the vertices should be highly inter-connected to each other

□Expression for Permanence for a vertex v is:

$$Perm(v) = \left[\frac{I(v)}{E_{max}(v)} \times \frac{1}{deg(v)}\right] - [1 - c_{in}(v)]$$

- $I(v)$: Number of internal neighbours of v within its own community
- E_{max}: maximum number of connections of v to neighbors in an external community
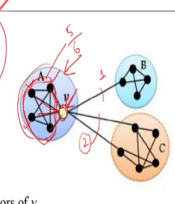- c_{in}: internal clustering coefficient of v

So, we combine these two things together to propose a metric called permanence. So, permanence is a local metric for a particular vertex v. We have two quantities. This is the first coordinate quantity corresponding to the first heuristics, second quantity corresponding to the second heuristics. Let us try to look at the components here. What is I v? I v is the number of internal nodes of v. What do you mean by internal nodes? number of internal connections of v.

(Refer Slide Time: 09:51)

Permance

$$Perm(v) = \left[\frac{I(v)}{E_{max}(v)} \times \frac{1}{D(v)}\right] - \left(1 - C_{in}(v)\right)$$

$I(v)$=Internal degree of v
$D(v)$=Degree of v
$E_{max}(v)$=Max. connection to an external community
$C_{in}(v)$=Clustering coefficient among internal neighbors of v

$Perm(v)=0.12$
$I(v)=4, D(v)=7, E_{max}(v)=2$
$C_{in}(v)=5/6$

So, let us say this is v. So, I v would be and let us say v belongs to this community. So, I v would be 1 2 3 and 4, I v would be 4 right. So, first quantity is I v. What is this one? This is E max v E max v is the maximum external pool right. In this case the there are two pools. So, the first pool is proportional to 2, the second pool is proportional to 1. So, the maximum pool is 2. So, this would be 2, this would be 4. So, as long as I v is higher than E max v which corresponds to the first heuristics.

My internal neighbor should be higher than my external neighbors; maximum external neighbors right, I am stable. So, this to higher this value higher the permanence value ok. Meaning higher the internal neighbors and lower the external neighbors higher the permanence value and this is normalized by the degree of v. Because I want that the permanence value should range between say minus 1 to 1 let us say ok or say whatever I mean 0 to 1 or minus 1 to 1 basically minus 1 to 1. So, we normalize this by degree. So, that this quantity will always be between you know 0 to 1 ok.

So, this corresponds to the fast heuristics, heuristics 1 and this is the second heuristics. So, what is C in v? C in v is the internal clustering coefficient of v. What do I mean by this? I measure the clustering coefficient of v with respect to the neighbors which are internal to the community where v belongs to. For example, in this case v belongs to this community and what are the internal neighbors? This one, this one, this one and this one, so I measure the clustering coefficient with respect to these 4 nodes, right.

So, you see there are total 5 connections among them right. So, this would be 5 and a what is the what is the total possible connections between 4 node 6 4 C 2. So, my internal clustering coefficient would be 5 by 6 right. So, what was the second heuristics higher the internal connection higher the permanence value. So, what I do here is basically I subtract this interconnection from 1. So, you can think of this quantity as a penalty ok.

So, now think about it higher the clustering coefficient lower this penalty right and is a minus right lower this penalty and higher the permanence value ok, clustering coefficient ranges between 0 to 1. So, if clustering coefficient is 1 then this would be 0 permanence would be maximum. So, when we will get the maximum permanence value? We will get the maximum permanence value when there is no external edge no external connection all internal connections.

So, therefore, this I v would be D v. So, this would be 1. Remember when this is 0 when no external connection this would be 0, then it would be 1 by 0 right. So, in that case we basically say that you know let us assume that when E max v is 0 we basically consider this as 1. So, that we can compute it easily otherwise 1 by 0 undefined and so on, we try to avoid that situation.

So, when a node has only internal connections no external connections and all the internal neighbors are connected, then this would be 1, this would be 0, this would be 1. So, 1 minus 0, permanence would be 1 ok.

(Refer Slide Time: 13:52)

Permanence and Community Detection

☐ Permanence of the entire network:

$$Perm(G) = \frac{\sum_{v \in V} Perm(v)}{|V|}$$
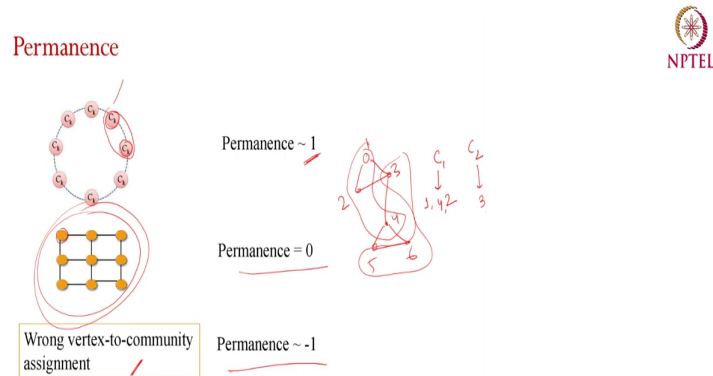
☐ permanence value ranges between -1 to 1
  ☐ when vertex v is a part of a clique, Permanence is 1
  ☐ when there is no appropriate community structure of a network (like a grid network), Permanence is 0
  ☐ when $I(v) \ll deg(v)$ and $c_{in}(v) \approx 0$, Permanence tends to -1

So, let us look at the limiting and then when we measure the permanence for a particular vertex we take the average permanence value of all the vertices right and that would give you the permanence of the graph. So, sum of permanence values divided by the number of vertices. So, the beauty about permanence is that it also tells you the quality of the community structure ok. So, if permanence ranges between minus 1 to 1. If it tends to 1 it tends to be 1, then you have a very strong community structure like a ring of cliques.
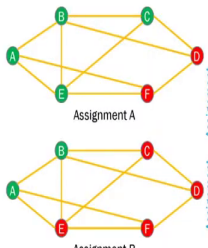
And interestingly if you use permanence maximization you will see that these cliques can be detected at separate communities. In contrast to the modularity maximization where modularity maximization tends to group small communities into big communities, but here you can detect small communities as well. When you have a structure like this a grid right which does not have any community structure I mean you can either consider each node as a community or the entire network as a community.

So, because grid does not have any community structure right permanence tends to be 0 ok and if you wrongly assign a vertex to a community right say a let us say; let us say you have a network like this ok 1 2 3 4 5 6 and you know that there are two communities C 1 and C 2, but you assign you assign node 1 node 4 and node 2 in one community. Node 3 node 5 and node 6 in another community like random group like this right which does not make any sense. Wrong grouping it tends to be minus 1 ok.

So, from the value of the permanence you basically you know you basically judge the I mean whether a network is at all qualified to be to pass into a community detection algorithm.

(Refer Slide Time: 16:04)

(Refer Slide Time: 16:09)

So, this is permanence. Now you know this is an example we have already discussed an example. Now how we can use permanence for community detection ok? So, we you know propose an algorithm called max perm which maximizes the permanence and it basically uses the same strategy the same strategy that we use for Louvain algorithm. So, we start so at

every pass if you remember the Louvain algorithm at every pass we have two steps modularity maximization and community aggregation.

In this case permanence maximization and community aggregation. So, at every pass we start with assigning nodes into different communities, then you start grouping. You keep on grouping and then you collapse groups into super nodes you create a super network. You keep on doing this thing until you see that permanence value decreases it is called permanence maximization.

So, in our paper we showed that if you use permanence maximization you will get better community structure compared to the modularity maximization. You will also be able to reduce the problem of resolution limit because permanence maximization can also detects small communities ok. For example, in case of ring of cliques you can detect each cliques as a separate community.

We also showed theoretically that it reduces the problem of degeneracy of solution and asymptotic growth right. But of course, it is not able to completely overcome these problems, but these problems will be reduced significantly if you use permanence maximization ok.

(Refer Slide Time: 17:48)

So, you can detect small size communities, you can reduce the problem of resolution limit and degeneracy of solutions. But the problem here is that in case of permanence since this is a vertex centric metric you actually need to measure permanence for all the vertices and that is

time taking. And the other problem is that clustering coefficient is order of n square I mean order of d square and the maximum value of d is n its order of n square.

So, if you can come up with a better way to detect to you know to measure clustering coefficient then max one would be a better metric better algorithm ok. So, this is all about disjoint community detection. In the follow up lectures we will discuss overlapping community detection, we will discuss two three algorithms in details and then we you know end this chapter by discussing how I mean how you can evaluate community structure ok.

Thanks.