

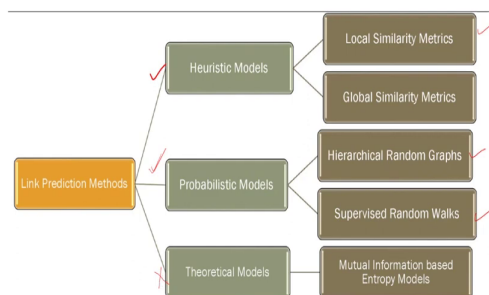
Social Network Analysis
Prof. Tanmoy Chakraborty
Department of Computer Science and Engineering
Indraprastha Institute of Information Technology, Delhi

Chapter - 06
Lecture - 03

Let us see some of the link prediction methods that we generally use although the methods that we are going to talk about here. These are generally used as baselines right for link predictions, but these are useful to understand ok.

(Refer Slide Time: 00:38)

Link Prediction Methods



So, if you look at the taxonomy of link prediction methods right you can see that mostly they are divided into this three categories.

The first one is heuristic based models where we essentially take some heuristics about the node attributes and you know the connection properties and then based on the heuristics set of heuristics we predict whether this link exists or not right. We will in this lecture we will discuss some heuristics based on local similarity local similarity measurement and some metrics based on global similarity measurement.

In fact, there are also a few metrics which fall in between those are called you know quasi semi quasi similar metric right quasi local similarity metric ok. So, the second category is

probabilistic models where in which we basically try to measure the probability of appearance of a particular edge right it can be supervised it can also be unsupervised.

For example this hierarchical random graph model that we did that we will discuss, here that is more of an unsupervised approach whereas, supervised random walk is a supervised link prediction approach that we will discuss. And we will skip this one this is more of a information theoretic model where we look at entropy based models and see whether mutual information based you know entropy based models can be useful for link prediction tasks.

(Refer Slide Time: 02:21)

Link Prediction: Heuristic Models



- Irrespective of the techniques used, the underlying idea of **link prediction** is
 - to successfully connect nodes that share some similarities, but are not linked as of now
 - closer/similar two nodes are, the more likely they are to be in agreement, and more likely they are to interact
- **Similarity between the nodes** can be derived using a combination of properties
 - Level of nodes ✓
 - Level of edges ✓
 - Level of metadata to the nodes
- **Heuristic measures of structural similarity**
 - Local Heuristics ✓
 - Global Heuristics ✓
 - Quasi-local Heuristics ✓



So, let us start with the you know metrics which are generally used to capture different heuristics ok. So, generally the similarity between nodes can be derived using a combinations of different properties it can be a combination of node level properties edge label properties as well as node and edge and metadata in related properties.

For example say right say node level structural properties include say degree clustering coefficient this kind of you know features. Whereas metadata of a node for example, in case of social network you can fetch the location information of a user or job or say you know say gender or some other interest political inclination and so on these are basically metadata related properties that we often use.

And as I mentioned we will look at three types of such metrics; one is called local heuristics, the second one is global heuristics and in between these two we have quasi local semi local kind of heuristics ok alright.

(Refer Slide Time: 03:46)

Link Prediction: Local Heuristic



- ✓ $G(V, E)$: an undirected dynamic network
- Three nodes $x, y, z \in V$ such that, at the current time instance
 - $(x, z) \in E, (y, z) \in E$
 - $(x, y) \notin E$
 - To decide the formation of the link (x, y) in near future
- Some local structural similarity base heuristic for the above
 - Common Neighbourhood ✓
 - Jaccard Similarity ✓
 - Preferential Attachment ✓
 - Adamic Adar ✓
 - Salton Index ✓
 - Hub Promoted Index ✓



So, local heuristics let us consider a graph an undirected graph $G(V, E)$ and there are say three nodes x, y, z these three nodes belong to V at the current of course, at the current time instance and. Let us assume that x and z are connected y and z are connected, but x and y are not connected. So, how do we decide that whether x and y will be connected in the future ok.

So, we will discuss this metrics, we will discuss something called common neighborhood metric common neighborhood similarity metric. Jaccard coefficient preferential attachment, Adamic Adar salton index, hub promotion index right and so on and so forth.

So, these are some of the examples example metrics that we will discuss today ok.

(Refer Slide Time: 04:36)

Local Heuristic: Common Neighborhood



$S(x, y) = A^2(x, y)$

- Triadic closure property
- By virtue of the common friend z, x and y are highly likely to be friends in future
- Common Neighborhood score between two randomly selected nodes x and y

$$S_{CN}(x, y) = |\Gamma(x) \cap \Gamma(y)|$$

Where $\Gamma(v)$: Neighbourhood set of node v

- Higher the number of common neighbours, more likely the node will be linked in future
- Example: In network G_1 , $S_{CN}(A, C) = |\{B, D, E, F\} \cap \{D, E, F\}| = 3$



So, the simplest one is the common neighborhood similarity.

It basically says that given two nodes x and y which are not connected. Let us look at the common neighbors ok x's neighbors and y's neighbors and the more this common neighborhood set the higher the similarity between x and y.

In this case here you see that A and C they are not connected, so this broken edge this does not exist. But A has node F, node B, node D, node E as neighbors node C has F E and D as neighbours. Therefore, the intersection of the neighbors of A and C would be 3 the size would be 3, so the similarity would be 3 ok.

So, this is the simplest measurement and in the matrix formulation if you think of if you take adjacency matrix A, essentially you are looking at the second of distance right. How many such paths how many such paths are there between A and C of size 2 right? So, this is one path ok, this is one path and this is one path.

So, essentially and how do we how do we measure the number of paths of size 2 of length 2? We just multiply a with itself a square. Now the similarity between x and y is basically the entry of x and y the x and x th and y th entry of a square matrix ok. So, this is common neighborhood the next one is Jaccard similarity.

(Refer Slide Time: 06:26)



Local Heuristic: Jaccard Similarity

□ Normalized version of common neighborhood score

□ Jaccard Similarity score between two randomly selected nodes x and y

$$S_j(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

□ The ratio of the number of common neighbors and the number of all neighbors of these two nodes

□ Example: In network G_1 , $S_j(A, C) = \frac{|\{B, D, E, F\} \cap \{D, E, F\}|}{|\{B, D, E, F\} \cup \{D, E, F\}|} = \frac{3}{4} = 0.75$



And so the problem in common neighborhood is that say for example, A has a lot of neighbors lot of neighbors and C has only one neighbor right. So therefore, automatically when we take the intersection right the intersection and say let us assume that the that is C's neighbor is also A's neighbor right. Say this is C and say and this is A right there is one common neighbour, but A has a lot of other neighbors.

So, automatically if you did the intersection it would be one right. So, you are, but A and C are not similar right. So but you are not penalizing the fact that A has a lot of degree right which lot of neighbors which are not common for C right. So, this matrix should be normalized ok, now Jaccard similarity is basically a normalized common neighborhood similarity in which the numerator is same, basically intersection of two neighbors and the denominator is the union of two neighbours.

Now when we take the union you basically penalize the fact that a node has a lot of neighbors and the other node has very few neighbors ok. Say in this case if you see node A has node A has four neighbors right, node B has three neighbors right intersection is three union is four therefore 0.75.

But if you take that example that I given that I have given earlier this one say this is C this is A and let us say there is a common neighbor x , but A has say 50 other neighbors ok. So, intersection would be one and the union would be what the size of the union would be would be 51, so 50 neighbors here and this one 51 ok 1 by 51 which is much much lesser right.

So, this is Jaccard index and then we say that ok right these two nodes are similar, then you basically rank all the node pairs based on Jaccard similarity and you return top five highly similar node pairs right and you say that ok you know this five pairs will be connected right in the future.

(Refer Slide Time: 09:18)

Salton Index

$$S(x,y) = \frac{|T(x) \cap T(y)|}{\sqrt{k_x \cdot k_y}}$$

Soham Index

$$S(x,y) = \frac{|T(x) \cap T(y)|}{k_x + k_y}$$

Hub Depressed Index (HDI)

$$S(x,y) = \frac{|T(x) \cap T(y)|}{\max\{k_x, k_y\}}$$

Hub Promoted Index (HPI)

$$S(x,y) = \frac{|T(x) \cap T(y)|}{\min\{k_x, k_y\}}$$

Leicht-Hofman Newman Index (LHN)

$$S(x,y) = \frac{|T(x) \cap T(y)|}{k_x \cdot k_y}$$



So, let us look at some other metrics. So, one is called Salton index ok, what is Salton index? Salton index is, so Salton index between x and y is neighbors of x right intersection neighbor of y ok and you take the cardinality of this and square root of $k_x \cdot k_y$. This is again another way of normalizing the common neighborhood right.

So, numerator is same as the Jaccard coefficient right the denominator is different here the denominator is essentially the degree right you are also considering. So, in the case of Jaccard coefficient the denominator was the size of the union right and when you take union you basically do not duplicate do not count the common neighbors two times ok common neighbors are counted one times right, but if you take the degree common neighbors are also connected two times.

So, in order to reduce the effect because it may happen that there are many common neighbors and you are counting this common neighbors double times right. So, you take the square root, so the effect will reduce ok; so this is Salton index.

Similarly there is another metrics , so these are look. So, I mean you can also come up with your own way you can say that look I will not take square root I will take write 1 by 3 ok or 1 by 4 does not matter right. It depends on the application and depending on that depending on the application you can think of your own ways. So, the other one is called Sorensen index ok, now these names are these names are coming from the inventor of this metrics ok.

So, this is S_{xy} is two times intersection of this one ok essentially you are taking the intersection and then you are normalizing it by the average degree of x and y $\frac{k_x + k_y}{2}$ right, this is called Sorensen index ok. So now, let us look at another interesting metric called hub depressed index ok hub depressed index ok HDI.

So, this HDI between x and y is neighborhood of x intersection neighborhood of y, you see that this numerator is common almost same right for all the cases only the denominator changes. And then you have max here we are not taking the average here we are taking the max.

So, this is called you know hub you know hub depressed index and similarly another metric called let me write these two metrics side by side hub promotion index, hub promoted index promoted index ok HPI. So, $S_{HPI, xy}$ is intersection of neighbors of x and y divided by mean of k_x and k_y what does it mean ok what does it mean, so ok. So, let us look at this denominators carefully ok.

So, this is called hub promoted index meaning that if x is connected to a node y which is a hub, so hub will promote you ok. So, let us say sorry I mean x and y are not connected ok. So, you are trying to understand whether x and y will be connected or not. So, if x is a x I mean if we are measuring right the similarity between x and y and one of them is hub. So, if one of them is hub then its degree would be huge right say y is hub.

So, S_{xy} would be much greater than S_{xx} because it is a k_x of x because it is a hub. So therefore, when you take the minimum of this always the other node which is not hub that will be considered here in the denominator. So, in other ways hub basically promotes the other node to use its own degree. So, you will basically be better off if you are if you are surrounded by hubs.


Similarly, this is the other one is called hub depressed because we are taking max right. So, if you are surrounded by hubs the hubs degree will be dominated. So, hub the hubs degree will

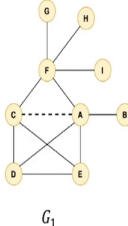
dominate this metric right and because you are you are taking max. So, so k of y will be considered here right.

So, this would be high ok your degree would not play any role. So now, these are some of the metrics which are which we use in not only link prediction (Refer Time: 15:37) also for I mean in general simulated measurement ok.

(Refer Slide Time: 15:42)

Local Heuristic: Preferential Attachment






G_1

- Derived from the concept of preferential attachment of scale-free networks
- Likelihood of a node x to obtain a new edge is proportional to k_x , the degree of the node
- Preferential Attachment score between two randomly selected nodes x and y

$$S_{PA}(x, y) = k_x \times k_y$$
- Future interaction between them depends on the existing degree of the individual nodes
- Example: In network G_1 , $S_{PA}(A, C) = k_A \times k_C = 4 \times 3 = 12$



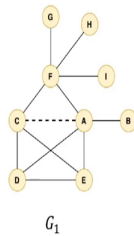
So, the next one is called preferential attachment, now preferential attachment is something that we already discussed in the network growth model chapter right. So, here the idea is that the philosophy behind preferential attachment is that two nodes are connected to due to I mean a node will get connected to a node based on its degree right.

So, it is highly likely that the degree both the degree would play an important role right. So, likelihood of a node x to obtain an a new edge is proportional to k_x this is called preferential attachment. Therefore the preferential attachment score between two randomly selected nodes x and y is just the multiplication of the degrees k_x times k_y ok.

So, this is the very simple idea and this kind of matrix this has a lot of applications in different you know to quantify different functional significance of links, subject to various say you know node based dynamics such as percolation right propagation synchronization transportation and so on and so forth.

(Refer Slide Time: 17:12)

Local Heuristic: Adamic Adar



□ Primary Objective: shift focus towards rare events

□ Assigns higher weights to less-connected nodes

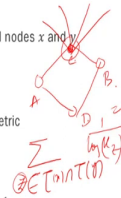
□ Adamic Adar metric between two randomly selected nodes x and y

$$S_{AA}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}$$

□ Resource Allocation Index is variant of the above metric

$$S_{RA}(x, y) = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z + 1}$$

□ Example: In network G_1 , $S_{AA}(A, C) = \frac{1}{\log 3} + \frac{1}{\log 3} + \frac{1}{\log 5} = 5.62$



So, now let us look at a very important metric ok, but before that let me you know discuss another metric which people generally use. Leicht Holme Newman index right LHN ok.

This is kind of similar intersection of the neighbors, but the denominator is k_x times k_y we have taken we have seen average max mean now multiplication ok alright. So, now, let us look at a very important metric right, now this is this has been used as a baseline for quite some time although it was proposed long time back. This is called Adamic Adar distance or Adamic Adar similarity right and two researchers Lada Adamic and Adar, so they proposed this metric.

So, Adamic Adar distance between two nodes x and y is sum over we look at the common neighbors ok. You see here $\Gamma_x \cap \Gamma_y$ right common neighbors and z is one such common neighbor and we take 1 by \log of k_z what is the intuition behind this ok, let us try to understand.

So, let us say A and B ok and they have common friends two common neighbors C and D . So, C you see that C has a lot of degree alright C is connected to many nodes whereas, D is connected to only A and B ok. It means that C is more D is more dedicated to A and B compared to C .

So, any information if it comes to C right if it comes to C the likelihood that C will pass that information to B say right, say an information moves from A to C the likelihood that C will pass the information to B its is lower than the same for D ok.

So, we basically penalize those nodes those common neighbors which have a high degree ok. In this particular case you see, C will be penalized because C is not dedicated to A and B whereas, D is dedicated ok. So, the metric basically says that we take the sum the intersection of the neighbors of x and y and z is one such common neighbour, we take $1/k_z$, k_z is a degree of z.

So, higher the degree lower the importance because we take $1/k_z$ right of course, we can take log because we want to dampen the effect, so we take the log ok. In fact, there is another metric called resource allocation index same as Adamic Adar, but here they do not take the logarithm of the degree they just take the degree ok.

Now this is a very interesting idea ok. You are basically saying that it is good to be connected with less nodes less to be connected with less number of nodes which are common neighbors which have less degree compared to the case where you are connected to a lot of nodes, but the common neighbors have higher degree ok.

(Refer Slide Time: 21:31)

Local Heuristic: Salton Index and Others



□ **Salton Index** (Commonly used metric to measure the similarity between a pair of documents or embeddings in a vector space) between two randomly selected nodes x and y

$$S_{SI}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_y}}$$

□ **Hub Promoted Index** (used to assign high scores to links adjacent to hubs) between two randomly selected nodes x and y

$$S_{HI}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min(k_x, k_y)}$$

□ **Hub Depressed Index** (used to assign low scores to links adjacent to hubs) between two randomly selected nodes x and y

$$S_{HDI}(x, y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max(k_x, k_y)}$$



So, Salton index hub promotion hub depressed we have already discussed.

(Refer Slide Time: 21:37)

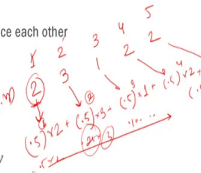
Global Heuristic: Katz Score



- Inspired by Katz centrality
- Takes into account the influence by neighbors beyond 1-hop
- However, longer the path length, less likely the end nodes influence each other
- Between two random nodes x and y, Katz score is given by

$$S_{KZ}(x, y) = \sum_{p=1}^{\infty} \alpha^p \cdot A_{x,y}^p$$

Here, $A_{x,y}^p$: number of paths of length p that exists between x and y
and α : damping factor that reduces the impact of longer paths



Let us look at the global heuristics. So, these are local heuristics the metric that we have discussed so far. In the local heuristics we only look at the neighbors and based on that we judge right. Now global heuristics with is with respect to the entire graph ok the first one is Katz centrality sorry Katz score right and this Katz score is motivated by the Katz centrality that we discussed earlier ok.

So, the Katz score between x and y is forget about this part right is the number of number of paths of length p ok and we take all paths all possible paths of all possible lengths. Say you have length 1, length 2, length 3, length 4, length 5 right and between x and y there are two paths of length 1 three paths of length 2.

Remember there can be there can be multiple shortest path between a pair of nodes ok. So, you have two shortest paths of length 1 I mean of course, let us not you know let us not use the term shortest path here. You have two paths of length 1, three paths of length 2, one path of length 3, two paths of length 4 and two paths of length 5 ok. We basically sum them up ok, but we also penalize the paths having higher length ok which is controlled by alpha say alpha is point say alpha is 0.5.

So, you penalize this in this way. So, the component the contribution of this part would be 0.5 to the power 1 times 2 plus this would be 0.5 to the power 2 times 3. Now this 1 is the length 2 is the length this would be so plus 0.5 to the power 3 times 1 this would be 0.5 to the power

4 times 2 and this would be 0.5 to the power 5 into 2 right. So, as you move further right you see the contribution will also decrease.

So, this is 0.5 times 2, this is 0.25 times 2³, this is 0.125 times 3. So, although you may have more number of shortest path more number of paths say 3² whatever, but the contribution will decrease because now you multiply it by 0.25 then 0.125 and so on. This is exactly same as the Katz centrality that we discussed, but here we take the pair of nodes right and this alpha is a damping factor same as the Katz centrality ok.

(Refer Slide Time: 24:55)

Global Heuristic: Hitting Time



- Based on random surfing model. A random surfer
 - a) starts at node x
 - b) moves to a neighbor of x chosen uniformly at random
 - c) repeats step (a) till it reaches y
 - Hitting time (HT_{xy}): Expected number of steps it takes for a random surfer starting at x to reach y
 - The Hitting Time score between nodes x and y is given by

$$S_{HT}(x, y) = -HT_{xy}$$
 - Smaller the hitting time between two nodes, closer in proximity the nodes, therefore higher the chances of their interaction in future
 - The Normalized Hitting Time score between nodes x and y is given by

$$S_{HT}^{norm}(x, y) = -HT_{xy} \cdot \pi_x$$
- Here π : stationary distribution of PageRank for the network



The next one is called hitting time, what is hitting time? So, hitting time before understanding this again is based on the this random surfing behavior right random walk process. So, you start your random surfing from a node x right you basically move to a neighbor of x again you just choose one of the neighbors uniformly at random and then you move there again from that node you choose another node uniformly at random again the neighbors one of the neighbors and then you move there.

So, hitting time between x and y is the expected number of steps a random walker needs for a needs to move from x to y . So, why do I say that expected number of steps because this is a random walk process and depending upon your trials the path length will change right.

So, you basically repeat these experiments again and again and you take the expectation of this and that would be your hitting time and we take the negative minus why because we

want that higher the hitting time. So, lower the hitting time better would be right better would be the similarity or higher would be the similarity.

So, lower the hitting time higher the similarity between x and y therefore, we take the minus because if we take the minus then higher I mean then we can say that you know higher the negative of hitting time higher the similarity right. So, higher the better, so the in order to satisfy this we basically do this thing.

So, sometimes what happens is that you know this measurement with the hitting time this is not normalized ok if this is not normalized what I mean you basically want to make this thing between zero to one. So, what you do you multiply this with the stationary distribution that we obtained from the PageRank hope you remember what is PageRank.

You basically repeat because PageRank always guarantees that this would be this would be the stationary distribution the stationary distribution the sum would be 1 and each would be between 0 to 1. So, we basically multiply this quantity with the stationary distribution of I mean the value of x from the stationary distribution of the PageRank process ok.

(Refer Slide Time: 27:54)

Global Heuristic: Commute Time



- Extending upon the random walk model. A random walker
 - a) starts at node x
 - b) moves to a neighbor of x chosen uniformly at random
 - c) repeats step (a) till it reaches y
 - d) travels back to x (not simply jumps to x)

□ The Commute Time score between nodes x and y is given by

$$S_{CT}(x, y) = -C_{xy} = -(HT_{xy} + HT_{yx})$$

□ Smaller the commute time between two nodes, closer in proximity the nodes, therefore higher the chances of their interaction in future

□ The Normalized Commute Time score between nodes x and y is given by

$$S_{CT}^{norm}(x, y) = -(HT_{xy} \cdot \pi_x) + HT_{yx} \cdot \pi_y$$

Here π : stationary distribution of PageRank for the network



The next one is commute time.

So, here idea again the idea is same, but here we look at both the things. So, in the hitting time we start from x and see when the random walk will hit random walker will hit y right

here you start from x and you measure the hitting time to reach y you also start from y and you measure the hitting time to reach x right.

So, HT_{xy} , HT_{yx} you take the sum and that is your commute time ok and of course, you take the negative the same due to the same reason. And if you want to make it normalized you multiply it by the stationary distribution of x and y in the PageRank process ok.

(Refer Slide Time: 28:46)

Probabilistic Models: Hierarchical Networks



- A network is said to be a **hierarchical network** if
 - the vertices can be divided into groups,
 - each of these groups can further be subdivided into groups of groups, and so on
 - each group formed in a logical order corresponding to a granular functional/social unit
- Can easily be rendered as a tree or a **dendrogram**: Nodes of a network form the leaves of the dendrogram
- Smaller the height of the links between the groups or the nodes, the higher the similarity between them**



So, this is pretty much about the local heuristics that we use local and global heuristics that we use for link prediction ok. For essentially for you know measuring the similarity between two nodes.

But if you remember with the in the link analysis chapter we also discussed measures like SimRank PathSim right. So, SimRank PathSim those are also used for measuring the similarity between two nodes right SimRank was SimRank in SimRank if you remember, it was basically based on the idea that how similar the neighbors of your nodes of the given pair of nodes right.

In case of PathSim we look at the heterogeneous graph and then we look at the meta path and we look at number of meta paths of certain types and so on and so forth. So, those matrix can also be used for link prediction ok. So, we stop here we will discuss this algorithm in the next lecture ok.

Thank you.