


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

Chapter - 04
Lecture - 25
Lecture - 06


So, we have we have been able to discuss quite a few things in this chapter. We have discussed PageRank, DivRank. Even before that, we discussed some simple approaches to measure the link right. So, in this lecture, we will discuss another important metric or quantities called SimRank. And the intuition of SimRank is very different from say PageRank or related measures.

(Refer Slide Time: 00:50)



Measuring Similarity of Objects

- Metadata used to measure similarity between objects are often hard to determine and quantify in practice
- Contextual information may be used for the purpose
 - Two objects are similar if they are related to similar objects
 - Easier to determine in practice
- SimRank follows the above paradigm to measure similarity between entities
- SimRank is efficient as well
 - For a network of size N , we require N^2 similarity scores, one per each pair of objects
 - For the same network, a score like SimRank would form a list of length N !



So, here the idea is that you know an two objects are similar, if they are related to similar objects. Now, the difference between PageRank, DivRank, and this kind of SimRank kind of methods is that PageRank DivRank they basically produce ranks of nodes ranking of nodes, right. It basically measures the prestige of a node. It is a node centric measure whereas, SimRank is a measure for a pair of nodes, for a pair of objects ok.

So, the idea is that two nodes, two objects are similar if they are referred, they are related or they are cited by similar types of nodes ok. So, here what is the idea? The idea is that we are measuring the similarity between a pair of nodes right. And our task is to return you know

return those pairs which are similar whose similarity values are higher. In other words say for example, you are given a node and you are asked that hey, look at other nodes and tell me that out of the other nodes which four nodes are similar to the given node ok?

So, your node is given. So, what you would do? You would basically measure the similarity between this node and this node, these and these, these and these, these and these and so on and so forth. You cannot do this thing using PageRank right. So, therefore, this SimRank kind of approaches were proposed. So, given a network of size N, you need to compute N square such similarity measures ok.

(Refer Slide Time: 02:39)

**SimRank:
Measuring Similarity of Objects**


- Paper E cites papers C and D
- Papers C and D appear similar
- Paper H cites papers B and G
- Papers B and G appear similar
- What about the similarity of papers A and B?
- $\Gamma(A) = \{C, F, G\}$ and $\Gamma(B) = \{D, H\}$
- SimRank can answer such question

So, here is the idea. So, this is the network and say let us say this is a citation network ok. And node E cites paper C and D. Therefore, C and D although they are not connected to each other directly, but they are related because they are cited by the same paper E. Node H cites both B and G. So, B and G are related because they are cited by same paper H ok. What about A and D?

So, A if you look at the you know the papers who are citing A you have C, F and G. So, the neighbors of A is C, F and G, but neighbor of D is what? Neighbor of D is sorry this is neighbor of B ok. This is A and B ok. So, neighbor of B is D and H. So, there is no neighborhood intersection. Since, there is no neighborhood intersection one may say that you know A and B are not related ok, but if you think carefully; if you think carefully they are related in some ways.

For example, B is cited by H right; G is also cited by H ok. So, B and G are somehow related ok. Now, G is citing A. So, in indirectly H is also citing A in two hops right. Now, therefore, A and B are cited by G in some ways. So, G is citing A, A and B are cited by H in some ways. So, H is citing B directly H is citing A indirectly. So, in that sense A and B are connected, A and B are related similar ok.

(Refer Slide Time: 05:02)



SimRank: Basic Formulation

□ For a node v in the network $I(v) = \{i, (v) | 1 \leq i \leq |I(v)|\}$ and $O(v) = \{o, (v) | 1 \leq i \leq |O(v)|\}$ denote the sets of indegree and outdegree neighbours, respectively.

□ Formulate the similarity score $s(u, v) \in [0, 1]$ as follows:

$$s(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } I(a) = \emptyset \text{ or } I(b) = \emptyset \\ \frac{|I(a)| |I(b)|}{|I(a)| |I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b)) & \text{otherwise} \end{cases}$$

• A node is **maximally similar** to itself
 • No way of determining the score for a neighborhood that does not exist
 • Similarity between two randomly selected nodes is proportional to the **average similarity** between their neighbors

Now, how do we; how do we do this thing computationally? So, let us assume that you know there are two quantities. If the graph is directed then you have inward edges and outward edges; inward neighbours and outward neighbours right. So, $I v$ is the set of inward neighbours and $O v$ is the set of outward neighbours ok. So, v this sets. So, this node this node these are inward neighbours of v and these are outward neighbours of v ok, outward neighbours of v .

So, the similarity between a and b is 1; if a and b are same because the I mean an item is I mean similar to itself with at most quantity and that is 1 ok. If $I a$ and $I b$ meaning the inward neighbour of a or the outward neighbour of b ; any of these sets is empty then this is 0. You cannot compare right. Let us say there are two nodes right; something like this right.

How do you compare between these and these nodes? Although they are related in terms of the outward neighbour, but they are not related in terms of the inward neighbours ok. So, we are discussing this thing in terms of inward neighbours. So, otherwise if this set and this set

are not empty, then the similarity of a and b is the sum of similarities of all their neighbours, all their pairwise neighbours.

What do you mean by this? Let us say; let us say ok. So, this is a, this is b, this is c, d and e ok. So, I of a is I of b is sorry c, d, e ok, right. So, I take all pairs of inward neighbours meaning I will take c, d, c, e, d, d and d, e. So, this summation the summation would be over all pairs you see here double summation and they are simulated.

So, similarly between c d, similarity between c e, similarity between d d similarly between d e, and then you take the sum ok. And then you normalize it by the total number of pairs. So, how many pairs are possible? So, mod I a times mod I b right. In this case 2 times 2 right. Forget about this c, I will discuss what is c ok; c is a constant I will discuss what is c. So, this is the similarity ok.

(Refer Slide Time: 08:36)

SimRank: Naïve Solution

□ An iterative solution for SimRank is as follows:


$$R_0(a, b) = \begin{cases} 0 & \text{if } a \neq b \\ 1 & \text{if } a = b \end{cases}$$


and

$$R_{k+1}(a, b) = \begin{cases} 1 & \text{if } a = b \\ \frac{c}{|I(a)| \cdot |I(b)|} \sum_{i \in I(a)} \sum_{j \in I(b)} R_k(I_i(a), I_j(b)) & \text{if } a \neq b \end{cases}$$

with

$$\lim_{k \rightarrow \infty} R_k(a, b) = s(a, b)$$






Now, let us make it even more concrete. What I say is that I start with a simulate. So, here as and R as same; assume that as and R as same ok. So, I denote s I denote similarity by R. So, similarity at 0 iteration 0th iteration between a and b is either 0 or 1. If a and b are same then 1, if a and b are not same then 0. And over time over time you basically update it right. And this is the updation rule if a and b are same then 1, if a and b are different then the quantity that we discussed. And then we repeat it. We stop when R of k plus 1 a, b is kind of same as R of k a comma b ok.

(Refer Slide Time: 09:35)

SimRank in Heterogeneous Bipartite Network




□ In a heterogeneous network of users and products, the similarity of products and users are **mutually-reinforced**

- two users can be considered similar if they buy similar products
- two products can be considered similar if they are bought by similar users

□ Similarity between two distinct users can be expressed as:

$$s(u_1, u_2) = \frac{C_1}{|O(u_1)| \cdot |O(u_2)|} \sum_{i=1}^{|O(u_1)|} \sum_{j=1}^{|O(u_2)|} s(O_i(u_1), O_j(u_2))$$

□ Similarity between two distinct products can be expressed as:

$$s(p_1, p_2) = \frac{C_2}{|I(p_1)| \cdot |I(p_2)|} \sum_{i=1}^{|I(p_1)|} \sum_{j=1}^{|I(p_2)|} s(I_i(p_1), I_j(p_2))$$


And, if you have a kind of a bipartite network say you have users and products right. You can use the same quantity ok. So, you say that. So, this is product. So, a product is basically the similarity between two products is the similarity between users who have bought these two products; whereas, the similarity between users is the similarity between products which these two users have bought in the past, same ways right.

So, similarity between users is the similarity between products ok. Here outward neighbours of users is basically products you see these are outward neighbours for a user ok. And for a product you basically look at the inward neighbours and their similarity. So, two users can be considered similar if they buy similar products two products can be considered similar if they are bought by similar users ok.

(Refer Slide Time: 10:53)


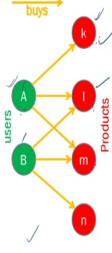


Illustration: SimRank in Heterogeneous Bipartite Network

To calculate the similarity between users A and B




$$O(A) = \{k, l, m\} \text{ and } O(B) = \{l, m, n\}$$

$$I(k) = \{A\}, I(l) = \{A, B\}, I(m) = \{A, B\}, \text{ and } I(n) = \{B\}$$

$$s(A, B) = \frac{1}{3 \times 3} (s(k, l) + s(k, m) + s(k, n) + s(l, l) + s(l, m) + s(l, n) + s(m, l) + s(m, m))$$

$$s(A, B) = \frac{c_2}{3 \times 3} (s(A, l) + s(A, m)) = \frac{c_2}{9} (1 + s(B, B)) = \frac{c_2}{9} (1 + \frac{c_2}{2})$$




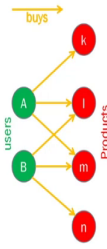
Let us take an example say this is a graph and bipartite graph and you have users A, B you have products k, l, m, n ok. And you want to calculate the similarity between A and B. So, for node A what are the outward neighbours? k l m. So, $O(A)$ is k, l, m, $O(B)$ is l, m, n right. For product l, the inward neighbour is A; product l A, B; product m A, B; product n B ok.

So, now, the similarity between A and B is basically we look at these two things all pairs. So, similarity between k l, k l, k m, k n, right; l l, l m, l n, m l, m m and m n 9 pairs ok. One term is missing this is m n. So, I need to calculate all these things all these numbers one by one and this is normalized by this one 3 times 3 some c 1 ok.

(Refer Slide Time: 12:28)

Illustration: SimRank in Heterogeneous Bipartite Network





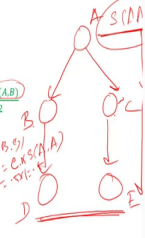
Similarly, $s(k, m) = \frac{c_1}{2} + \frac{c_2 s(A, B)}{2}$, $s(k, n) = c_2 s(A, B)$


$s(l, l) = 1$, $s(l, m) = \frac{c_1}{2} + \frac{c_2 s(A, B)}{2}$, $s(l, n) = \frac{c_1}{2} + \frac{c_2 s(A, B)}{2}$

$s(m, l) = \frac{c_1}{2} + \frac{c_2 s(A, B)}{2}$, $s(m, m) = 1$, $s(m, n) = \frac{c_1}{2} + \frac{c_2 s(A, B)}{2}$

Solving, $s(A, B) = \frac{3c_1 c_2 + 2c_1}{9 - 4c_1 c_2}$

Further, setting $c_1 = c_2 = 0.8$,
 $s(A, B) = 0.547$





So, now, let us look at this quantity ok. So, let us you know let us look at; let us look at this one. Let us look at s of, let us look at the easiest one this is the easiest one s of l, l. This is 1, this is also 1 ok. So, what is s of k, l? Let us try to understand s of k, l the first quantity. So, s of k, l is some sort of c 2 right. So, k and l their inward neighbours are A and B. So, this would be s of A, A k, l right k and l.

So, A A and A B, A, A plus s of A comma B. And this is 2 into 1 ok. So, c 2 into 2, this would be 1. And this would be s of A, B which you do not know right. So, c 2 by 2 plus c 2, s A comma B by 2 ok. What is s of, what is s of k, m? s of k, m s of k, m is same as we look at; we look at k and m. So, same pair A, A and A, B. So, this would also be this one ok. So, this is done, this is done. Now, let us look at s of k, n; s of k, n means you have this and this.

So, you have only one pair A, B right. So, s of k, n is c 2 times s of A, B ok. Try to understand what is happening. So, I am writing all these similarity in terms of s of A, B; s of k, l is a function of s of A, B; s of k, m is a function of s of A, B; s of k, n is a function of s of A, B right. Now, let us look at s of l, l. This is 1. What is s of l, m? s of l, m is A, A; A, B; B A; B B; s of l, m right. So, you have you have all pairs. So, let me check less of l, l, m right l, m ok. So, s of l and m. So, l and m l and m.

So, you have you have basically same pairs A and B. So, you have A, A; A, B; and B, B ok. You can also write in terms of s of A, B ok. So, if you think about it carefully, you will see that all these similarities can be either 1 or a function of s of A, B ok. So, finally, s A, B I


mean if you add everything you will see that s of A, B would be mean if you when you add and move all s of A, B to the left side you will get this one right based on $c_1 c_2$.

Now, $c_1 c_2$ are constant that we can fix. Let us say c_1 is 0.8 and c_2 is also 0.8 this is given. So, therefore, s of A, B would be 0.547 ok. So, this is the idea behind SimRank. Now, what is this c ? The c is something that allows you to you know let us take an example and you will understand ok. So, say this is A, B, C, D, E. So, the similarity between A and A is 1, the similarity of the similarity of B and C is basically would not be 1 because this is maximum.

So, would be something lesser than 1. Now, this lesser factor is constant C. So, this ranges between 0 to 1. And let us say this is 0.5. So, similarity of B, C is C times similarity of A. So, 0.5 into 1 is 0.5. So, the similarity of A, B is less than similarity of A, A similar to D, B similar to D, E is what? Similarity of D, B D is C times similarity of B, C and that is C square times similarity of A, A.


So, it was C here this is now C square. So, 0.5 times 0.5 which is 0.25. So, the C is kind of gives you less and less lesser and lesser weights as you move you know further from A to the I mean to the given node to the rest of the nodes. This factor C actually ensures these thing ok. And C can be defined based on your requirement alright.

(Refer Slide Time: 18:47)



Heterogeneous Networks

- A tuple of the form $(V, E, \mathcal{A}, \mathcal{R}, \phi, \psi)$ represents an information networking system if
 - V is the set of vertices
 - E is the set of edges
 - \mathcal{A} is the set of different node types present in the network
 - \mathcal{R} is the set of different link types present in the network
 - $\phi(v): V \rightarrow \mathcal{A}$ maps each vertex to a node type
 - $\psi(e): E \rightarrow \mathcal{R}$ maps each edge to a link type
- If $|\mathcal{A}| = 1$ as well as $|\mathcal{R}| = 1$, then the system is termed as a homogeneous network
- On the contrary, if $|\mathcal{A}| > 1$ or $|\mathcal{R}| > 1$, or both, then the system is termed as a heterogeneous network



So, this is about SimRank. Now, we are almost at the end of this chapter and we will discuss the last method last metric called PathSim ok. And this is based on something very interesting

ideas based on something called heterogeneous network and meta path ok. So, what is heterogeneous network? Heterogeneous network you already discussed; nodes can be of different types, edges can be of different types ok.

So, V is the set of nodes, E is the set of edges right, A is the set of different node types. It can be papers authors E R is the separate edge types citations co author shape and so on. And then you have a mapping which maps a node to node type; and there is a mapping which maps an edge to an edge type ok. Now, if $\text{mod of } A \text{ equals to } 1$, meaning that you have only one type of nodes right. This is node homogeneous. If $\text{mod of } R \text{ equals to } 1$, this is edge homogeneous. And if both of them are greater than 1 then this is both node and edge heterogeneous network ok.

(Refer Slide Time: 20:12)

Heterogeneous Networks: Variants

□ When $|\mathcal{A}| > 1$ and $|\mathcal{R}| = 1$, then we have a heterogeneous network consisting of vertices of more than one types, and only one types of links

□ A typical example is **consumer-product purchase network**, where

- $\mathcal{A} = \{\text{users, products}\}$, and
- $\mathcal{R} = \{\text{user} \rightarrow \text{products} \mid \text{user buys product}\}$

So, think of this user product bipartite network. This is actually node heterogeneous network. You have different nodes user's nodes and product nodes, but edges are same. All these edges indicate some sort of buy relationship ok.

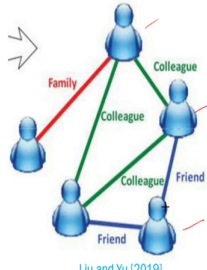
(Refer Slide Time: 20:33)

Heterogeneous Networks: Variants



□ When $|\mathcal{V}| = 1$ and $|\mathcal{R}| > 1$, then we have a heterogeneous network consisting of vertices of one type, but there are more than one type of links between these vertices

□ A typical **online social networking platform**;

- only one type of vertices, viz. users of the network;
- There are more than one type of links: friends in real life, family members in real life, office colleague in real life, and so on.



Liu and Yu (2019)



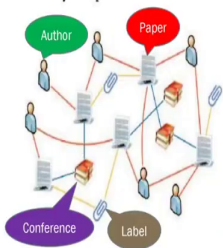
So, and this is a kind of edge heterogeneous network where nodes are users, but edges can be friendship relation, can be family relation, can be colleague and can be anything right. So, edges are heterogeneous.

(Refer Slide Time: 20:50)



Heterogeneous Networks: Variants

□ When both $|\mathcal{V}| > 1$ and $|\mathcal{R}| > 1$, then we have a heterogeneous network consisting of vertices of one type, but there are more than one type of links between these vertices

□ A typical **bibliographic network** consisting of authors, papers, conference venues, etc., and various kinds of relationship between these entities



<https://www.semanticscholar.org/paper/11Rank%3A-A-Path-based-Ranking-Framework-in-Network-Li-Shi/185687269aa420ca87ba4538739226e0a3c9d>



We you can have both nodes and heterogeneity for example, citation network nodes can be papers or authors links can be citations or co authorship and so on.

(Refer Slide Time: 21:05)

Heterogeneous Networks: Network Schema



APA

□ A meta-data level outline for a heterogeneous directed network $G(V, E)$ and the information tuple $(V, E, \mathcal{A}, \mathcal{R}, \varphi, \psi)$, where $\varphi: V \rightarrow \mathcal{A}$ is the object type mapping, and $\psi: E \rightarrow \mathcal{R}$ is the link type mapping. The corresponding network schema is given by $T_G = (\mathcal{A}, \mathcal{R})$

(A) DBLP network with a star network schema

(B) Douban Movie network with a general network schema

https://www.researchgate.net/publication/314120795_General_network_schema_agnostic_sparse_tensor_factorization_for_single-pass_clustering_of_heterogeneous_information_networks



So, on the heterogeneous network we define something called network schema ok. What is network schema? Network schema is a relationship between node types not nodes. The schema is defined on the node types ok. You basically say that you know say you define a schema like this on a citation network which means that an author can only be connected to a paper through citation through the you know edge type right.

So, author can write a paper a venue a publication venue can be connected to a paper through the relation publish a keyword a key term can be connected to paper in terms of contains, paper contains the term and in the particular term. So, this is basically a network schema. So, network schema is basically the is giving you an idea about how different types of nodes, different node types are connected through which types of links ok.

You cannot connect an author to a term, you cannot connect like, this is not allowed because this schema has to be defined earlier. And then based on the schema you can create your network ok. Those who have studied DBMS this is very related to entity relationship diagram EER diagram if you remember correctly ok. There also you have entities they have relations you have schema right and so on.

(Refer Slide Time: 23:02)

Heterogeneous Networks: Meta-Path

- A meta-path is a meta-level description of the structural connectivity between the entities
- Different paths deliver varying semantic similarity/differences or measure different topological connectivity
- A meta-path is a path \mathcal{P} of length ℓ defining a composite relation over the ℓ links $\mathcal{R} = \mathcal{R}_1 \circ \mathcal{R}_2 \circ \mathcal{R}_3 \circ \dots \circ \mathcal{R}_\ell$ and $\ell \neq 1$ objects $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \dots, \mathcal{A}_{\ell+1}$ denoted in the form $\mathcal{A}_1 \xrightarrow{\mathcal{R}_1} \mathcal{A}_2 \xrightarrow{\mathcal{R}_2} \mathcal{A}_3 \dots \mathcal{A}_\ell \xrightarrow{\mathcal{R}_\ell} \mathcal{A}_{\ell+1}$

<https://www.researchgate.net/publication/339302745/figure/fig/1/339302745>

So, now we define a new concept called Meta-Path ok. Meta path is same as path, but on the heterogeneous network ok. So, meta path of length 1, this is defined by a composite relationship of length 1. And the meta path is defined on the network schema not on the graph ok. If I say that there is a meta path from there is a meta path from author to paper to author meaning I am indicating this meta path ok. If I say. So, this meta path has a certain semantics APA, what does it mean?

This paper this author is written this paper has written this sorry this author has written this paper; and this author has also written this paper. So, it this APA relation indicates the co authorship relations between these two authors ok. Let us say there is another meta path author to paper to venue to paper to author. So, author has published this paper. This paper is published in this venue say this venue see KDD right. This paper is also published in this venue, and this paper is written by this author.

Meaning that these two authors are not direct co authors, but they are they tend to publish papers in same venues ok. So, this meta paths different types of meta paths indicate different semantics ok. So, if I say that if I say that say there is a meta path of length 3, there is a meta path of length 3 from between author between A 1 to A 3, meaning there is a; there is a node type A 2 and the link a particular link type R 1 and R 2 ok.

So, therefore, it is a composite relation it is a composition of all the relation. So, in this case the relation is author right published whatever publish relation or written relation write or


written relation and this is also write or written relation. Remember one thing, that if there is a link from A to author to paper and the relationship type is right, there is another link from paper to author with the relation type written by or write inverse ok.

There is relation from venue to paper through publish the same relation exists from paper to venue with the link publish inverse or published by and so on right. So, if I tell you that when I move from author to paper am I be allowed to move from paper to author? Yes, through this link ok. If I say that there is a meta path of type this one author writes paper contains term right contains term contain inverse paper right written by author this is also possible ok.

So, when I say that I am interested in APA kind of meta path it means that in the original graph in the original heterogeneous graph I have interested in all such path which connects two authors through a paper relation through a paper node. So, now, in the original graph you have author 1, author 2, author 1 and author 2 are connected to paper through paper 1 right; author 7 and author 8 author 7 and author 8 are connected through paper 5.

So, I am actually referring to all those paths which follow this meta path APA ok. So, I can also count the number of instances of the number of instances number of paths of this type APA right.

(Refer Slide Time: 27:58)




Object Similarity via Meta-Path

Path Count: It indicates the number of path instances p of \mathcal{P}_p , which begin at x and end at y . The similarity score is

$$s(x, y) = |\{p \in \mathcal{P}_p | x \in \mathcal{A}_t, y \in \mathcal{A}_{t+1}\}|$$

Random Walk: For a random surfer starting at x and following the path \mathcal{P}_p , what is the probability of it ending at y

$$s(x, y) = \sum_{p \in \mathcal{P}_p} \text{Prob}(p)$$




So, hope you understood what is meta path. Now, we define called path count. Path count it basically indicates the number of path instances p which begin at x and ends at y ok. So, I say

that I am interested in path like APA. So, I can easily count the number of such paths in the original graph which starts from an author and it ends with an author and at the middle we have a paper. We can count it, random walk.

So, we can define all bunch of things on the on now on the heterogeneous network using this meta path concept. Random walk you can say that ok I want to start from node x, I want to follow a particular type of meta path P I right. So, what is the probability that I start from x, I reach y following the meta path P I. So, the so in this way you can also measure the similarity. Similarity between two nodes x and y using random walk is basically the probability of existence of a path p of type P I, we take the sum.

(Refer Slide Time: 29:25)

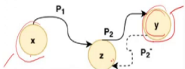



Object Similarity via Meta-Path

□ **Pairwise Random Walk:** For a concatenated meta-path $\mathcal{P} = (\mathcal{P}_1, \mathcal{P}_2)$ with instances starting at x and y , if we reverse the second sub-path to have two sets of random walkers starting at x and y and reaching a mid-point z , it forms a valid instance as $(x \rightarrow z \leftarrow y)$. Here, the similarity score is given by

$$s(x, y|z) = \sum_{p_1 \in \mathcal{P}_1, p_2 \in \mathcal{P}_2} \text{Prob}(p_1) \cdot \text{Prob}(p_2)$$


here p_2^- is the **reverse path instance** of the path p_2

Pairwise random walk: pairwise random walk the idea is same as random walk, but here the idea is that you basically start from two different nodes. So, again similarity between x and y given a particular intermediate node z . So, what is the probability? That I follow a particular meta path P_1 from x and I reach z following P_1 and what is the probability that I follow P_2 as start from y and reach z .

In other words, what is the probability that I start from x follow a meta path P_1 reach z follow a meta path P_2 inverse and reach y ? What is the comp the this combined probability? This is called pairwise random walk ok.


(Refer Slide Time: 30:27)



PathSim: Similarity in Terms of “Peers”


Path count and Random walk (RW)

- Favor highly visible objects (objects with large degrees)



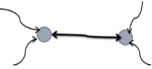
Pairwise random walk (PRW)


- Favor pure objects (objects with highly skewed scatterness in their in-links or out-links)



PathSim


- Favor “peers” (objects with similar visibility and strong connectivity under the given meta path)





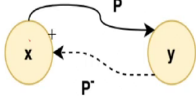
So, path count random walk you can define right in I mean in this particular aspect. Now we define something called PathSim.


(Refer Slide Time: 30:46)



PathSim: Formulation

- A measure of similarity search scoring and ranking in heterogeneous information networks
- Use the notion of meta-paths for the formulation
- A meta-path of the form $\mathcal{P} = (P, P^{-1})$ where the starting and ending object is the same, is termed as a round-trip meta-path. By default, it is always symmetric.






So, what is PathSim? So, PathSim is a measure of similarity search scoring and ranking in a heterogeneous information network. It basically uses the notion of meta path ok. A meta path of the form as I already mentioned meta path of the form $A \mid A^{-1}$. So, say I move from x to y with P and then I move from y to x with the meta path P^{-1} right and this is

called round trip meta path ok. Because you follow the same meta path and you move from x to y and then y to x ok.

(Refer Slide Time: 31:25)



PathSim: Formulation

□ A meta-path based symmetric similarity measure, PathSim, between two objects x and y of the same type can be given as follows:


$$s(x, y) = \frac{2 * |\{p_{x \rightarrow y} | p_{x \rightarrow y} \in \mathcal{P}\}|}{|\{p_{x \rightarrow x} | p_{x \rightarrow x} \in \mathcal{P}\}| + |\{p_{y \rightarrow y} | p_{y \rightarrow y} \in \mathcal{P}\}|}$$

here $p_{x \rightarrow y}$ is path instance between x and y , and $p_{x \rightarrow x}$ and $p_{y \rightarrow y}$ are roundtrip path instances

□ The salient features of PathSim

- Symmetric: $s(x, y) = s(y, x)$
- Normalized: $s(x, y) \in [0, 1]$
- Self-Maximized: $s(x, x) = 1$

(Handwritten red annotations on the slide include a diagram of a roundtrip path, a circled '2' in the numerator, and a circled '2' in the denominator.)



So, now, PathSim between two nodes x and y ok is defined by the you know total number of round trip paths with respect to x , total number of round trip paths with respect to x , total number of round trip paths with respect to y , and how many paths are there from x to y , remember these paths are all meta paths ok. So, we look at we will we say these are two nodes.


We look at all types of meta paths and let us fix a particular meta path ok. So, we basically want to understand how many meta paths how many round trip paths are there round trip meta paths are there for x because we already defined what is what do you mean by a meta path P . I follow P and I come back x right, it may move from x to z x to p does not matter.

So, how many sides round trip paths are there which actually start from x and follow the meta path and again come back to x ? How many round trip paths are there for a y starting from y moving through some nodes and again come back to y ? And how many such meta paths actually you know how I mean how many such round trip paths cover both x and y ok?.

And y 2 because when you count this one you are basically double counting things because for y let say there is a there is an edge like this round trip edge round trip meta path like this. So, for x also you are counting one times for y also you are counting one times; you are

double counting. Therefore, the denominator is divided by 2 ok because in the numerator you are counting one time ok.

(Refer Slide Time: 33:44)



PathSim: Illustration


The table below depicts the venue based publication frequency of some authors. To find the author most similar to Mike

Author	MOD	VLDB	ICDE	KDD
Mike	2	1	0	0
Jim	50	20	0	0
Mary	2	0	1	0
Bob	2	1	0	0
Ann	0	0	1	1

S. (M, J)

$$\frac{2 \times 0 + 0 \times 20}{(2 \times 2 + 1 \times 1)} + \frac{(2 \times 2) + (1 \times 1)}{(2 \times 2 + 1 \times 1) + (2 \times 2)}$$

Handwritten notes: Mike → MoD, VLDB, ICDE, KDD; Jim → VLDB, ICDE, KDD; Mary → MOD, ICDE, KDD; Bob → MOD, VLDB, ICDE, KDD; Ann → ICDE, KDD.



So, this is the formulation now let us look at an example ok. So, let us say we are interested in this type of meta paths author, venue, author. We have these authors Mike, Jim, Mary, Bob and Ann. And these authors published papers in SIGMOD, VLDB, ICDE see KDD, these are different conferences ok. Say Mike published 2 papers in SIGMOD, 1 papers 1 paper in VLDB, 0 paper in ICDE and so on.

So, the task is that given Mike as an author, who are the similar authors of Mike? Find the authors who are similar to Mike in terms of publications? We are only interested in author venue, author meta path ok. So, if you look at the visibility of think of the graph right. So, in the heterogeneous graph you have one entry called Mike, you have another entry called Jim, you have another entry called Mary right, you have mod is also bipartite network right VLDB, ICDE something like that and there are 2 papers. So, Mike published 2 papers in SIGMOD meaning that you can think of these 2 edges, right.


So, how many; so, how many round trip paths? Remember in the denominator we need to count this one right. So, when I measure the similarity between Mike and Jim, I need to count the round trip paths for Mike and the round trip paths for Jim. How many time round trip paths are there with respect to SIGMOD for Mike, this and this one this and this two this and this three right and this and this four ok. So, right.

So, how many such round trip paths are there? 2 times 2 with respect to MOD with respect to VLBD, 1 times 1 ICDE, 0 plus 0 times 0, and KDD 0 times 0. So, we got the denominator. So, this is 2 times 2 plus 1 times 1. Now, we are interested in similarity between Mike and Jim. What about Jim? 50 times 50 plus 20 times 20 ok. This is denominator, how many pairs are how many such paths are there? We start from Mike and end at Jim right.

And follow this follow SIGMOD for example, think about it. And so 50 is difficult to explain. Let us take let us assume that right let us assume that you have 2 edges here and there is 1 edge here. So, Jim has published only 1 paper in SIGMOD and Mike has published 2 papers, how many round trip paths are there? Round this is not round trip paths starts from Mike and ends at Jim.

So, you follow this path and you come here. And second time and another path you follow this and come here how many? 2 times 1. So, here in this case 5 times 50 sorry 2 times 50, 2 times 50 plus 2 1 times 20 right plus 0 plus 0. So, this is the similarity between Mike and Jim ok.

(Refer Slide Time: 37:53)



PathSim: Illustration

Similarity scores in terms of V_p and C_p are as follows


$$s(\text{Mike}, \text{Jim}) = \frac{2 \times 120}{5 + 2900} = 0.0826$$

$$s(\text{Mike}, \text{Mary}) = \frac{2 \times 4}{5 + 5} = 0.8$$

$$s(\text{Mike}, \text{Bob}) = \frac{2 \times 5}{5 + 5} = 1.0$$

$$s(\text{Mike}, \text{Ann}) = \frac{2 \times 0}{5 + 5} = 0.0$$

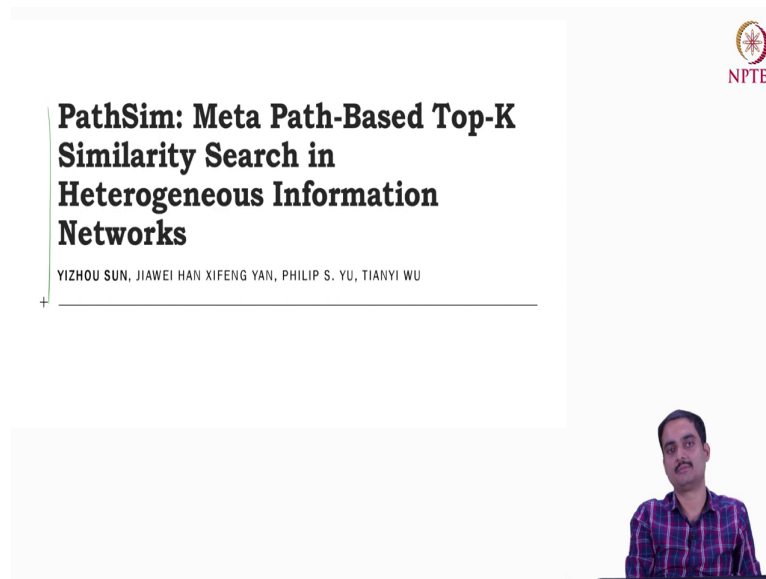
+



This is the similarity between Mike and Jim. Similarly, say Bob do the same thing. Then denominator would be if in case of Mike and Bob 2 times 2 plus 1 times 1 plus 2 times 2 plus 1 times 1. And numerator would be 2 times 2 plus 1 times 1, 2 times 2, 1 times 1 ok. So, if you do this calculations what you will get? You will see that the highest similarity is between Mike and Bob which is 1 ok.

Let us see; let us see, what about Mike and Jim? Look at this Mike and Bob, Mike and Bob the rows are exactly same. So, they should be they should be similar, but if you look at Mike and Jim rows are kind of identical, but the quantities are very very high. So, they should not be same. So, this denominator penalizes the fact that one node publishes one author publishes more papers and other publishes less paper ok. Therefore, between Mike and Jim you see that the similarity is pretty less 0.082 ok, right. So, this is the idea behind PathSim.

(Refer Slide Time: 39:46)



So, if you want to read about PathSim this is the paper that you definitely go through. So, this ends the this chapter on link analysis we have learned many things, we have started off by connecting you know network science path related analysis with social science theory, then we looked at simple measures like, then we looked at we also looked at balance and status theory, we then moved into the matrix like PageRank and DivRank.

Then we looked at Sim SimRank, PathSim and how you can use meta path heterogeneous information in the node analysis, particularly for node importance calculation and node simulating measurements ok. I hope you understood this chapter. And, in the next chapter we will discuss community analysis and we will understand how this edges play an important role in community detection.

And the next chapter we will discuss link prediction, there you will see how exactly you know all these methods that we discussed you know play an important role for recommendation for link prediction, ok.

Thank you.