Artificial Intelligence for Economics

Prof. Dripto Bakshi

Humanities and Social Sciences

Indian Institute of Technology Kharagpur

Week – 01

Lecture - 01

Lecture 01 : Network Data - Some Stories !!

Welcome to the first lecture of Artificial Intelligence for Economics. In this segment, in this lecture and the few lectures coming up, I will try to introduce to you a few different topics which will get you warmed up, hopefully. So in the first lecture, what I have planned is, I will try to give you a few examples of data which we see all around us. from politics to finance, especially network data. That's what we'll deal with in this particular lecture. And we'll try to see how we can interpret that and what stories they reveal.

Let's move on. First, let's start with history. Let's start with marriage alliances. We know that in history, Marriage alliances have been a very common strategy when it came to forming political liaisons.

Marriage was a key tool for political alliances or marriage between royal families was a very common tool. occurrence and it played an important role when it came to power sharing. So let's roll back the clock and let's go back to Florence. In fact 14th century, 15th century Florence. Well these were the most influential families of Florence back then.

Now and this is the Florentine marriage network. So, consider any two families let us say the Salviati and the Medicis. So, an edge existing between them means that one member of the Salviati family has been married to somebody in the Medicis family. So, if two families are connected via marriage then there exists an edge between them in this network, where the families are represented as vertices or nodes. Now, can we take a look at this network? and guess something about the power structure of Florence or the power distribution of Florence? Can we guess which family or which families were ruling Florence or were the most powerful in Florence? Before we do that, by the way it turns out that it was the Medicis.

I don't know, you can pause, whether you can guess whether it's the medicines by merely looking at the network. Maybe you can. But why the Medicis? Is there a mathematical foundation which tells us by looking at the network, by interpreting the network, that the Medicis will turn out to be the most important families? Most important family? Before we get into the math, let's get into the history first. You can watch this Netflix show if you wish. A few words on the Medici's, the Medici dynasty or family, they went to banking.

It was founded by Cosimo de' Medici in the late 14th century or early 15th century. The Medici's became extremely popular, they almost occupied many of the important positions in the assembly. to the extent that Catherine became the Queen of France in 1547. So they were that powerful. The Medicis also played an important role in patronizing all these Renaissance artists, Michelangelo, Raphael, Leonardo da Vinci.

Now let's formalize, let's try to understand, let's try to look at the network History tells us that yes the medicis were the most powerful, but can we simply look at the network and mathematically infer that the medicis will, medicis are powerful? Can we make the history and the math coincide? Let's try to see, let's try to formally understand if we can do so, let's formalize. Before we get back to the Florentine marriage network, let me define a few things and then we'll get back. Consider this network, a very simple one. A few definitions. In a network, two nodes i and j are called neighbors if there exists an edge between them.

For example, 1 and 3 are neighbors, 4 and 5 are neighbors. Nodes i and j are connected if there exists a path between i and j, not necessarily neighbors. For example, 1 and 4 are not neighbors, but 1 and 4 are connected because there exists a path between 1 and 4. The shortest path between node i and j is the shortest route or the number of hops. So it's the minimum number of hops required to read j from i.

Degree of a node is the number of neighbors a node has. For example, the degree of 4 will be 2. The degree of 5 will be 4. So on and so forth. Sorry, degree of 5 is 3.

Degree of 4 is 2. Degree of 3 is 3 again. Okay? Great. Now that we know this, let's define a particular metric called betweenness centrality. So now we will define two measures or two metrics if you may call them, which in a way depicts the importance of any particular node in a network.

So what is betweenness centrality? Let's understand. So if I have two nodes, i and j, any two nodes, I define $t_{i,j}$ as the number of shortest paths between i and j. Number of shortest paths. For example, between 1 and 4, what is the shortest path? It's 1, 3, 4. But I have another path, 1, 2, 3, 4.

But 1, 3, 4 happens to be the shortest path. So there is only one shortest path between 1 and 4 in this case. $p_{k,i,j}$ is let's say the number of times a node k lies in the shortest path connecting i and j, okay. For example in the shortest path connecting 1 and 4 there is only one shortest path and 3 appears in that path, so $p_{3,1,4}$ is going to be 1, so p_3, so if you look at 3 here, so 1, 4 will be 1 and $p_{3,1,4}$ will also be 1, right. Anyway, between the centrality of a node k is the number of times k features in the shortest paths between any two nodes in the network.

okay so this tells you that how many times if any two nodes have to connect to each other how many times they'll have to connect via k okay so between s means between i and j how many times between how many i's and j's k features Let's formally define this now. Let's define this set S_k. This is the set of all pairs of connected nodes. Set of all pairs of connected nodes other than k. So, what is S_k? So, S_k is pairs of all connected nodes other than k.

Between this centrality is defined in the following manner. I take $p_{k,i,j}$ that is how many times k features in the shortest paths between i and j for all i, j belonging to S_k and I divide $p_{k,i,j}$ by $p_{i,j}$ so the numerator $\frac{p_{k,i,j}}{p_{i,j}}$ is the number of times k features in the shortest path divided by the total number of shortest paths. and that divided by the cardinality of S_k. So, this is the betweenness centrality of k, that is the betweenness centrality of k, great. Let us compute the betweenness centrality of the different nodes in this particular Let us consider 1 and 3.

Are 1 and 3 connected? Yes, they are. What about $p_{4,1,3}$? I want to compute the betweenness centrality of node 4 now, let us say. So, I will take all pairs of nodes other than 4. So, that is my set S_k.

So, I take 1 and 3. 1 and 3 are connected. So, $p_{1,3}$ is 1. And not only connected, how many shortest paths are there between 1 and 3? It's only 1 because they are directly connected, there exists an edge. So P is 1. Does 4 feature in that shortest path? Answer is no.

So P is 0. What about 1 and 2? 1 and 2 are neighbors again. So $p_{1,2}$ is 1, so number of shortest paths is 1 because they are direct neighbors. What about $p_{4,1,2}$? Does 4 feature in the shortest path between 1 and 2? Answer is no, absolutely not. Similarly, I can find out for all other pairs p, so I can find out $p_{i,j,4}$ and $p_{i,j}$ for all other nodes $i, j, i \neq j \neq 4$, that is what I have done in this slide and once I do that I can find the summation which is this, which is what we have seen in a few slides before this, okay and if we compute that we get that the betweenness centrality of 4 is 9 by We can proceed similarly for 3, it turns out that the betweenness centrality of 3 is 8 by 15 and proceeding for all of them it turns out that the betweenness centrality of 4 is the highest.

So, in this network 4 happens to be the most important network by if we consider betweenness centralities. followed by the betweenness centralities of 3 and 5, followed by 6, 7, 1 and 2. Now let's introduce another measure of importance in a network. Another measure which depicts the importance of a particular node in a network.

It's called the CADS prestige. So let's understand what is CADS prestige. The power of a node comes from connecting to a powerful node and the powerful node derives its power from connecting to other powerful nodes and so on and so forth. So, it means let us say I have a node i and n i is the set of all neighbors of i. So what is the prestige of node i, what is the cat's prestige of node i or player i or family i, whatever you might call it. The cat's prestige of this vertex i is given by the cat's prestige of its neighbors divided by their degrees.

So i scan through the set of all neighbors of i. compute their, see what their prestige is divided by their degree and add them up. Why divided by a degree? What's the rationale for that? So let's say if you and I are connected and you are extremely powerful and if you are also connected to other people then your influence gets dissipated amongst others. So the fraction or share of the power which I derive by being associated by, with you gets diminished, it is inversely proportional to the number of other associates you have got, okay.

Great, so this is Katz prestige. So, for this network, so yeah this is what I was mentioning, the prestige of node I depends on both the prestige of its neighbors and the degree of the neighbors as well, the time and resources that the prestigious node can share to an individual node reduces with increase in its degree. Okay, now let's try to compute the cat's prestiges of this particular network. Let's see. Let's first normalize.

Let's say prestige of 1 is 1. Let's start with that. Then what is the cat's prestige of 2? Well, it is 2 has 2 neighbors, right? 1 and 3. So the cat's prestige of 2 is going to be cat's prestige of 1 divided by the degree of 1. Degree of 1 is 2. plus Katz prestige of 3 divided by the degree of 3, degree of 3 is 1, 2, 3.

So, this is 3. What about Katz prestige of 3? 3 has 3 neighbors 1, 2 and 4. So, Katz prestige of 3 is prestige of 1 divided by the degree of 1 which is 2. plus the prestige of 2 divided by the degree of 2 which is again 2 plus the prestige of 4 divided by degree of 4, degree of 4 is again 2. Similarly we can write down the equations for all the cat's prestiges. Now we have a system of equations, we have 7 equations and 7 unknowns, we can solve them, right.

Let's solve them and this turns out to be the cat's prestiges of the different knowns. okay but here we initially considered we had to if we have a system of equations with no particular initialization that will yield to nothing right then we will have infinitely many solutions But here the initialization was p1 equal to 1. And then we wrote down our system of equations for Katz prestige and found it. But we could have started with p2 equal to 1, and then we'll get another vector of Katz prestiges. $p_3 = 1$, we'll get another vector of Katz prestiges.

So we should remove this initialization sensitivity. So just to do that, we follow the following algorithm. So, we first initialize one, compute all the catch prestiges, then initialize two, compute all the catch prestiges and finally, we take all average of all those prestiges. Great. So, we have learned two important measures, two important metrics which signify the importance of a particular node in a network.

What were they? Betweenness centrality and cat's prestige. So let's roll back to Florence and let's understand what is the betweenness centrality. Let's delve deep into the Florentine marriage network and understand what is the betweenness centrality or cat's prestige of the different families. It turns out that the Katz prestige of the Medici family is 47.

5 and that's the highest. Sorry, the betweenness centrality. The Katz prestige is also highest for the Medici family again. So the Florentine marriage network, if we look into the Florentine marriage network and compute betweenness centrality and Katz prestige, it turns out that they give us a picture that the Medici family is the most powerful family in this network. And it also turns out in history that the Medicis were indeed the most powerful in Florence. So we see how a little interpretation of the network gives us a peek into the history.

Great. So, so much so for an example from history. Now, a little bit of finance quickly. So, this is from a paper by Demiror, Debald and Liu and Yelmaz. So, let us connect it. This is about connectedness of financial institutions. So, what are they doing? I will not get into the technical details.

So let's understand. So the study basically takes in 96 banks and these are all chosen from the world's top 150 banks by assets. 82 are from developed economies and the remaining 14 are from emerging markets. And all of these banks which are chosen are globally systematically important banks, GSIBs. First, let's define something called volatility or let's call it volatility of a bank. So total volatility of a bank, they have defined it in this manner.

This is the definition which has been used. let's say σ^2_{it} is the volatility of bank i at time t, time t is period t, day t let's say, it is simply on day t the volatility of bank i, volatility means how much it's fluctuating, it's simply given by this complicated expression where H_{it} is the highest stock price, L_{it} is the lowest stock price, C_{it} is the closing stock price and O_{it} is the opening stock price of bank i in day t. Simply observe on day t or on time period t, I observe the highest lowest closing and opening stock prices of bank i and then i apply this complicated equation or expression and we get the volatility of bank i in time period t now we use something called variance decomposition we try to see how the return volatility of bank i is influenced by other banks j okay so the total volatility which is sigma i well technically it's σ^2_i this can be decomposed into this θ_{ij} so what is θ_{ij} theta i j is or as we will define it very soon θ_{ij} is the effect of effect of bank J on the volatility of bank I and we can when we talk about this effect, this effect could be an immediate effect or it could be into the future effect. For example, I can talk about θ_{ij} which basically tells you that this is the effect which bank j has on the volatility of bank i eight step forward. That is what happens in bank j, how will it affect the volatility of bank i h periods from now, h periods from now.

Great. Now we can do this by using something called VAR which is vector autoregression and a lasso regression technique. This is what the authors of the paper have done but I'm not getting into the technical details because this is an introductory lecture. I just want to get you or give you a glimpse of what's going on.

Great. So let's see. So this is firm J's contribution. to firm i's eight step ahead variance this is given by $\theta^g_{ij}(H)$ okay great and g is the network g is the network of banks but let's forget about networks now so what is $C^H_{j\to i}$ well this is $\frac{\theta^g_{ij}(H)}{\sum_{j=1}^{N}\theta^g_{ij}(H)}$. So, this is basically the proportion of the total volatility which, proportion of the total volatility of i which comes about due to the effect of j. So, this is the proportion of, I repeat once more, this is the proportion of the total volatility of i which is being brought about by tank j. And by construction of course, so that is c j to i and we have we are taking it for 8 step ahead. This h could be anything 10 day ahead, 3 day ahead whatever your interest of the empirical study is.

So, clearly this summation will by construction this summation will be equal to 1 and this summation is going to be n. If you sum this up over all j's it will simply be 1 as you can imagine. Now we define something called total directional connectedness, what is this? So if you consider any firm i or any bank i, the total directional connectedness is the total amount of effect other banks are having on bank i. on an average. So let's say there are n banks, $C^H_{j \to i}$ is the effect j has on i and if you sum it over all j's, $j \neq i$, so that is the effect all other banks are having on bank i or firm i.

So $C^H_{\cdot \to i}$ is the total directional connectedness to firm i from all other firms. Similarly, the total get directional connectedness from firm i is $C^H_{i \to \cdot} = \frac{\sum_{j=1, j \neq i}^{N} C^H_{i \to j}}{N}$. So, this is the total effect which i has on all other banks on the volatility of all other banks on an average. What is system-wide connectedness? Well, system-wide connectedness is every bank or every firm has a total directional connectedness to that firm. The system-wide connectedness is the average of those total directional connectedness to the firms.

In this case, I'm computing CH, so I've taken H, so H is the 8th step ahead system-wide connectedness. Okay? Remember, H could be anything. It could be 1, 2, whatever your interest is. In this study now, the authors have proceeded with H equal to 10. So, if we look at H equal to 10 and if we do the, if we look at the volatility connects, this is what we see.

Well, this seems like a very complicated network. We cannot seem to make a head or tail of this, but I think we can interpret this table a little more carefully now. Look at this. So, what story comes out from this? Let's look at Africa first, Africa is getting influenced by who the most? Well 45 is the total influence let's say, it's the total volatility of all African banks, so I'm aggregating over continents now, so if this is the total volatility of the African banks, it is coming from where? The total directional connectedness or the total influence Africa is being brought about by Europe and North America the most. What about Asia? Asia again is getting influenced by who? Again Europe and America the most.

The remaining influences are not much see 30, 21, 4, nothing. What about Europe? Europe is getting affected by who the most? North America the most. 581 is the total volatility, 431 of that, so that's a huge proportion is coming from North America.

Similarly, when it comes to North America, who is influencing North America the most? Europe. So, it seems North America and European banks are ruling the world.

They are impacting the volatility of banks all over the world. What about Asian banks? Well the total volatility or the total input into Asia is 480, 418 that is the amount of total directional connectedness into Asia, to Asia. What about the total direct directional connectedness from Asia? It is 214, 214 see. So, which means Asia is getting influenced more. rather than influencing more right that is the picture we loosely get. So, two key takeaways from this from the study we see that in this network of banks North America and Europe are large and they are transmitters of future volatility uncertainty to the rest of the world.

Okay, so all everybody is connected, the banks are connected to each other. North America and European banks affect the volatility of other banks globally to a much larger extent and they also form a cluster amongst themselves. They are also connected to each other massively. Remember, Europe is affected by North America massively and Europe is, North America is affected by Europe massively. So, they are extremely interdependent and they also have tremendous volatility spillovers globally in other zones.

Asia on the other hand has noticeably large total directional connectedness into, more into than from Asia. So, it is a net receiver of volatility, if you may think about it that way. So, this was a regional thing. let's look at time wise, if we compute this 8 step ahead volatility spillovers, the total directional connectedness, it turns out that in September 1, 2008 that's before the Lehman crisis, this is how the connectedness scenario looks like, whereas in this situation which is post November 21 after Lehman went bankrupt, The connectedness is all the more, which means that the banks were failing together now. They were impacting each other much more after the crisis than before the crisis.

In fact, it turns out if you compute the system-wide connectedness of the bank volatilities, remember system-wide connectedness? This is what it is. This is system-wide connectedness. so if you compute system-wide connectedness of all the banks globally it turned out that the system-wide connectedness went up and it peaked during the Lehman crisis and then again it started coming down which means that when there is a global crisis the volatilities of the banks move together which makes the crisis all the more grave And that's what we saw during the Lehman crisis. When Lehman went bankrupt, every other bank, there was a spillover effect and every other bank started falling down and we slipped into a recession and a complete economic meltdown.

Thank you. I think I have given you a starter to look into data. First was interpreting a network of marriages from 15th century Florence and the second one I talked a little bit about connectedness of financial institutions or banks. I hope this got you started. In the next lecture I will talk about something completely different. I will talk about