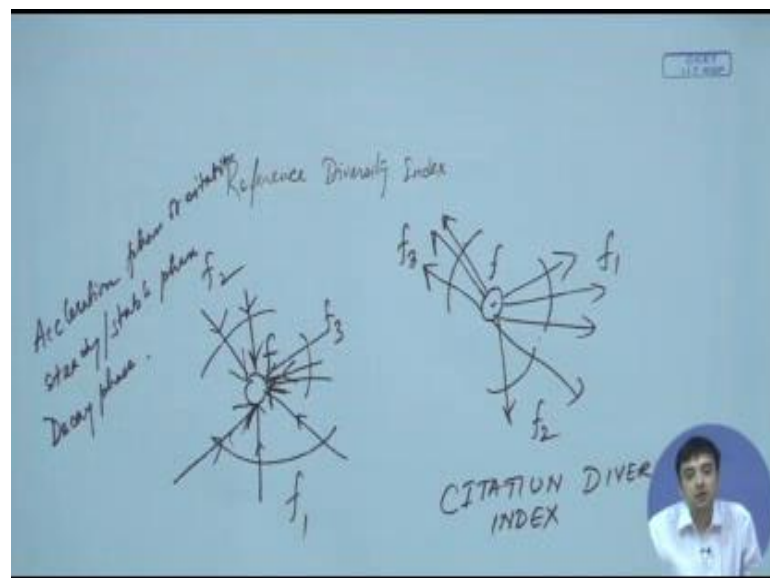


Complex Network: Theory and Application
Prof. Animesh Mukherjee
Department of Computer Science and Engineering
Indian Institute of Technology, Kharagpur

Lecture – 19
Citation Analysis – II

Last day we looked into Citation Analysis in general, and we started off with this idea of quantifying interdisciplinarity in computer sciences. And we already saw this measure of Reference Diversity Index.

(Refer Slide Time: 00:38)



So, using this measure we try to see if there is a field f what is the proportion of references that this field f is giving to different other fields. And if this proportion is very diverse in terms of the entropy value then we assume that particular field f is more interdisciplinarily, than some field which puts all its references to one or two other single file. So that is one type of a notion for quantifying interdisciplinarity.

Now, since we know that the citation network is a deduced graph. If there are references which is like outgoing edges, so you are assuming that there is a field f and there are outgoing edges like this which we were dividing into proportions. If this is the field f for which we were trying to quantifying interdisciplinarily, we were saying like how many of its outgoing edges go to some field f_1 some another field f_2 , a third field f_3 and so on. And in this way, based on these proportions we try to quantify the reference diversity

index.

Similarly, if you consider another field the picture of the same field f , but now receiving citations from different other fields. So what I am trying to say is that these two factors are like the two sides of a coin. So, one is where you are looking at the outgoing edges basically the references going out of f , the references that are made by the papers in f to all other different fields. Whereas, here you see what is the citations that the papers of field f receive from all other different fields, so may be from field f_1 it receives 3 citations from field f_2 it receives 2 citations and then there is another field f_3 from which it again receives 3 citations.

So, this is also another way which one could use to quantify the notion of interdisciplinarily. That is, if you are getting citations from various other fields then there is a higher chance that you are interdisciplinarily field, where as if you are getting most of your citations from one particular field or (Refer Time: 03:19) two fields then there is a chance that you are not so interdisciplinarily in nature. This can be again quantified by something called the Citation Diversity Index. So here, what you do you find out the proportion of citations that are coming from the different field, the fraction of citations that are coming from the different fields to the field f . And once you have derived all this proportions you can again find out the entropy of these fractions and that will tell you how diverse its citation is.

But, then there is one difference between the reference diversity index and the citation diversity index. What is this difference? Note that the outgoing edges once a paper is written whatever references you put in are fixed forever. So, once the paper is written and it is published the set of references does not change, that means the number of outgoing edges from a particular paper once the paper is published do not change. This remains fixed over time

So, the reference set here of f does not change over time, in this particular figure the reference set does not change over by definition. Whereas, citation on the other hand might increase over time, that is like f might get 10 citations from different fields in the first year, another 10 in the next year, another 10 in the third year and so on and so forth. In this way the volume of citations actually increases over time for f whereas, the volume of references going from f remains same from a particular paper.

So, if you consider one particular paper in the field f the reference set of the paper is fixed and it does not change. Whereas, the citation set for this paper might increase over time, this paper might be more and more cited over time. This is the prime difference. So, for this citation diversity index we not only look at these raw values, but we look at the temporal differences of the citation diversity index. We compute this citation diversity index say at the current year y then again after 2 years we compute the citation diversity and then we find the difference.

Now you can keep a window of 1 year, 2 year, 3 year, 4 year and so on and so forth. So every year you might calculate the citation diversity index and find out the difference. This here the time window assumed is 1 year. Now, if you try to identify the differences after 2 years the time window will be 2 years and so on and so forth. For our analysis we assume that the time window is set to 1 year, that is we are trying to see what is the increase or decrease in citation diversity over a period of time taking 1 year gap.

(Refer Slide Time: 06:31)

CITATION DIVERSITY INDEX

- CDI of a paper X_i at time t_i

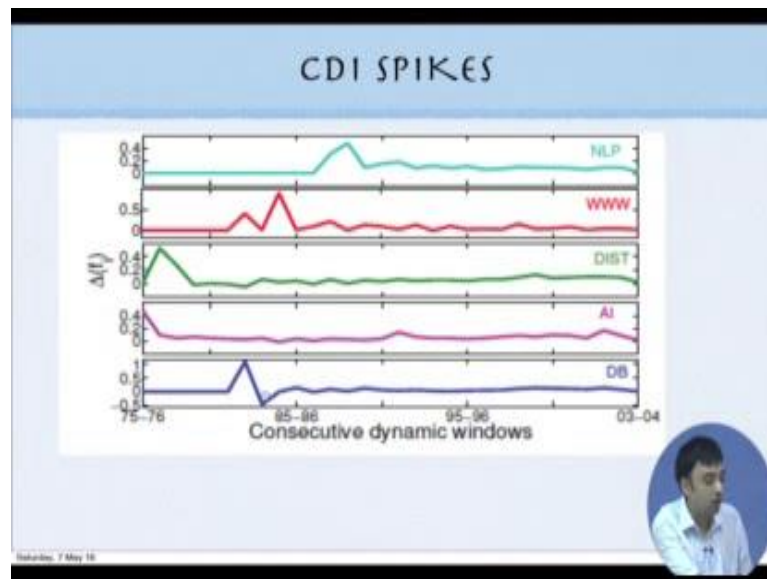
$$CDI_{t_i}(X_i) = - \sum_j p_j \log p_j$$

where, p_j is the proportion of citations received by X_i from the papers of field f_j
- Drift of CDI between two successive time windows
$$\Delta_{t_i}(f_i) = CDI_{t_{i+1}}(f_i) - CDI_{t_i}(f_i)$$

So, we have already discussed how to define citation diversity it is again the proportion of papers from different fields. Citing the papers in a particular field $\log p_j$, so sum over all such proportions some over all $p_j \log p_j$ where j is a particular field. Now this actually defines the entropy the notion is exactly similar to the reference diversity index. Now since this is a time warring quantity what we do is we find out difference in CDI values over consecutive time periods. Here in this case for our analysis we have kept the

time period equal to 1.

(Refer Slide Time: 07:14)



Now, if you try to plot this CDI differences you see certain interesting patterns as you see here; so what happens is the difference initially remains low, then there is a point where the difference spikes up and then again it stabilizes. And you will see it for all the different fields mostly those which are assumed to be in general interdisciplinarily like that the language processing, which is like a mixer of knowledge from linguistics, knowledge from information retrieval, knowledge from algorithmic techniques, knowledge from machinery learning so all this actually develops the field of natural language processing. So that is really a truly interdisciplinarily field.

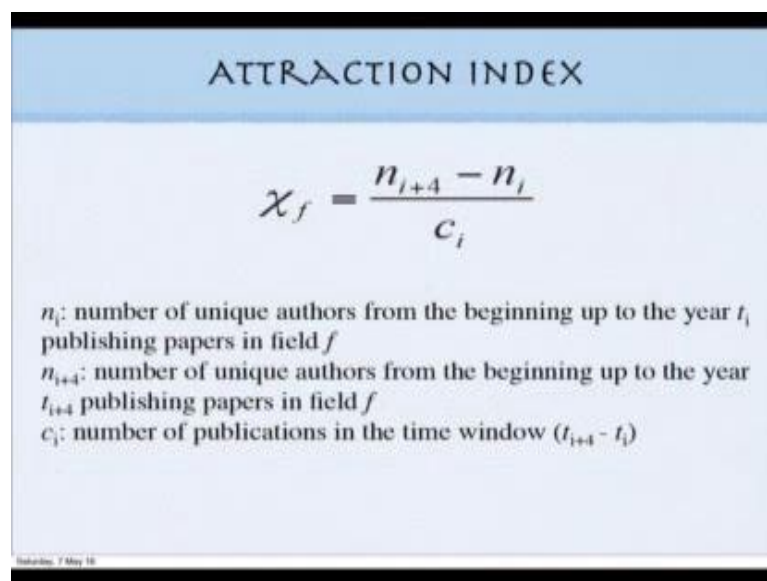
Similarly, the World Wide Web which actually borrows concepts form graph theory, security distributors system, and a lot of other areas. Now, for each of this what you see is once you move across the time line since as we say that the CDI is the time waning quantity, so as you move across the time line each of them show a spike at certain point. Now what can we interpreted from this. You see that initially the citation diversity index is low at one point it rises therefore the differences spikes up. At this point the citation diversity suddenly grows, that is the field has suddenly become highly interdisciplinarily. So, at this point the different between these CDI values at the previous point and the next point is high so the different spikes up, the value spikes up.

Now then in the immediate next time point the difference again falls down which means

that whatever CDI was achieved at this point is maintained that is why the difference again falls down. So here say, the CDI value was some number x, here the CDI value becomes some number y so the difference of y minus x is very very high that is why there is a spike. Now at this point probably the CDI value become z and the difference z minus y is again small because z is as high as y, so the high CDI is maintained. Since, the high CDI is maintained you see that there is a drop in the difference, so the difference becomes again low. As the two consecutive time points the CDI values are high.

Here between these two consecutive time points this was the lowest CDI this was the highest CDI that is why there was a spike. Now this spike falls because after that point roughly the same CDI value is maintained and therefore the rest of the differences as you see are the stable curve. So, similar observation holds from for all of the different fields.

(Refer Slide Time: 10:02)



ATTRACTION INDEX

$$\chi_f = \frac{n_{t+4} - n_t}{c_t}$$

n_t : number of unique authors from the beginning up to the year t publishing papers in field f
 n_{t+4} : number of unique authors from the beginning up to the year $t+4$ publishing papers in field f
 c_t : number of publications in the time window $(t+4 - t)$

Innovation, 7 May 18

Now, in this way you can define various matrixes actually to identify or to quantify the notion of similarity. The next measure that we will get introduced to is the number of new authors coming into a particular field. We have already said that interdisciplinarily is the talk of the day. So people are driven more and more to towards interdisciplinarily research. New researchers who are entering in the field are trying to be more and more interdisciplinarily in nature and if that is so then there will be more number of new authors who are writing interdisciplinarily papers.

So, what we try to do is that we try to see that in very year what is the total number of

new author that are entering into the system of writing papers. And we see in what fields they actually join much. So, we see like whether most of the new authors who have joined the computer science field, where they most of them have natural language processing, World Wide Web, or operating systems, that is the questions that we are trying to ask. The bunch of new authors who start doing research and start writing papers in computer sciences we try to ask out of this what proportion are writing papers in known interdisciplinary areas, whereas what proportion are writing papers in the core areas of computer science.

(Refer Slide Time: 11:40)



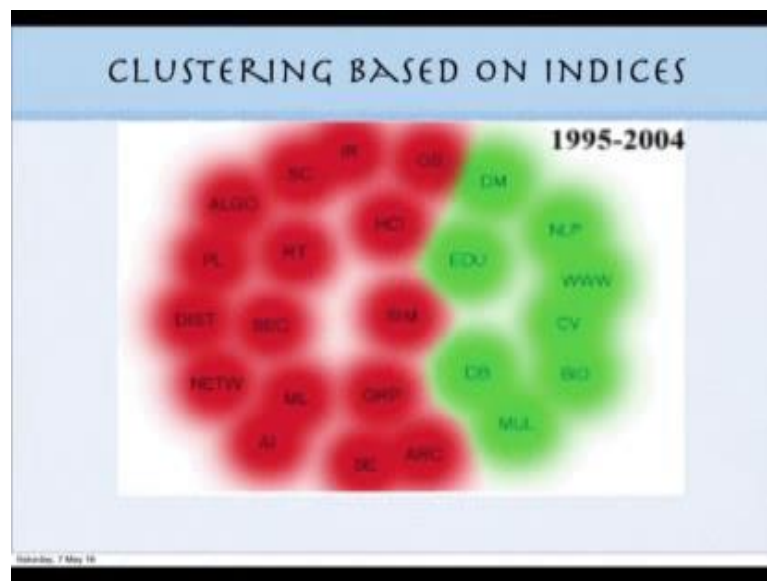
So, if you try to look at this proportion you see again some interesting patterns what do you see is that in the earlier years of computer science like in 1975 to 79 most of the new authors who joined the field, who joined this some domino of computer science, joined in areas of operating systems, distributive systems, security, etcetera. However, as time progresses more and more interdisciplinary areas come at the top rank. These fields here are ranked by the proportion of new authors that are joining that field. In computer science what is the proportion of new authors that are joining a particular field that actually is used as a quantity to decide this ranking. As you see initial years mostly core fields like operating system, distributed systems, security, etcetera, or net fox are the areas in which actually more and more new authors have joined.

However, as time progresses in the year of 2004 to 2008 what you see is that mostly

interdisciplinarily areas like bio informatics, World Wide Web, data mining, natural language processing, these are the areas which are known to be more interdisciplinarily in nature. There is more and more fraction of new authors now this author joining these areas, so that is again an indication of interdisciplinarily. So, now we use all these factors; the reference diversity index, differentiate in the citation diversity index, the attraction index which is the total number of new authors joining a particular field in a year so all these factors we use to actually identify a particular field.

So, each field actually is represented by a vector of numbers. Now these vector consist of; the reference diversity index, the delta citation diversity index, and the attraction index. These three factors actually define each particular field. Now based on this vector space, now each field can be represented by these three numbers which is like a vector.

(Refer Slide Time: 13:56)



Now if you represent all the 24 fields that we have of computer science in the vector space and do a class studying you see an interesting structure emerging like this. The red bubbles here in this picture are the ones which are known to be core fields of computer science like for instance; algorithms, operating systems, architecture these are more and more the core areas. These are known to be core areas of computer science. Whereas, the green bubbles here are known to be more interdisciplinarily like; World Wide Web, natural language processing, computer vision, bio informatics, and fields like that.

So, fields which actually incorporate ideas from various different domains by informatics

actually incorporate ideas from biology from different branches of computer science in fact data mining algorithms, graph theory and so on and so forth. These fields are truly interdisciplinary. And as we see that given these three factors, reference diversity index, ideal difference of the citation diversity index, and the attraction index given these three quantities representing each field as a vector nicely differentiates the core fields from the interdisciplinary fields of computer science. So that is an unsupervised method by which we can identify those fields which are interdisciplinary in computer science compared to those which are the more known to be core fields.

(Refer Slide Time: 15:29)



So, there is another interesting thing that happens so look at this figure. What I show here is basically the evolution of the group of two different fields. There are two fields that we consider; in the top panel we consider the field of World Wide Web and in the bottom panel we consider the field of programming languages. And the bubbles here show the fraction of citations that are going from the World Wide Web to different other fields. So, what we see in 75 to 84 in this particular decade what happens is that the most of the citations from the World Wide Web papers; actually go to the data bases, most of the references from the World Wide Web actually go to the data bases.

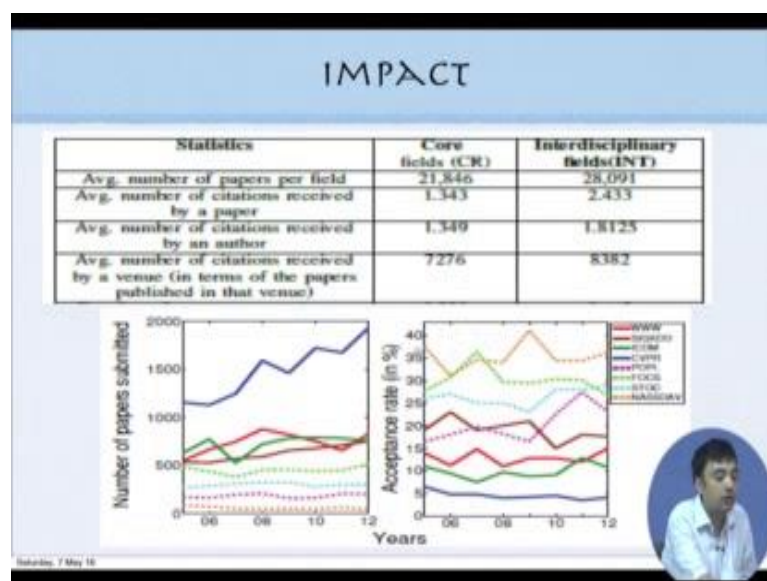
In the next decade what you see is its not only data basis, but also information retrieval where many of the citations from World Wide Web go. However, in the last decade what you see is that while there is already some fraction of citations that is going to data basis,

data mining, etcetera, there is a large fraction of citations that is going to the World Wide Web itself. This is the point where we see that World Wide Web itself has emerged to be a big field.

Since, 1975 to 1994 what we observe is that World Wide Web is probably still a small field in its infancy most of the papers that are written in this field they do their references, they give most of their references to other fields like, data bases or data mining etcetera. Whereas, in the last decade what we see that World Wide Web itself has grown into an entity and many of these references are held within World Wide Web itself, many of the references do not go outside World Wide Web.

Now, the fractions of citation that World Wide Web was providing to the other fields have reduced, while the fraction of citations of World Wide Web to itself has increased. So, that is the typical evolution phenomena of an interdisciplinarily field. Whereas, if you look at the core field which is programming language. Programming language is more like a core field it has existed in the computer science domain for quite long time and there is hardly any such evolution that is observed there. You see it is more or less stable it is fractional citation to different fields is more or less stable over time mostly coming to the programming language field itself. So that is happening for all the three decades there. This actually shows a striking difference between the evolutions of the interdisciplinarily field compared to the evolution of a core field of computer science.

(Refer Slide Time: 18:24)



So, the next slide actually tries to motivate you why one should go for interdisciplinarily research in computer science. All this time I have been stressing on the point that interdisciplinarily is the talk of the day, one should actually try to do research in interdisciplinarily and there is a lot of fame and name in doing interdisciplinarily research. This particular slide actually gives you some indication of that fact. So what you see here is, in the top table you see some average statistics, so average number of papers written per field. If you look at the core field of computer science, so the core fields are the ones that are marked in red blubs here and the interdisciplinarily fields are the wants that are marked in green blubs here.

So, you see that the total number of papers or roughly the average number of papers written in an interdisciplinarily field is much higher than those written in the core field of computer science. Further, average number of citation received by any paper in the interdisciplinarily field is almost double that in the core field. So, the probability that your papers will get highly cited, that you will become highly reputed, your papers will be highly visible, your paper will be cited by other people that probability actually gets enhanced or almost gets doubled from whatever we see here it gets almost doubled if you are working in interdisciplinarily field compared to a core field.

Similarly, the average number citations received by an author. You can look at the average number of citation received by every paper in a field or every author in a field, so if you consider from the author prospective that also is actually higher for the interdisciplinarily authors than the core authors. And if you see the number average number of citations received by a venue in interdisciplinarily field, so interdisciplinarily venues are confidences where interdisciplinarily research is published, whereas core venues are places where core areas of computer science research is published. So, what you see the average number of citations in the interdisciplinarily venues is much higher than in the core venue.

But all these things look very (Refer Time: 20:48) that is doing inter doing research in inert disciplinarily actually favors the numbers of citations that every individual paper or an individual author actually receives. However, there is one point of caution here, doing good interdisciplinary research is actually very hard because it is growing more and more competitive over the years. So, what we see is that although it is the fact that receiving citations at the level of authors as well at the level of papers is more probable in

interdisciplinary areas. At the same time publishing it in a paper in a interdisciplinary venue is more difficult than publishing a paper in a core venue. That is what we see in the bottom figures.

So, in the bottom figure what we do is basically we short list four interdisciplinary venues and four core venues, so the four interdisciplinary venues are represented by the bold lines; World Wide Web is one of the top tier conferences in interdisciplinary research actually is dedicated to graph theory, security, social networks properties of complex networks, etcetera.

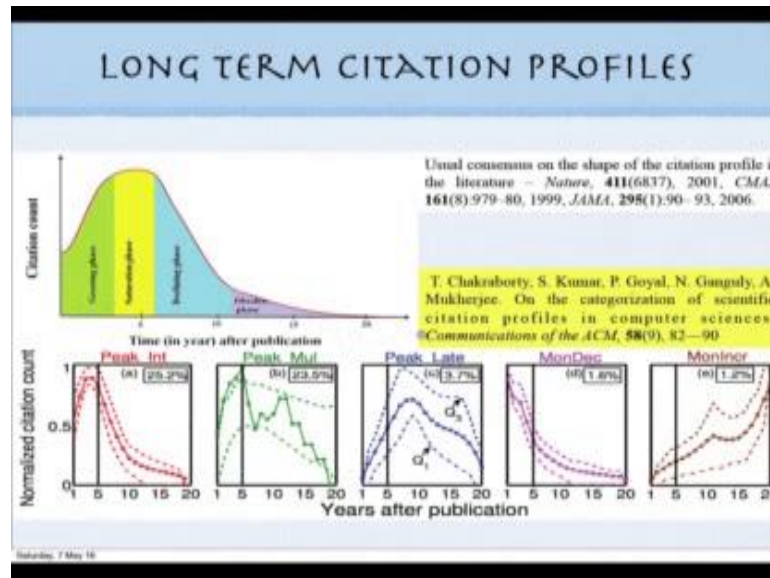
Then you have (Refer Time: 22:17) which is a data mining and data basis conference which is also known to be highly interdisciplinary, because it accepts papers again from graph theory, it accepts papers from social networks, etcetera. ICDM is again a data mining conference and it is known to be highly interdisciplinary in nature. And then there is this fourth conference which is called CVPR; Computer Vision and Pattern Recognition is also very highly reputed conference in computer vision, graphics, image processing, and better recognition. All these are like interdisciplinary areas known to be interdisciplinary venues. Whereas, the last four the broken once represented by broken lines are known to core fields like POPL, FOCS; POPL is the top tier conferences in programming language, FOCS and STOC are top tier conference in algorithms and theory of computation, and (Refer Time: 23:13) is a top tier conference in operating system.

Now if you see on the left hand side we show the number of papers that are submitted to each of this conference as you go over the years. So, what you see is that the number of papers that are submitted to the interdisciplinary areas is way higher than the number of papers that are submitted to the core areas of computer science. On the other hand, if you see the number of paper that is accepted in the core areas that are actually higher for the core areas. Whereas, it is much lower in the interdisciplinary areas.

So, basically this indicates these two figures together indicate that there is a high level of competition in the interdisciplinary areas. In interdisciplinary areas there are a large number of papers submitted, while there is only very less number of papers accepted. Whereas, in the core areas there are lesser number of paper submitted, but a reasonably high fraction of them are accepted. That means, these two figures together tells you that there is a larger level of competition in doing research in interdisciplinary areas.

So, although as I said that there is a high probability that you have citations, you gain citations by doing research in interdisciplinary area, but then publishing a paper in an interdisciplinary area is actually really harder. So, that actually gives you some idea about how to quantify interdisciplinarity and why one should be involved in more interdisciplinarity research.

(Refer Slide Time: 24:50)



The next thing that we will talk about is about the citation life cycle of a particular paper. So, what we try to show you here in this first figure on the slide is the average profile of a citation of a particular paper over the years. Initially, when a paper is published it tries to accumulate citations and it there is an acceleration in its citation and at some point in times say after 3 or 4 years the citation value actually stabilizes and then there is a steady decline, there is an exponential decay in the citation of the paper.

Basically, if you take any random paper from the computer science data set you will roughly see this particular behavior. So, what happens in this behavior, what we want to show is that initially there is a growing phase. Suppose a paper piece accepted in your particular venue and is published then from the point of publication initially there is an accelerating phase when the paper keeps on getting accumulating citations then there is a point like in 3 to 4 years there is a stabilization that happens, so the number of citations gets more or less stable and then the number of citations for the rest of time period actually is declining over time.

So, there is an acceleration phase, then there is a study or stable phase citations and then there is a decay phase. In fact, this observation was made long long back like, this observation was made as early as in 2000 when people try to study the individual citation profiles of various papers and from that derive measures like, impact factor which are based on the idea that a paper usually tends to get most of its citations in the first 3 years. After the first 3 years a paper hardly gets any more citation.

So, this 3 year time window is actually a very important factor and that actually goes into the definition of impact factor. So, impact factor if you look at the definition it tries to look into the citation history of the last 3 years and those 3 years is fixed 3 or 5 years. These 3 to 5 years is fixed based on the observation of the citation profile of different papers. Since researchers observe that most of the citations are accrued by a paper in the first 3 or to 5 years of its publication that is why the impact factor time window is also set to be 3 to 5 years.

However, since we had this huge data set we try to further reinvestigate that whether this particular phenomena is true across all papers. However, what we observe is that this scenario is not so straight forward. And we basically see that apart from the citation profile that the researchers have already observed, the one that I show here in the red in the first left hand side figure in red. This is one that is very similar to this particular figure. So, this is the one where the paper actually accrues most of citations in the first 3 to 5 years roughly and then there is a study data line.

However, apart from this particular citation profile, this particular behavior there are other at least four identifiable behaviors. So, what we see in the second one is that there are certain papers in our data sets which have two or more peaks. Here you have only a single peak, and the peak is seen in between 0 to 5 years. In this second one which we call multiple peaks or peak mul, you have more than two peaks. The third one is a late peak scenario where the peak observed nowhere in between 1 and 5 years, but much later than 5 years.

Then there is the 4th one which is like the monotonically decreasing scenario, where the citations to the paper actually only decrease over time. And as a researcher you will never want to be in this particular bucket, because you want your papers to have any how more and more citations not less over time. And then there is this 5th bucket which is a

very interesting bucket. What we see here is that the citations for the papers in this bucket actually only increase over time which we call the monotonically increasing. The citations never fall never decay, so this is actually like papers which are kind of seminal papers, here the these papers actually keep on gaining citations over time there is no decay of citations for this particular set of papers.

So, in the next part of the lecture we will see more interesting properties of these 5 different categories of papers.