Christos Faloutsos Speaks Out
on Power Laws, Fractals, the Future of Data Mining, Sabbaticals, and More

by Marianne Winslett

Christos Faloutsos
http://www.cs.cmu.edu/~christos/

Welcome to this installment of ACM SIGMOD Record's series of interviews with distinguished members of the database community. I'm Marianne Winslett, and today I have here with me Christos Faloutsos, who is a professor of computer science at Carnegie Mellon University. Christos recieved the Presidential Young Investigator Award from the National Science Foundation in 1989. He received the 1997 VLDB Ten Year Paper Award for his paper on R+ trees, and the SIGMOD 1994 Best Paper Award for a paper on fast subsequence matching in time series databases. Christos is a member of the SIGKDD Executive Committee, and he has wide-ranging interests in data mining, database performance, and spatial and multimedia databases. His PhD is from the University of Toronto. So, Christos, welcome!

Thank you very much, Marianne, for having me here.

Christos, you have been described as the master of collaborations: someone who can reach out to colleagues not only in the database area, but also in other disciplines, such as the sciences and in statistics, and to visitors from industry as well. What do you do to nurture such collaborations?

Thank you very much for the compliment. I think that the urge to collaborate is part of my personality. Some people prefer to stay in an area and do deep work there, and other people enjoy collaborations. I was extremely lucky to have wonderful colleagues from industry, statistics, and machine learning; the collaborations just developed by themselves and I didn't have to do anything special to nurture them.

How do you learn the basics of so many techniques in so many different fields, and bring them to bear on database problems?

The criterion is that if I see a method being applied two or three times, or being reinvented two or three times, then it is probably a method that could have application also in databases. That is what happened with fractals. Now we are trying to do the same with singular value decomposition, with independent component analysis, because these techniques have the potential for deep influence in many disciplines.

I have also been told that you are "the nicest guy in the whole wide world" and "completely altruistic," so I am sure that that also has something to do with the success of your collaborations. Someone suggested that I ask you how you always manage to smile!

You have two brothers who are also computer scientists, and with them you wrote what has become known as the "Faloutsos cubed paper," a very influential paper about power laws that hold with respect to the internet topology of 1997-8. What are these power laws? Do they still hold today? Why should we in the database community care about them?

The power laws still hold today. I have graphs that show the power laws holding for several years after 1997-8. And they seem to hold not only for computer networks! It is like Zipf's Law: some words appear very often, and most vocabulary words appear only once or never at all. The same is true for internet connections: there are a few nodes that are very popular. Everybody wants to connect to AT&T or IBM or Sprint, and nobody wants to connect to a tiny little ISP. The same is true also for company sizes. There are huge companies with a quarter million employees, but the vast majority of companies have only one or two employees. So these power laws will hold for not only for networks, as in the Faloutsos cubed paper, but also for many other settings---and for several centuries, not only recently.

Where should we apply the power laws in the database field?

We can apply them for selectivity estimation, for histograms. Yannis Ioannidis got the VLDB 2003 Best Paper Award for his paper on histograms. Histograms are superbly successful exactly because of these Zipf distributions: if you keep the frequency counts of the few most important attributes, then the rest don't matter that much.

Power laws have a very deep connection with fractals. Power laws, fractals, and self-similarity appear in many settings, and we can use self-similarity to battle the dimensionality curse. In databases and in data mining, if we have a lot of attributes, we say we have the problem of the dimensionality curse. High dimensionality is a problem because the running times of most of the data mining algorithms explode exponentially if there are many attributes. But it's not the dimensionality that matters; it's the fractal, the intrinsic dimensionality, of the data set that matters. The fractal dimensionality is usually much lower, exactly because of the skewed distribution of the importance of attributes. There may be a hundred attributes, but only the first three or four or five of them are the most important ones, and therefore the problem is not as hard as we would fear.

*So, **fractal** is the answer. What is the new and most important question?*

There are a lot of questions that will benefit from fractals, such as analysis of graphs and social networks. There are definitely power laws on almost any type of graph we get, and probably self-similarity also. Social networks (who knows whom), biological networks, and food web networks (who eats whom) all have self-similarity and fractals. The sensor time series analysis that we are working on also has self-similarity. Time series have burstiness, which can be very well described with self-similarity. You have silent periods, explosions, bigger silences, bigger explosions, as opposed to the standard Poisson distribution that says that every now and then you have an event happening. No, instead events are very clustered and self-similar. So I think that fractals will be the answer to multiple questions, not only one.

The field of data mining is young, vibrant, and quickly evolving. What do you see as the major future directions in data mining---where is the field going?

That's another very good question. Definitely there are a lot of possibilities with the web, computer networks, social networks, biological networks, regulatory networks; there is a lot of emphasis on networks. Time series analysis also, because we'll be flooded with measurements from sensors, and we want to find patterns there. We want to find intrusions if we have a network, and so we measure how many packets or how many pings we get per time unit. Bioinformatics should also become a hot area. Actually, it *is* a hot area.

It sounds like you are saying that the field will be concentrating on the application areas for a while.

Yes, but that is a biased opinion, because I'm more oriented toward the practical areas. I'm sure my more theoretically-minded and statistics-oriented colleagues from data mining will have a different opinion. They will be looking forward to studying deeper mathematical problems. So mine is a biased opinion from an application person.

With social networks, what kind of patterns are we mining? What are we looking for?

We want to find common patterns, like the way Jiawei Han is trying to do for AIDS virus molecules, where he is trying to see what submolecules occur often. We want to find what groups of people occur often in a company network and figure out whether the appearance of a particular group is destructive or constructive for the department. We want to figure out which are the outlier edges. So if we have, say, a group of researchers who usually don't talk to each other, then if we see edges between them, these are important edges. These edges are either suspicious, because they shouldn't be happening, or else they are very valuable because these are the bridges that make the department work harmoniously together. The problem with data mining is that we are not looking for something specific. We try to look for something that we don't know yet, a pattern that will help us compress this data set.

The data mining field includes people with backgrounds in statistics, AI, and databases. I have heard that people from one area can't understand the conference talks of people from the other areas. I have heard that a statistics person said to you, after one of your talks, that he could always come up with a query that would break your index structure, so what was the point of having the index? What will happen to the field, with this uneasy alliance of subdisciplines that don't really understand each other?

I think what will happen is what is happening already: conferences try to put all these people in the same room and after the first few uneasy years, they will understand each other's mentality, as is the case now. So in database classes, we are teaching about Chi-squared tests because it's useful for statistics, and I'm sure that statisticians are teaching about B-tree indices. I don't remember the specifics of the situation that you mentioned, but there is a lot of cross-fertilization. Yes, the first few years will be uneasy, but eventually the goal is worth the initial pain.

It seems that many people from Greece do database research. What are your views on this heritage---is it an opportunity, a burden, or something else altogether?

I think that it's a happy coincidence. It's the 80/20 law and fractals in action, a clustering effect. A few database professors came back to Greece when I was an undergraduate, like Dennis Tsichritzis, and of course they were all enthusiastic about databases. Then we all went out to the states or Canada, and the same thing repeated itself a few times, creating exponential growth. Now we have a huge number of Greek students and Greek professors doing database research. I'm sure this is the case with other nations too; there are a lot of Indian and Israeli database professors too. So it's a happy coincidence.

You first came to Carnegie Mellon as a sabbatical visitor, and then stayed on to become a professor there. What was it like to make the transition from Maryland, a database-oriented department, to CMU, which had never had a database faculty member when you arrived? How do you cope with having to justify your entire discipline to everyone you meet?

Actually, it was a very pleasant transition, because Carnegie Mellon was actively trying to build a database group. So yes, I had to do some justification, but it was mainly what I had to do was education. I had to tell people that a database is not a collection of information, it's a collection of tables with SQL on top. People at Carnegie Mellon are extremely nice and very cross-disciplinary; they are all hand picked to be cross-disciplinary by the hiring process. So it was very easy.

How did you convince them that databases were important? Were you arguing that databases are of economic importance, or just intellectually interesting?

I didn't have to argue at all, because they were already convinced that database research is important. That's why they invited me.

Natassa Ailamaki says she has had to justify her discipline over and over again.

We have had more to *explain* than to justify, because they all have large data sets and they want to do something with them. Even during my year as a visitor, people would say, "Oh, so you know databases, great! I have this problem, can you help me? I have time series for monkey brains. How can I store them and try to find similarities?" They want to do neurobiology and figure out how the monkey brains operate when you show them visual stimuli. So it was not a matter of justification, it was a matter of a quick crash course.

I have heard that you take many sabbaticals, and that you like to spend them in a particular way. What recommendations do you have for faculty members considering a sabbatical?

I think sabbaticals at an industrial lab are very valuable because they help us get in touch with real problems, real customers, and get in touch with the trenches. That's why I spent two sabbaticals at IBM with Rakesh Agrawal and Bill Cody, and at AT&T with Avi Silberschatz and H. V. Jagadish when they were both at AT&T. So my suggestion is to try to find out what real customers are complaining about.

To really be in touch with the real customers, wouldn't you need to go a development group?

Actually, no, because the collaborators have direct contact with customers or with other people within the company who are in direct contact with customers. It's close to customers but not extremely close, not face-to-face interviews with customers, but their complaints percolate eventually to the research labs.

Do you have any words of advice for fledgling or mid-career database researchers or practitioners?

I think the major advice is for people to make sure that they enjoy what they are doing. If you find a topic interesting, then other people will find it interesting too. Before tenure, of course it's important to focus on the rules of the game: if the university wants journal publications, let's make sure that we have the appropriate number, and so on.

After tenure, I think people are free to do what they enjoy most, which is a matter of taste. Personally, I prefer to work on problems that have practical importance and also can use some nice theoretical solutions. Other people prefer to focus on practical issues; as long as the problem is important for companies or society, then they work on it. And at the other extreme, some people work on completely theoretical issues that may or may not have applications. I think all three modes are valuable. People should pursue whatever keeps them up at night.

The problem that I see with your philosophy is that after you get tenure and have the freedom to do whatever you want, you still have those grad students. In order to get a job, many of them feel that they have to behave a lot like an assistant professor in producing all these papers. So how do you get off the treadmill once you get tenure?

I don't think you get off the treadmill ever! Unfortunately, people think, “If I get tenure I’ll get relaxed.” No, nobody will relax. Just more freedom and more peace of mind, but still it’s still the same amount of work. I think the amount of work the person is doing when he or she is a graduate student will be the same until this person retires. It’s a matter of state, it’s a matter of mentality.

We shouldn’t tell them that, should we? Won’t all the grad students who read this get depressed?

I don’t think so. It is true, they are smart and they see their professors (assistant, associate, full, emeritus) working 10 and 12 and more hours a day. But this is *fun* work, this is something that people enjoy doing, and I don’t think it’s bad.

That’s a good lead-in to my next question: if you magically had enough extra time at work to do one thing that you are not doing now, what would that thing be?

Nothing different.

More of the same?

More of the same: playing with drafts, collecting data sets, trying to find patterns, trying to figure out what is the next best tool to use for all the problems mentioned before.

Among all your past research, what is your favorite piece of work?

It is probably the PODS 94 paper on how to use fractals to characterize non-uniformity of a cloud of points, so that we can figure out the performance of R-trees and other spatial access methods.

If you could change one thing about yourself as a computer science researcher, what would it be?

Maybe get better organized. For now, things are not so well organized.

Sometimes people have their secretary or their postdoc do all their organizing, to keep them on track. Have you tried that?

No, I should. That’s a good idea.

I have been told that you are a resource for jokes for the Greek database community. Can you tell us a joke to conclude our interview?

Of course! The shortest joke I know: I'm an atheist, thank God!

Thank you very much.

Thank you very much!