Interview with Jon Kleinberg

Gregory Piatetsky
KDnuggets
editor at kdnuggets

ABSTRACT

Interview with Jon Kleinberg, a pioneer in web mining, social network analysis, and other fields and a winner of many awards, including 2 KDD Best Papers and a MacArthur 'genius' award.

Keywords

Links, Authorities, Social Network Analysis, Small-World.

1. Kleinberg – a Brief Bio

Jon Kleinberg is currently a Professor of Computer Science at Cornell. He received his B.S. from Cornell in 1993 and his Ph.D. from MIT in 1996. Jon Kleinberg is one of the most interesting, talented, and wide-ranging researchers today. He has made significant contributions and written papers in numerous areas, including Web Analysis and Search: Hubs and Authorities, Small-World Phenomena and Decentralized Search, Social Network Analysis, Algorithms and Complexity, Clustering and Data Mining; and Genomics and Protein Structure Analysis

He received many awards, including the MacArthur 'Genius Award' (2005), Nevanlinna Mathematics prize (2006) and KDD-2003 and KDD-2005 Best paper awards.

To data miners he is best known as one of the founders and leaders of the "social network analysis" field of computer science, as well as the author of HITS (Hubs and Authorities) algorithm, which was very influential in web search and conceptually similar to Google's PageRank. Some credit Kleinberg's work with being one of the inspirations for PageRank.

2. First Program

Gregory Piatetsky-Shapiro: What attracted you to computers? What was the first interesting program you wrote?

Jon Kleinberg

Kleinberg: I've always been interested in mathematics, and I discovered computers when I first used an Apple II around sixth or seventh grade. At school, we all used them to play games, and the temptation quickly developed to try writing our own variations on the games. So these were the first interesting programs I wrote -- text-adventure games, games that used the simple graphics features, things like that. In a funny way, the limited power of personal computers then made it possible to feel like you were really using the machine to its full potential, even if in retrospect one can see all the additional subtleties that were hiding in the background.

The evolution of information technology over the past few decades has been so fast that you can look at professional computer scientists who differ in age by very little, and discover that their formative experiences with computers were radically different. It slices the age groups into micro-generations that I think helps give the field some of its great diversity of perspectives on problems. And by now, of course, you have freshman entering college who were four years old when the Mosaic browser appeared.

3. Links, Authorities and Hubs

Gregory PS: Around 1996 you developed the HITS algorithm (Hypertext Induced Topic Selection), which is a link analysis algorithm that rates Web pages for their authority and hub values. Authority value estimates the value of the content of the page; hub value estimates the value of its links to other pages. Using links rather than keywords was a major advance for web search engines. How did you come to the idea of using links and authorities and hubs ?

Kleinberg: Improving the quality of Web search was a research problem on the minds of many people in 1996, and Prabhakar Raghavan, who was in the process of recruiting me to spend a year post-PhD at IBM Almaden, had very interesting ideas on this issue, and he encouraged me to join his group in thinking about it. From my perspective, there was considerable motivation to explore Web search as something more than a pure text retrieval problem: I was coming from a theoretical computer science background, focusing on graph algorithms, and it was hard to avoid noticing that in the Web we had a system where enormous numbers of people were, through their browsing, busily engaged in the traversal of an enormous graph, harvesting information as they went.

A few researchers, including Ellen Spertus and Peter Pirolli, Jim Pitkow, and Ramana Rao, had begun considering the link structure of the Web in their work on search. Understanding the network structure of the Web seemed to be both scientifically compelling, and also a potential route for improving Web search.

The idea that there were two kinds of interesting pages -- essentially, resource lists and the authorities themselves -- was an intuitively fundamental part of one's Web browsing experience in mid-1990's. It was a common experience that you would try to learn about a topic, gradually find people who had assembled hub-like resource pages with extensive information and links, and eventually discover a kind of consensus in where these links pointed -- in other words, you would find the authorities that these hubs consistently endorsed, and from this you'd find both the good authorities and also the good hubs that had led you there. This intuition from everyday experience suggested that there was a more general rule here that could be formalized.

What's nice is that a direct way of formalizing the idea works well: you simply start from the premise that you want to assign scores to pages as authorities and hubs, in such a way that the good authorities tend to receive links from good hubs, and the good hubs tend to point to good authorities. The equilibrium inherent in this notion -- that the quality of a page depends on the quality of other pages -- is related to ideas in the area of spectral graph theory, and from this body of work one knows that such equilibria can be made precise using eigenvectors. Brin and Page's definition of PageRank has an analogous kind of eigenvector-based equilibrium in it, in which you roll the notions of quality into a single "goodness" measure: roughly, good pages are those that receive links from other good pages.

From the perspective of designing search tools, there is clearly a huge amount of information contained in both the text content and the links -- as there also is in a searcher's pattern of clicks, and a number of other sources of data. Over time, commercial search engines have built on the different threads of research in this area to take advantage of numerous features of all these types.

4. Social Network Analysis and Small-World Networks

Gregory PS: You have made major contributions to social network analysis and "small-world" networks. For readers that are not familiar with this field, can you summarize the key ideas in 2-3 paragraphs?

Kleinberg: While on-line information networks such as the Web are relatively recent in origin, social networks extend back to the earliest parts of our history. In a social network, nodes represent people or other social entities, and links indicate some kind of social interaction (for example, friendship, collaboration, or influence). Social networks have been central objects of study in the social sciences for a long time, since they have the potential to help illuminate how social outcomes can arise not just from the properties of individuals in isolation, but from the pattern of interactions among them -- in other words, from the structure of the network.

When you look at large-scale social networks drawn from different settings, you see a number of recurring patterns. A central one is the "small-world phenomenon" -- the fact that most pairs of people in a large social network are connected by very short paths. The social psychologist Stanley Milgram provided perhaps the first strong empirical evidence in support of this in the 1960s, using a now-famous experiment in which he asked people in the Midwest to forward letters through chains of friends to a "target" person in a suburb of Boston. Among the paths that successfully reached the target, the median length was six, a number that has since entered pop culture as the "six degrees of separation." There have been a number of mathematical models aimed at capturing the abundance of short paths in social networks; the 1998 work of Duncan Watts and Steve Strogatz in particular catalyzed a huge amount of research along these lines.

I became interested in the small-world problem because of what I viewed as the striking algorithmic content of Milgram's experiment: that the people taking part in the experiment were in fact performing a kind of decentralized routing. Each person had only local information about the network, but collectively they were able to route the message to a far-awar destination. My work on this problem centered around the development of social network models, building on the Watts-Strogatz framework, in which one could quantify the power of such decentralized algorithms.

More generally, there are huge opportunities at the moment in the study of social networks, since the digital traces of on-line communication have produced huge datasets on social interactions. There has also been the creation of new social structure on-line, through social networking sites such as Facebook and many other kinds of collaborative on-line media. A deep understanding of such data will require the collaboration of computing and information science with the social sciences -- a kind of synthesis that we're actively pursuing at Cornell.

5. Kleinberg - a "Rebel King" ?

GPS, Q4: Your students sometimes call you "Rebel King" (an anagram for Kleinberg). Tell us about the "Rebel King" phenomenon in the classroom.

This started in a large undergraduate course on discrete math for computer science majors in Fall 2000. The course was a lot of fun to teach -- there were more than 200 students in the class, but they were still willing to participate and discuss ideas during lecture, which made things much more lively. That was 7 years ago, so it's interesting how the tradition you mention persists, beyond the time span of any one Cornell undergraduate. (It's more than just the effect of having it preserved on Wikipedia, though that likely helps.)

But the fact that one can teach large classes and have students who are motivated and enthusiastic about participating -- this is a great thing. The experience of teaching and working with the undergraduates here is definitely one of the powerfully appealing features of being at Cornell. This year I'm again reminded of this, as David Easley and I finish teaching our large introductory undergraduate course on networks in the social, technological, and natural worlds. It was a great experience -- we had over 200 students from 25 different majors or intended majors -- and it was a reminder of how quickly research-level ideas can make their way down to the introductory college level, and eventually enter the general vocabulary more broadly.

6. Key insights from 2 KDD Best Papers

Gregory Piatetsky-Shapiro, Q5: Can you summarize the key insights of your 2 KDD Best Papers:

- David Kempe, Jon Kleinberg, and Eva Tardos, Maximizing the Spread of Influence through a Social Network, KDD-2003 Best Paper.
- Jure Leskovec, Jon Kleinberg, and Christos Faloutsos, Graphs Over Time: Densification Laws, Shrinking Diameters, and Possible Explanations, KDD-2005 Best Paper

By way of background, the mechanisms through which information and influence spread across a social network is a topic that has a rich history in the social sciences. For example, a large sub-field of sociology is concerned with empirical studies of the diffusion of innovations, in which one analyzes how new technological and social practices cascade from a set of early adopters, through personal and professional networks, and ultimately to success or failure. (Word-of-mouth effects and "viral marketing" are two other reflections of this phenomenon.) Formal models for these kinds of processes have also been studied by mathematical economists and sociologists.

This topic has increasingly been attracting the interest of computer scientists over the past few years -- in part this is because we are collecting on-line data on the spread of ideas and technologies, and in part it's because the kinds of probabilistic models for studying networks that our community has been developing turn out to fit very well with these types of questions.

In our KDD 2003 paper, David Kempe, Eva Tardos, and I were motivated in particular by the confluence of these social science models with a paper that Pedro Domingos and Matt Richardson had just published at KDD 2001. Pedro and Matt raised the following style of question in their paper: if we have the ability to market a new product to a small number of initial adopters in a social network, whom should we target if the goal is to create as a large a cascade of further adoptions as possible? What David, Eva, and I were able to show in our 2003 paper is that, while the general problem is computationally intractable, one can find strong approximation algorithms for selecting an influential set of initial adopters in some of the most fundamental models. The basic idea was to exploit a subtle kind of "diminishing returns" property (known in discrete optimization as "submodularity") that enables hill-climbing heuristics to produce close-to-optimal solutions. This analysis tool then became a guide for developing richer models that were still computationally manageable; some of these subsequently proved useful to researchers modeling the spread of other kinds of information, including the spread of news stories through the links among weblogs.

Over longer time scales, the network structure itself is subject to dynamic processes and change, and this was the subject of the KDD 2005 paper with Jure Leskovec and Christos Faloutsos. We looked at a number of different kinds of networks (including citation and co-authorship networks from the scientific literature, for which extremely detailed temporal data is available), and we found that they exhibit some recurring but unexpected phenomena as they grow. First of all, they densify -- with the average number of links per node increasing according to a fairly simple pattern -- and second, the average distance between nodes tends to shrink slowly over time. Subsequent work by a number of other groups has identified these two phenomena in a range of further networks over time, suggesting their generality. We also considered a class of models that -- through simulation at least -- exhibit densification and shrinking distances; the models operate by positing that nodes form new links through a sequence of rapid, cascading processes on the network structure. This hints at an interesting potential connection between the 2003 and 2005 papers, suggesting some of the ways in which dynamic network behavior over short time scales might have structural effects over much longer time scales.

7. Facebook and other social networking sites

Gregory PS: There are many social networking sites, like MySpace, Facebook, LinkedIn, etc that are trying to ride the Web 2.0 wave. What can a researcher like you learn from their success and do they do any interesting link/network analysis?

Kleinberg: Social networking sites are seeking to accomplish many goals at once -- interpersonal, informational, economic, and many others. It will be very interesting to see whether these sites converge to a dominant style of use, and what that will be.

The informational and economic aspects raise many appealing questions. Essentially, when viewed as tools for finding information and getting questions answered, these sites are part of the broader trend toward information-seeking based on the knowledge possessed by people you know, supplementing the much vaster but less verifiable knowledge that resides on the Web at large. In other words, the friends of your friends may not always be able to answer your questions, but when they can, the answer is endowed by this network of friendships with a chain of trust. The trade-offs between this and the more traditional notion of Web search -- and some of the economic consequences of this trade-off -- is a question that Prabhakar Raghavan and I explored in our work on Query Incentive Networks in 2005.

The interpersonal aspects of these sites raise challenging questions as well. If you look at the experience of college undergraduates using Facebook, for example, one of the striking things is the way it really puts the idea of a social network on center stage. For many generations, people have moved away from home to places like college campuses and been implicitly aware of the ways in which they were forming connections, embedding themselves in a social structure. But the underlying social network itself was always latent, invisible, and to some extent, unknowable. Now, to have constant access to Facebook, to have your place in the social network depicted so explicitly on a computer screen, updating itself as you form these connections, following your progress like a kind of scoreboard -- this foregrounding of the social network is really a new phenomenon. It will be interesting to see where the feedback effects of such an explicitly visualized network will take us.

8. On current work and start-ups

Gregory PS: What new ideas you are working on and excited about? Do you see a start-up in your future?

Kleinberg: I'm continuing to collaborate with people in the social sciences, particularly economics and sociology, and I think the opportunities here continue to expand rapidly. There is a lot we can learn from these areas about building models of human populations -- particularly, the decisions people make and the population-level feedback effects that arise from these decisions. At the same time, there are deep questions about what form such models should take, and how the data should inform the development of models, and these are kinds of questions where the techniques we specialize in can be very useful.

It is also very interesting to build stronger connections between the types of network studies going on in data mining on the one hand and sociology on the other. The bulk of empirical network research in sociology has focused on relatively small systems (up to a few hundred individuals) for which researchers had a good

understanding of the people being studied and the meaning of the connections between them. Now, when you go and collect on-line data -- say, from a social networking site -- so as to study a network with a few million people, you end up a very different situation. You know much less about who the people are and what the links mean; but at the same time, you can potentially identify patterns that, while genuine, are too subtle to be visible at smaller scales. Thanks to increasing communication between the disciplines, these approaches are moving closer together; and the ultimate goal, of course, is to find the point at which the two lines of research converge -- giving us the ability to ask complex, nuanced questions from the social sciences on social networks of enormous size.

In a slightly different direction, there are many interesting questions in the development of tools to mine one's own personal data. As we accumulate digital archives capturing the history of our personal reading, writing, listening, and communication habits, we're going to be increasingly interacting with tools that mine this in the course of shaping our on-line experience. And of course, one can't go very far down this line of thinking without running into fundamental questions about the privacy of this kind of data, a problem I've been thinking about recently with Lars Backstrom and Cynthia Dwork in the context of social networks.

(GPS: see L. Backstrom, C. Dwork, J. Kleinberg. Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. Proc. 16th Intl. World Wide Web Conference, 2007.)

As for your question about start-ups -- that's of course a recurring issue for anyone working in this area. Personally, I like being in academia -- both the opportunity to interact and collaborate with people in very different disciplines, and the impact that comes from dealing with students at different levels. But part of my view is based on the fact that we're talking about an area in which the movement of ideas between academia and industry has been particularly fluid; people at both large and small companies are aware of what's going on in academic research, we all attend the same conferences, and there's a sense that you can do things that will really get used.

9. Recent books read

Gregory PS: What was the last book that you read and liked ?

Kleinberg: In preparing for my class on networks this spring, I re-read Thomas Schelling's "Micromotives and Macrobehavior", one of our assigned readings for the course. It's a very nice exploration into the power of mathematical models in the social sciences, and the ways in which one can use them to identify common themes at work in many different settings. It's also a book that rewards re-reading, since you see fresh things in it when you know more about the technical context surrounding the examples.

I also recently finished reading Fernand Braudel's "On History"; it was recommended to me by a friend who's a historian as a glimpse into something of that field's view of time and temporal processes at a relatively abstract level. This is part of my long-term interest in trying to think about the role of time in complex data; since much of the challenge here lies in finding the right structures and definitions, many of which are quite subtle, it's useful to look as broadly as possible into what other disciplines have done when reasoning about the importance of temporal phenomena.

10. Advice to data mining students and young researchers

Gregory PS: What advice would you give to data mining students and beginning researchers who are just starting to work in this area ?

Kleinberg: First of all, it's a great time to be working in this field. The problems are of fundamental interest, the connections to other areas run deeply, and the solutions have the potential for real impact -- just the ingredients you hope to find in a research area.

As for specific pieces of advice -- I've gotten lots of wonderful advice from my own professional mentors, and have tried to pass this on to students. Much of this has been about how to choose research problems to work on -- one of the basic tasks we all face.

As with many fields of research, some of the most important open problems are simple to pose but hard to start making progress on. In approaching such questions, it's useful to think about the possibility of employing methods from radically different fields. This is one of the reasons why it pays to stay informed about what people are doing in areas that may seem superficially distant from your own; progress often happens not through a direct assault on a problem, but through experimenting with a novel technique.

And finally, it's always worth seeking out research projects that strike you as inherently interesting, aesthetically appealing, and fun. It's much easier to make progress on something when you're having fun with it.

11. NOTES and REFERENCES

A first version of this interview appeared in KDnuggets,

www.kdnuggets.com/news/2007/n11/3i.html

For more information see Jon Kleinberg home page

and Jon Kleinberg Wikipedia entry

Gregory Piatetsky-Shapiro is the Editor of KDnuggets and Chair of ACM SIGKDD (www.kdnuggets.com/gps.html)