Tutorials

MA1

How to Analyze Massive Social Network Datasets without a Cluster

Presented by Derek Ruths

For all the wonderful things we hear about how compute clusters enable the analysis of massive datasets, the sad truth is that few researchers can use them. This is due to practical issues: clusters are expensive, their administration requires nontrivial time and technical knowledge, and the tools for using them aren't user friendly. Nonetheless, the average size of network-based datasets is growing. This means that more and more researchers need to do large network analysis but don't have a cluster at their disposal.

In this tutorial, we will teach attendees several easy-to-use strategies for analyzing large social network datasets (upwards of millions of nodes and edges with metadata) on a desktop or high-power laptop. These strategies involve using a combination of tools and techniques that capitalize on their individual strengths.

Over the course of the tutorial, attendees will first learn how to use several python-based libraries for large-scale data and network analysis as well as easy ways to store data to flatfiles to minimize retrieval time. Then we will discuss how to combine these tools together. Finally, we will use the newly learned approaches to analyze two large network datasets.

Derek Ruths is a computer science professor at McGill University. Much of his current work involves analysis and modeling of large networked datasets from online social platforms and general ontologies. He likes clusters, but believes that laptops and desktops (but not iPads) can do large-scale network analysis too.

MA2

Charting Collections of Connections in Social Media: Creating Maps and Measures with NodeXL

Presented by Marc Smith

Networks are a data structure commonly found across all social media services that allow populations to author collections of connections. The Social Media Research Foundation's NodeXL project makes analysis of social media networks accessible to most users of the Excel spreadsheet application. With NodeXL, Networks become as easy to create as pie charts. Applying the tool to a range of social media networks has already revealed the variations present in online social spaces. A review of the tool and images of Twitter, flickr, YouTube, and email networks will be presented.

Marc Smith is a sociologist specializing in the social organization of online communities and computer mediated interaction. Smith leads the Connected Action consulting group and lives and works in Silicon Valley, California. Smith cofounded the Social Media Research Foundation, a nonprofit devoted to open tools, data, and scholarship related to social media research. Smith received a B.S. in International Area Studies from Drexel University in Philadelphia in 1988, an M.Phil. in social theory from Cambridge University in 1990, and a Ph.D. in sociology from UCLA in 2001. He is an adjunct lecturer at the College of Information Studies at the University of Maryland. Smith is also a distinguished visiting scholar at the Media-X Program at Stanford University.

MP1

Evidenced-Based Social Design of Online Communities: Getting to Critical Mass and Encouraging Contributions

Presented by Paul Resnick and Robert Kraut

To become or remain successful, online communities and networks must meet a number of challenges that are common to many groups and organizations, offline as well as online. For example, online communities must handle the start-up paradox, when early in their lifecycle they have few members to generate content and little content to attract members. Throughout their lifecycle, they must recruit and socialize newcomers, encourage commitment and contribution from members, solve problems of coordination and encourage appropriate behavior among members and interlopers alike.

The social sciences tell us a lot about how to make thriving online communities. Economics and various branches of psychology offer theories of individual motivation and of human behavior in social situations. The theories generalize from observations of naturally occurring behavior, from controlled experiments, and from abstract mathematical models. Properly interpreted, they can inform choices about how to meet the challenges described. This tutorial will focus in particular on problems of encouraging contributions and getting new communities to critical mass.

It is based on selected sections of *Building Successful Online Communities: Evidence-based Social Design,* coauthored with Robert Kraut, on using the social sciences as a guide to designing online communities.

Paul Resnick is a professor at the University of Michigan School of Information. He received the master's and Ph.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology. His researches focus on online communities, recommender systems, and applications in politics and health.

MP2

Sentiment Mining from User Generated Content

Presented by Lyle Ungar and Ronen Feldman

The proliferation of user generated content on the web is driving a new wave of work on the determination of user sentiment from web texts such as message boards, blogs, tweets, and Facebook status updates. Both researchers and practitioners are developing and applying new methods to determine how users feel about everything: products and politicians, friends and family, scientific articles and celebrities. This tutorial will cover the state of the art in this rapidly growing area, including recent advances that combine information extraction with sentiment analysis to improve accuracy in assessing sentiment about specific entities. We will present several real world applications of sentiment analysis. Special emphasis will be given to lessons learned from years of experience in developing real world sentiment analysis systems.

Lyle H. Ungar is an associate professor of computer and information science at the University of Pennsylvania. Ungar received a B.S. from Stanford University and a Ph.D. from the Massachusetts Institute of Technology. He directed Penn's Executive Masters of Technology Management Program for a decade, and is currently associate director of the Penn Center for BioInformatics. He has published over 100 articles and holds eight patents. His current research focuses on developing scalable machine learning methods for data mining and text mining.

Ronen Feldman is one of the world's most recognized experts in the field of text mining, link analysis and the semantic analysis of data. In 1997, he founded ClearForest, a Boston-based business intelligence company later acquired by Reuters. He coined the term "text mining" in 1995, and his textbook, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data* is considered the world's premier authority on this complex topic. He currently serves as the head of the Information Systems Department at the Business

School of the Hebrew University of Jerusalem and was an adjunct professor at New York University's Stern Business School. He has given over 30 tutorials on text mining and information extraction and has written numerous scholarly papers on these topics. He received his Ph.D. in computer science from Cornell University and his B.Sc. in mathematics, physics and computer science from the Hebrew University of Jerusalem. He began his career in the military assigned to the elite Talpiot Group, and served for eight years as an officer in the Israel Air Force.

MP3

Information Extraction for Social Media Anaylsis

Presented by Denilson Barbosa

More and more regular users use the blogosphere to express and discuss their opinions, the facts, events, and ideas pertaining to their own lives, their community, their profession, or society at large. It goes without saying that being able to extract reliable data from this medium opens the door to the most varied kinds of analysis and using datasets of massive proportions. As a result, a great deal of attention has been devoted lately to applying information extraction to the blogosphere. In this tutorial, I focus on a specific subproblem: extracting information networks that act as summaries of the blogosphere as a whole. These networks consist of nodes representing entities and edges representing the relationship between such entities. I will cover fundamental tools from NLP and network science that allow the unsupervised extraction information networks from social media content.

Denilson Barbosa is an associate professor at the University of Alberta, working on databases, information retrieval, and the management of linked data. He received a PhD from the University of Toronto (2005), and is a member of the NSERC Strategic Network on Business Intelligence and the Canadian Writing Research Collaboratory.