Overview of the 2003 KDD Cup

Johannes Gehrke
Cornell University
Department of Computer Science
Ithaca, NY 14853, USA
johannes@cs.cornell.edu

Paul Ginsparg
Cornell University
Departments of Physics, and Computing and Information Science
Ithaca, NY 14853, USA
ginsparg@cornell.edu

Jon Kleinberg
Cornell University
Department of Computer Science
Ithaca, NY 14853, USA
kleinber@cs.cornell.edu

ABSTRACT

This paper surveys the 2003 KDD Cup, a competition held in conjunction with the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) in August 2003. The competition focused on mining the complex real-life social network inherent in the e-print arXiv (arXiv.org). We describe the four KDD Cup tasks: citation prediction, download prediction, data cleaning, and an open task.

1. INTRODUCTION

The 2003 KDD Cup competition was concerned with the mining of complex networks; it was based on a dataset from the e-print arXiv (arXiv.org), together with a set of tasks designed to capture some of the challenges inherent in analyzing large social networks. There has been a surge of recent interest in the study of networks across many domains, ranging from communication and information networks such as the Internet and the Web to social and biological interaction networks (for recent surveys, as well as books aimed at more general audiences, see e.g. [1; 3; 5; 7; 10]). This work has been characterized by an emphasis on networks as complex phenomena to be studied, rather than simply as artifacts to be designed; it has been stimulated by a growing awareness that complex networks arising in many different settings have a surprising number of qualitative features in common.

Our focus in the KDD Cup was on *social networks*. Roughly speaking, a social network is a structure in which the nodes represent social entities such as people or organizations, and the edges represent some form of relationship or interaction among them (see e.g. [8]). Social network analysis is a natural domain in which to apply data mining techniques, and a growing body of work has explored some of the potential applications of social network mining. By way of illustration, consider the following pair of examples.

- A large community of Internet users engaged in chat or instant messaging naturally defines a social network, and understanding the structure of this network can help one design services that improve the experience of users in such a system.
- The set of collaborations among employees in a large organization or company naturally defines a social network; understanding its structure can help one improve the efficiency of information flow and collective problem-solving, and to identify the people who play a critical role in the functioning of the organization. In some cases, as in recent studies of *terrorist networks*, the organization is a structure that tries to remain hidden from observation, and the goal is to infer enough about its functioning so as to effectively disrupt it.

A major obstacle in the evolution of social network mining research is the lack of datasets that are simultaneously complex, realistic, and reasonably complete. Most work to date has dealt with datasets that only exhibit two of these three properties: studies in sociology have thoroughly mapped social networks in the real world, but the enormous effort involved in such activities has necessarily limited their size; simulation can produce large, complex networks, but it is not clear how well they map onto real social networks.

In view of this, a promising line of work has studied *collaboration networks*, whose nodes are people working in a particular field or profession, and whose edges join pairs of paper who have collaborated on a project together. Examples that have by now received extensive study are the collaboration graph of movie actors and actresses drawn from the Internet Movie Database (see e.g. [9]), and the co-authorship graphs of researchers who have papers appearing in the physics e-print arXiv, Medline, or other venues [4; 6]. Such networks have the advantage that they are very large and complex, yet they have been completely "mapped". Their applicability to general network mining research is based on the argument that they capture many of the fundamental properties of large-scale social networks. Thus, even if our goal is to study social networks in general (rather than the behavior of movie performers or scientific researchers *per se*), such networks can serve as high-quality "model systems" for experiments and for the evaluation of data mining algorithms.

It is in this spirit that we employ the e-print arXiv as a rich dataset for network analysis. By way of background, the arXiv was started by Paul Ginsparg in 1991; while it was initially created to serve a small research community, it has since grown enormously to become the primary means of disseminating research results in many areas of physics. As of this writing, it contains roughly a quarter of a million papers, with about 40000 new papers added each year. Unlike systems such as CiteSeer, which collect papers by crawling the Web, the arXiv receives papers directly by author submission. The arXiv has an active readership as well;

it receives roughly 10 million requests per month, with the full text of each paper downloaded over 300 times on average within the first few years of submission. (The most popular papers can receive tens of thousands of downloads.)

At the heart of the KDD Cup data is the citation graph of papers in the high-energy physics section of the arXiv; this is the graph whose nodes are papers and whose (directed) edges connect papers to the other papers that they cite. (The full text of the papers contains an implicit representation of the citation graph, and an easily usable, "ground truth" representation of the graph is maintained by a group at SLAC/SPIRES.) Like a collaboration graph, a citation graph encodes information about the social structure of the scientific community; the field of *bibliometrics* has a long history of studying citation networks using the tools of social network analysis [2].

Building on this citation structure, we also provide for the KDD Cup the full text of the papers, and some limited download data. Thus, we have a simultaneous view of the content (paper text), structure (citations), and usage (downloads) of the arXiv network, and hence have the ability to pose a wide range of data mining questions. The KDD Cup tasks make explicit use of all aspects of this data.

2. THE TASKS

The Cup consisted of four tasks. We will describe the tasks out of order, starting with Tasks 1 and 3 and then moving on to Tasks 2 and 4. Overall, we received 57 submissions from countries around the world including Australia, China, France, Germany, India, Japan, South Korea, Slovenia, Switzerland, and the United States. Most groups had three or fewer members; the largest group had 12 members.

2.1 Task 1: Citation Prediction

In the first task we asked the question: Can we understand how the network evolves over time, and which nodes gain and lose in importance? We implemented this by asking participants to predict changes in the number of citations received by well-cited papers over time; we defined a paper to be well-cited if it received six or more citations from February to April 2003. (There were 441 well-cited papers overall in the selected database.) The specific task was to predict, for each well-cited paper, the difference between (a) the number of citations it received during the period from February 1, 2003 to April 30, 2003, and (b) the number of citations that it received during the period from May 1, 2003 to July 31, 2003. A vector of predicted changes was formed, with one coordinate for each well-cited paper, and the evaluation metric was the L_1 distance between this vector and the vector comprised of the true changes.

The task used a subset of papers categorized as "hep-th", which is the High Energy Physics Theory section of the arXiv. There were 30,119 papers written by 57,448 authors comprising in total 1.7GB of LaTeX sources with 719,109 total citations in the papers. 363,812 of these citations were external citations (citing a paper outside hep-th), and 355,297 citations were internal citations.

Contestants had information about both the content and structure of the citation graph. SLAC/SPIRES provided the citation graph, and from the arXiv contestants received the LaTex sources of each paper, including a separate user-submitted abstract and the arXiv submission date.

There were three entries with L_1 difference scores between 1300 and 1400, and four between 1400 and 1500. This task was very difficult; for example, simply predicting zero change for all papers would have been placed 11th among all entries. In addition, the L_1 difference was not dominated by the most highly-cited papers; the 20 most-cited papers accounted for only a few percent of the overall L_1 difference.

First place in Task 1 went to J N Manjunatha, Raghavendra Kumar Pandey, S R Sivaramakrishnan, and Narasimha Murty from the Indian Institute of Science; they describe their approach in the paper "Citation Prediction Using Time Series Approach," which appears later in this issue. Second place went to Claudia Perlich, Foster Provost, and Sofus Kacskassy from New York University; their paper "Predicting citation rates for physics papers: Feature construction for an ordered probit model" also appears later in this issue. Third place went to David Vogel from A. I. Insight, Inc.

2.2 Task 3: Download Estimation

In Task 1, the goal was to predict the evolution of explicit links in the network — the citations that are created as new papers are submitted to the arXiv. But these links are only manifestations of past activity "behind the scenes," namely the usage patterns of the arXiv. The question that we asked in Task 3 was: Can we find patterns in this usage behavior? The goal of Task 3 was to predict the number of downloads that a paper receives in its first two months in the arXiv.

Estimating downloads is interesting since it creates a link between the explicit structure of the network and its usage — between an ephemeral activity and something long-lasting. For example, when looking at both download and network structure, we can observe that growth in download activity is followed by growth in citations. Similar to fingerprints at crime scenes or frozen specimens in glaciers, we can see citations as frozen evidence of usage. A second tie between content and usage is provided by the arXiv's topic-specific mailing lists, which causes papers on similar topics to have similar download histories.

For Task 3, the contestants were provided with the data from Task 1 as well as the following download data: for each paper submitted to the arXiv in February 2000, March 2000, February 2001, April 2001, March 2002, and April 2002, they were given the total number of times the paper was downloaded from the main arXiv site in the first 60 days after its submission. The contestants needed to submit estimates of the corresponding quantity for papers submitted in April 2000, March 2001, and February 2002. The evaluation metric was the absolute difference between the predicted and true download counts, summed over the 50 papers with the most downloads from each period. (Contestants were allowed to make predictions for all papers.)

This task was also very hard since there were external influences that were essentially impossible to predict just from the papers alone. For example, one paper, `hep-th/0103239`, received 7160 downloads – significantly more than any other. This spike resulted from a story about that paper that appeared in *New Scientist* on April 14, 2001; the story included a link to the paper, bringing a large number of readers outside the "standard" pool to the arXiv.

First place in Task 3 went to Janez Brank and Jure Leskovec from Jozef Stefan Institute in Slovenia; their paper "The Download Estimation Task on KDD Cup 2003" appears later in this issue. Second place went to Joseph Milana,

Joseph Sirosh, Joel Carleton, Gabriela Surpi, Daragh Hartnett, and Michinari Momma from Fair Isaac Corporation; their paper "Model Builder for Predictive Analytics & Fair Isaac's Approach to KDD Cup 2003" also appears later in this issue. Third place went to Kohsuke Konishi from the University of Tokyo in Japan. The 150 most downloaded papers predicted by his submission had the highest intersection with true top 150.

2.3 Task 2: Data Cleaning

Task 2 was designed as a challenge for data cleaning tools. Recall that SLAC/SPIRES provides the actual citation graph of the papers. This citation graph is created by automated heuristics followed by human post-processing; the automated portion of this process achieves reasonable accuracy, having been refined over many years using extensive domain knowledge. The challenge is that the actual citations in the papers are unclean; for example, they contain spelling variations on author names, abbreviations, typos, and other sources of noise. Moreover, citations in the physics community usually do not contain the paper title. To illustrate, here is an example of three citations to the same article:

- Lisa Randall and Raman Sundrum, Physical Review Letters, 83(17):3370–3, 25 October 1999
- L. Randall and R. Sundrum, PRL 83, 3370 (1999)
- Lisa Randall, Raman Sundrum, Phys.Rev.Lett. 83: 3370–3373, 1999

Note that this is an easy example in which all the information is present; there are many more obscure examples in the LaTeX sources. In the physics community, citations often contain unique arXiv identifiers, for example:

- Lisa Randall, Raman Sundrum, Phys.Rev.Lett. 83: 3370-3373, 1999, hep-ph/9905221

— these identifiers are explicitly used by the SLAC/SPIRES heuristics.

The goal of this task was to recreate the citation graph for papers from a category called hep-ph (High Energy Physics Phenomenology), using only the LaTeX sources of the papers. Concerned the task might be too "easy", we removed occurrences of unique paper identifiers by running a perl script over the sources that removed arXiv id's (alphameric plus 7 consecutive digits), thus effectively making Task 2 harder than that faced by SLAC/SPIRES. The evaluation metric was the size of the symmetric difference between the true and submitted sets of citation links.

With limited time and no domain knowledge, this problem turned out to be very difficult. The true citation graph for hep-ph has about 421,000 edges between 35,000 papers. Only one entry outperformed the empty graph on this evaluation metric. This entry by David Vogel from A.I. Insight, Inc. consisted of only four hand-crafted citations, and it won first place. Second place went to Sunita Sarawagi, Kapil M. Bhudhia, Sumana Srinivasan, and V.G. Vinod Vydiswaran from IIT Bombay, who had the highest number of correct citations with about 40,600 correct citations out of over 175,800 predicted citations. Their paper, "Resolving citations in a paper repository," appears later in this issue. Third place went to Martine Cadot and Joseph di Martino from LORIA, the Laboratoire Lorrain de Recherche en Informatique et ses Applications in France. Their paper, "A Data Cleaning Solution by Perl Scripts for the KDD Cup 2003 Task 2," also appears later in this issue.

2.4 Task 4: The Open Task

In Task 4, we left the definition of the problem to the participants: The goal was to take the data, define the most interesting questions possible, and mine the answers to these questions from the data. A committee of judges selected the winning entry based on novelty, soundness of methods and evaluation, and relevance to the arXiv dataset. The committee consisted of the the three KDD-Cup co-chairs, together with Mark Craven, David Page, and Soumen Chakrabarti. First place went to Amy McGovern, Lisa Friedland, Michael Hay, Brian Gallagher, Andrew Fast, Jennifer Neville, and David Jensen from the University of Massachusetts, Amherst. They describe their work in the paper titled "Exploiting Relational Structure to Understand Publication Patterns in High Energy Physics". Second place went to Shou-de Lin and Hans Chalupsky from the University of Southern California, for their paper "Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset". Third place went to Shawndra Hill and Foster Provost from New York University for their paper "The Myth of the Double-Blind Review", and the fourth place went to Grigorii Pivovarov and Sergei Trunov for their paper "EqRank: A Self-Consistent Equivalence Relation on Graph Vertexes".

Acknowledgements

Manuel Calimlim at Cornell helped in evaluating the KDD Cup entries. Travis Brooks at SLAC/SPIRES provided the cleaned citation graphs and made several updates through the tasks available. Soumen Chakrabarti from the IIT Bombay, Mark Craven from the University of Wisconsin-Madison, and David Page from the University of Wisconsin-Madison were part of the small program committee that evaluated the entries for Task 4. We thank Pedro Domingos, Christos Faloutsos, and Ted Senator for valuable suggestions. More information about the tasks can be found at the website of the KDD Cup (`www.cs.cornell.edu/projects/kddcup`).

3. REFERENCES

[1] A.-L. Barabsi. *Linked: The New Science of Networks*. Perseus Publishing, 2002.

[2] L. Egghe and R. Rousseau. *Introduction to Informetrics*. Elsevier Science, 1990.

[3] J. Kleinberg and S. Lawrence. The structure of the web. *Science*, 294, 2001.

[4] M. Newman. The structure of scientific collaboration networks. In *Proceedings of the National Academy of Science USA*, volume 98, 2001.

[5] M. Newman. The structure and function of complex networks. *SIAM Review*, 45, 2003.

[6] M. E. J. Newman. Who is the best connected scientist? a study of scientific coauthorship networks. In E. Ben-Naim, H. Frauenfelder, and Z. Toroczkai, editors, *Networks: Structure, Dynamics and Function, CNLS 23rd annual conference (Santa Fe, NM, 12–16 May 2003)*. Springer, Berlin, 2004.

[7] S. Strogatz. Exploring complex networks. *Nature*, 410, 2001.

[8] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[9] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393, 1998.

[10] D. J. Watts. *Six Degrees: The Science of a Connected Age*. W.W. Norton & Company, 2003.