KDD-Cup 2003

Paul Ginsparg, Johannes Gehrke,
and Jon Kleinberg

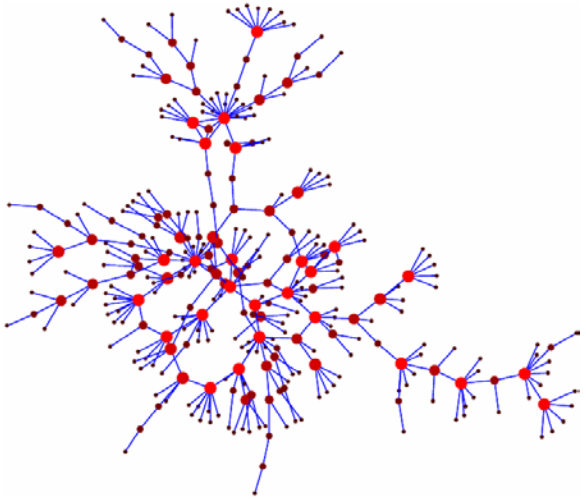
Department of Computer Science

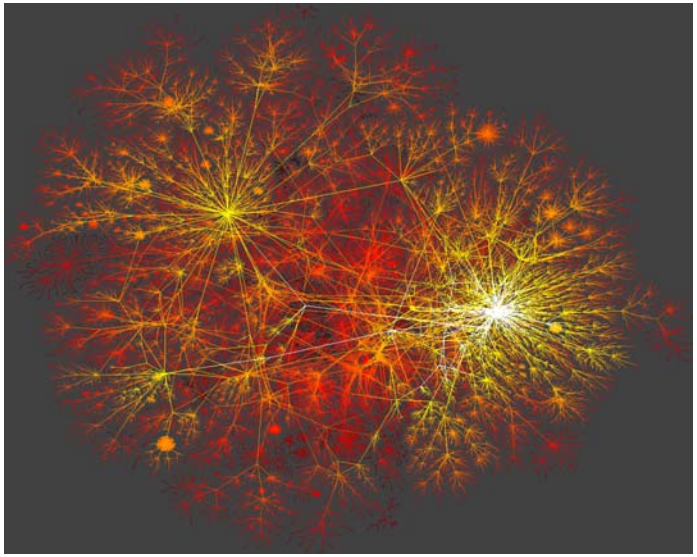
Cornell University

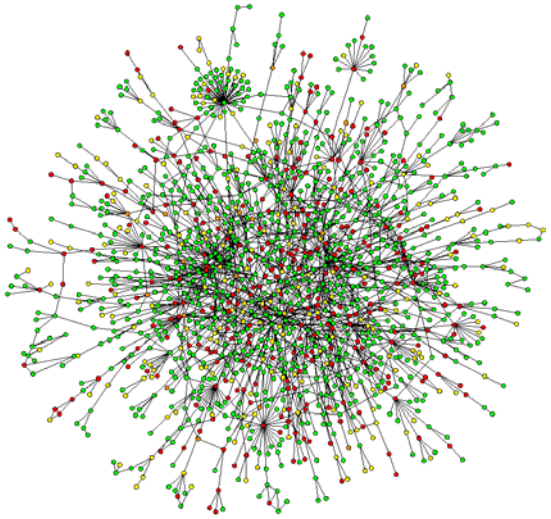
CORNELL

KDD Cup 2003

- Complex networks are everywhere
  - The Internet and the Web
  - Biological interaction networks
  - Social networks
- Emphasis on networks as phenomena rather than as designed artifacts







Social Networks

- Can be defined by influence, collaboration, information flow, …
- Examples:
  - Trust relationships in a financial market
  - Key players in large organizations
  - Trend setters in scientific communities
  - Partially observable activities in terrorist networks

Obtaining Social Network Data

First challenge: Hard to obtain social network datasets that are simultaneously:

- Large and complex
- Realistic
- Reasonably complete

Traditional sociology has primarily studied networks that have (2) and (3)

The arXiv: A Complex Network

- KDD Cup 2003 data: arXiv.org
  - Full text of research papers
  - Explicit citation structure
  - (Partial) download data
- Started by Paul Ginsparg in 1991
- Papers are submitted, not crawled as in CiteSeer
- Has become the main venue for disseminating results in many areas of physics

arXiv.org e-Print archive

Automated e-print archives [physics] Search | Form Interface | Catchup | Help

6 Jul 2003: Cumulative "What's New" pages.
Robots Beware: indiscriminate automated downloads from this site are *not* permitted.

Physics

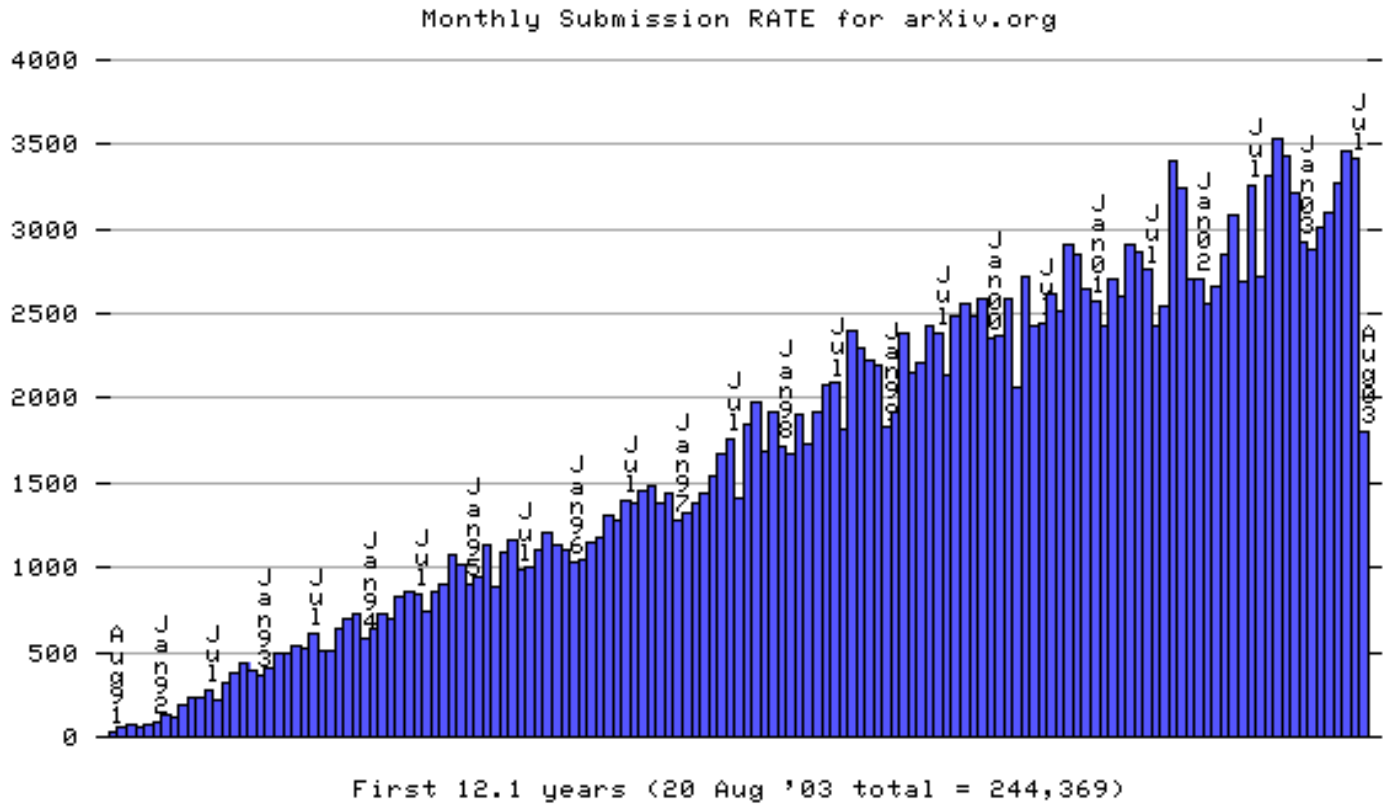
- Astrophysics (**astro-ph** new, recent, abs, find)
- Condensed Matter (**cond-mat** new, recent, abs, find)
  includes: Disordered Systems and Neural Networks; Materials Science; Mesoscopic Systems and Quantum Hall Effect; Soft Condensed Matter; Statistical Mechanics; Strongly Correlated Electrons; Superconductivity
- General Relativity and Quantum Cosmology (**gr-qc** new, recent, abs, find)
- High Energy Physics - Experiment (**hep-ex** new, recent, abs, find)
- High Energy Physics - Lattice (**hep-lat** new, recent, abs, find)
- High Energy Physics - Phenomenology (**hep-ph** new, recent, abs, find)
- High Energy Physics - Theory (**hep-th** new, recent, abs, find)
- Mathematical Physics (**math-ph** new, recent, abs, find)
- Nuclear Experiment (**nucl-ex** new, recent, abs, find)
- Nuclear Theory (**nucl-th** new, recent, abs, find)
- Physics (**physics** new, recent, abs, find)
  includes (see detailed description): Accelerator Physics; Atmospheric and Oceanic Physics; Atomic Physics; Atomic and Molecular Clusters; Biological Physics; Chemical Physics; Classical Physics; Computational Physics; Data Analysis, Statistics and Probability; Fluid Dynamics; General Physics; Geophysics; History of Physics; Instrumentation and Detectors; Medical Physics; Optics; Physics Education; Physics and Society; Plasma Physics; Popular Physics; Space Physics
- Quantum Physics (**quant-ph** new, recent, abs, find)

Mathematics

- Mathematics (**math** new, recent, abs, find)
  includes (see detailed description): Algebraic Geometry; Algebraic Topology; Analysis of PDEs; Category Theory; Classical Analysis and

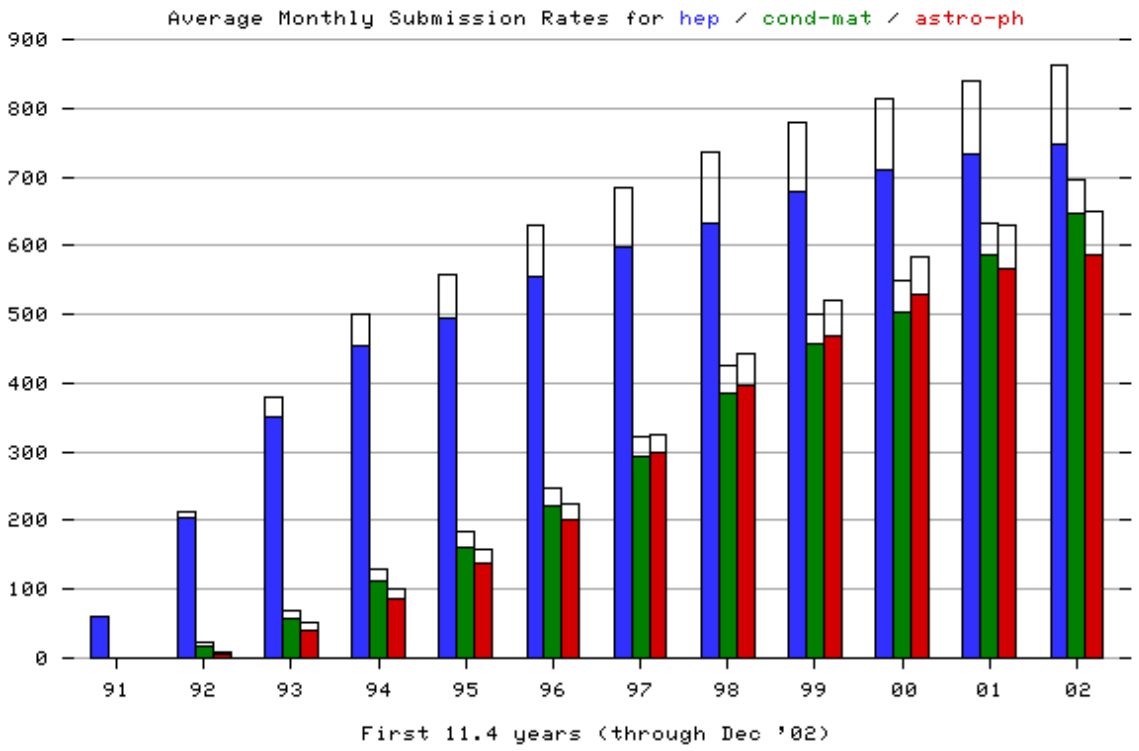
The arXiv (Contd.)

- Originally anticipated submission rate: about 100 papers
- About 233,000 papers, now about 40,000 new papers every year.



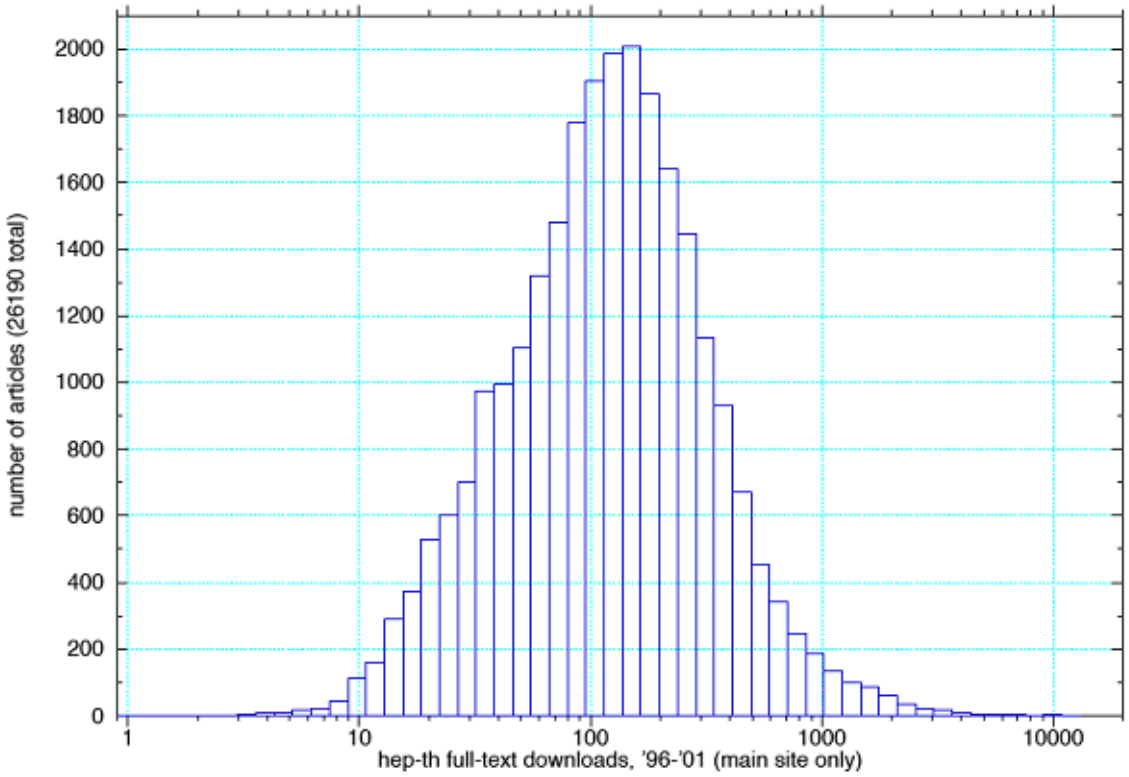
The arXiv (Contd.)

- Computer science is in there, but CS has had much smaller growth (CiteSeer)
  - Sub-communities submission rates
  - Blue is hep; people turn on and then they saturate (hep and astro-ph flattens out, cond-mat will overtake next year?)



The arXiv (Contd.)

- 10 million requests per month with tens of thousands of queries per day
- Web usage logs from 1993
- On average the full text of each paper has been downloaded over 300 times
- The most popular papers have been downloaded tens of thousands of times



CORNELL Data Mining

The arXiv as KDD Cup Dataset

- Full text of the high-energy physics theory (oldest and most active category) and high-energy physics phenomenology papers
- Full citation graph with "ground truth" from SLAC/SPIRES
- Limited download data

→Simultaneous view of content, network, and usage

Submissions Statistics

- Four tasks total
- 57 submissions from around the world including Australia, China, France, Germany, India, Japan, Korea, Slovenia, Switzerland, and the U.S.
- Most groups had three or fewer members; largest group had 12 members

Task 1: Citation Prediction Task

- The citation network evolves over time. Can we predict its trajectory?

- Goal: Predict changes in the number of citations to well-cited papers over time.
- Predict the *change* between
  - the number of citations that papers received during the period February 1, 2003 - April 30, 2003
  - the number of citations that papers received during the period May 1, 2003 - July 31, 2003
- Metric: Absolute difference between submitted change and the true change summed over all well-cited papers

Task 1: Data

- Uses subset of papers categorized as hep-th (High Energy Physics – Theory)
- 30,119 papers, 1.7GB of LaTeX sources
- 719,109 total citations
  - (363,812 external citations)
  - 355,297 internal citations
- 57,448 authors

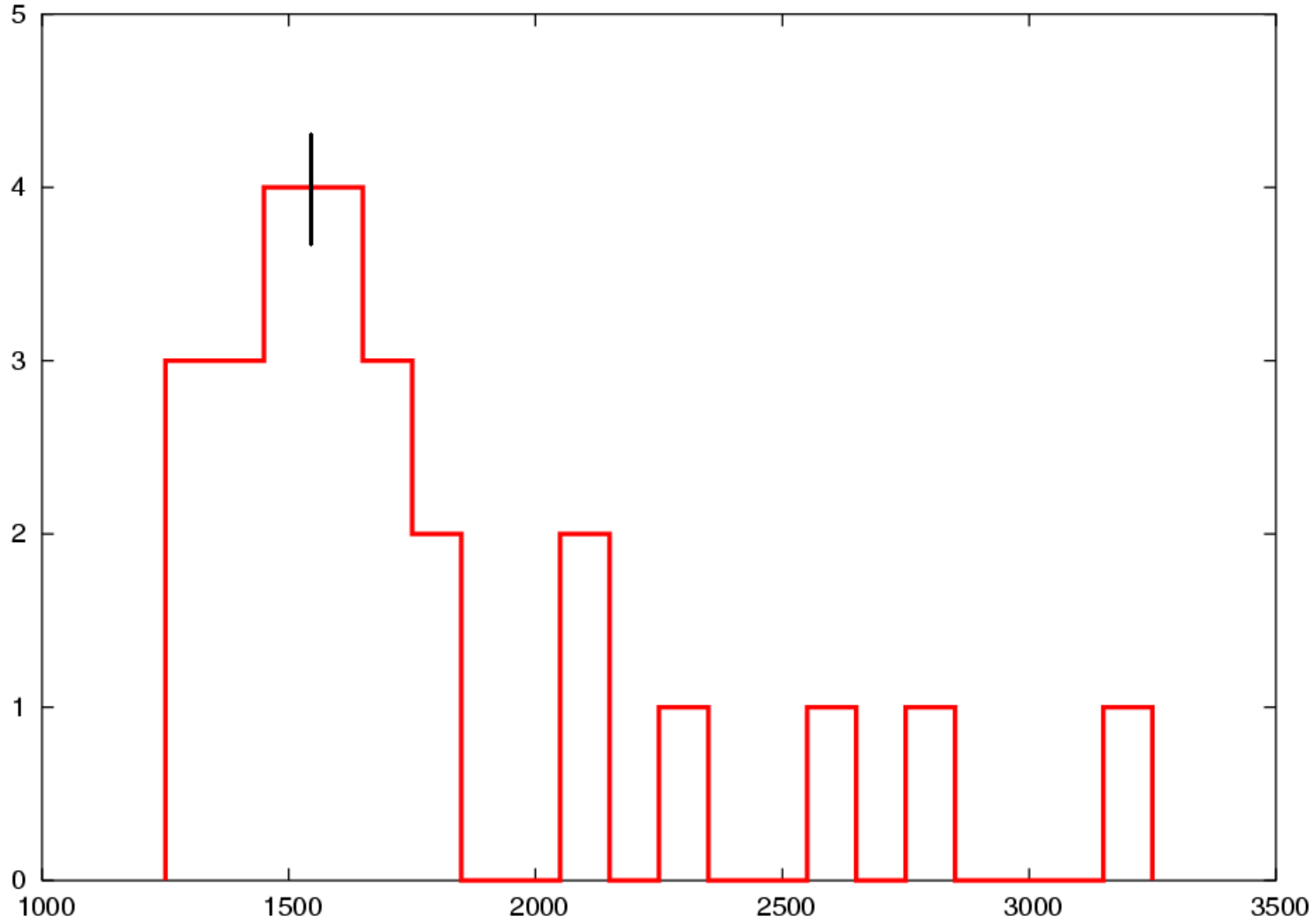
Task 1: Data (Contd.)

Available for contestants:

1. LaTeX sources of each paper
2. Separate user-submitted abstracts for each paper
3. The arXiv submission date
4. Hep-th citation graph from SLAC/SPIRES

Task 1: Histogram of Scores

All entries: bucketed by hundreds



Task 1: Winners

1. J N Manjunatha, Raghavendra Kumar Pandey, S R Sivaramakrishnan, Narasimha Murty
   (Indian Institute of Science)
2. Claudia Perlich, Foster Provost, Sofus Kacskassy
   (New York University)
3. David Vogel
   (A.I. Insight, Inc.)

Task 3: Download Estimation Task

- More generally, the arXiv evolves through usage by the physics community. Can we find patterns in this usage behavior?
- Goal: Estimate the number of downloads that a paper receives in its first two months in the arXiv.

Task 3: Download Estimation (Cont.)

Network structure and textual content (visible data) are inter-twined with usage (partially hidden)

- Growth of citations in sub-areas follows growth in download activity. (Citations as frozen evidence of usage…)
- Papers on similar topics tend to have similar download histories (e.g. due to topic-specific mailing lists)

Task 3: Download Data

Contestants were provided with:

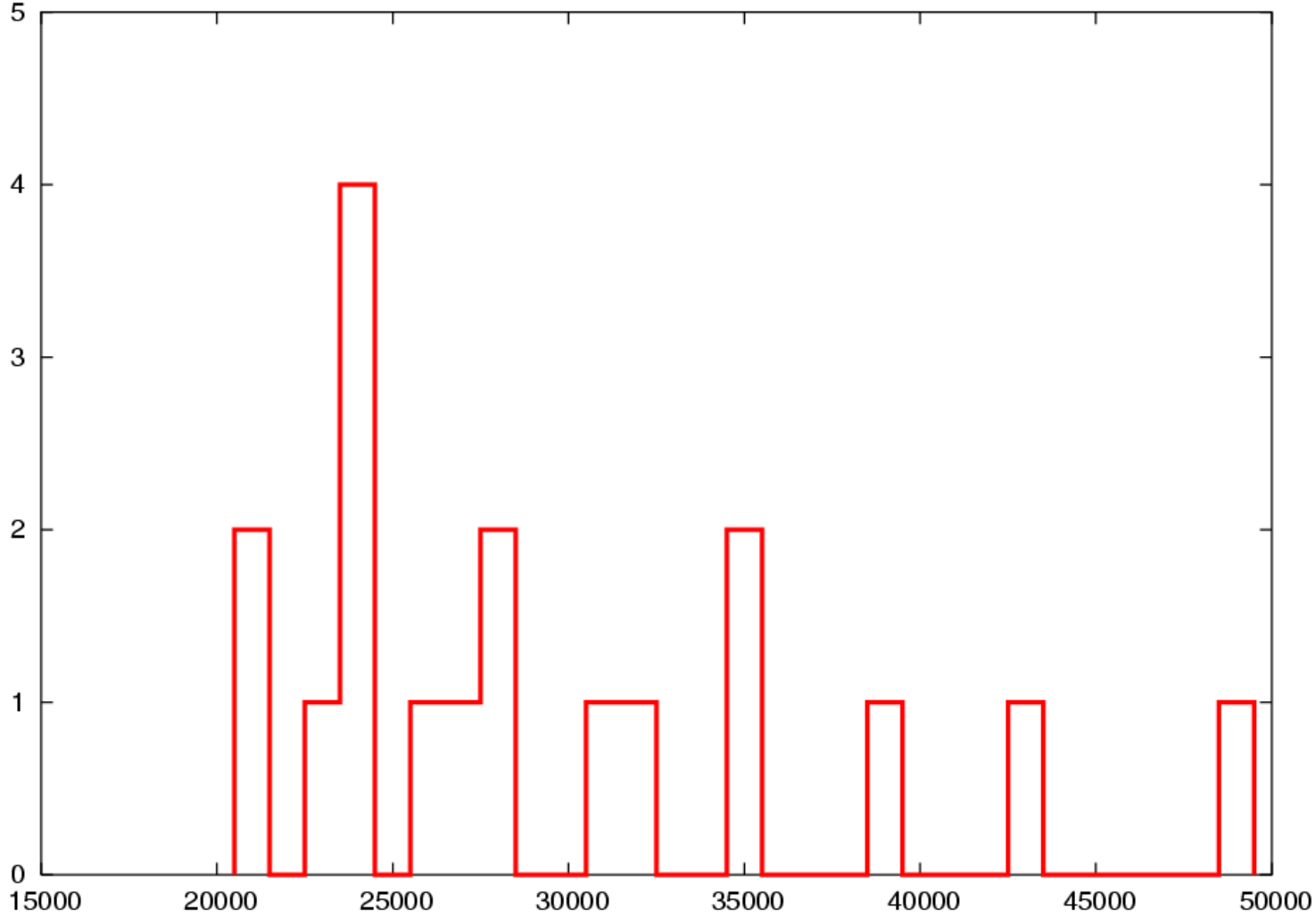
- For papers published in the following months, the downloads received from the main site in each of its first 60 days in the arXiv.
  - February and March of 2000
  - February and April of 2001
  - March and April of 2002

Contestants should submit estimates for April 2000, March 2001, February 2002.

Evaluation metric: Absolute difference with true download counts summed over top 50 papers from each period.

Task 3: Histogram of Scores

All entries: bucketed by thousands



CORNELL Data Mining

Task 3: Winners

- 1st place: Janez Brank, Jure Leskovec
  (Jozef Stefan Institute, Slovenia)
- 2nd place: Joseph Milana, Joseph Sirosh, Joel Carleton, Gabriela Surpi, Daragh Hartnett, Michinari Momma
  (Fair Isaac Corporation)
- 3rd place: Kohsuke Konishi
  (University of Tokyo, Japan)
  *Highest intersection with true top 150

Task 2: Data Cleaning Task

- Uses a category of papers called hep-ph (High Energy Physics – Phenomenology)
- SLAC/SPIRES provides a citation graph created through automated heuristics followed by human post-processing
- Unclean citations:
  - Spelling variations on author names
  - Abbreviations, typos, etc.
  - Citations in physics usually do not give the paper title
    - 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.

Task 2: Data Cleaning Task (Contd.)

- Contestants must recreate this citation graph using only the LaTeX sources
- Unique paper identifiers have been removed
  - Lisa Randall, Raman Sundrum, Phys.Rev.Lett. 83: 3370-3373, 1999, hep-ph/9905221.
- Evaluation metric: Size of symmetric difference between true and submitted sets of citation links
- SLAC/SPIRES's automated tools achieve reasonable accuracy:
  - Refined over many years
  - Based on extensive domain knowledge
  - Make use of unique identifiers

Task 2: Data Cleaning Task (Contd.)

- With limited time and no domain knowledge, problem was very difficult
- True citation graph has 421K edges over 35K papers
- Only one entry outperformed the empty graph on the evaluation metric
  - Consisted of four citations

Task 2: Winners

- 1st place: David Vogel
  (A.I. Insight, Inc.)
- 2nd place: Sunita Sarawagi, Kapil M. Bhudhia, Sumana Srinivasan, V.G. Vinod Vydiswaran
  (IIT Bombay)
  *40.6K correct out of 175.8K predicted
- 3rd place: Martine Cadot, Joseph di Martino
  (Loria)

Task 4: Open Task

- Take the data
- Define the most interesting question you can
- Mine the answer

- A committee of judges selected the winning entry based on novelty, soundness of methods and evaluation, and relevance to the arXiv dataset.
  - Committee: the three KDD-Cup co-chairs, Mark Craven, David Page, and Soumen Chakrabarti

Task 4: Winners

- 1st place: Amy McGovern, Lisa Friedland, Michael Hay, Brian Gallagher, Andrew Fast, Jennifer Neville, and David Jensen

  (University of Massachusetts Amherst)

  "Exploiting Relational Structure to Understand Publication Patterns in High-Energy Physics"

- 2nd place: Shou-de Lin, Hans Chalupsky

  (University of Southern California)

  "Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset"

- 3rd place: Shawndra Hill, Foster Provost

  (New York University)

  "The Myth of the Double-Blind Review "

Acknowledgements

- Manuel Calimlim (Cornell)
- Travis Brooks (SLAC/SPIRES)
- Soumen Chakrabarti (IIT Bombay), Mark Craven (U. Wisconsin), David Page (U. Wisconsin)
- We thank Pedro Domingos, Christos Faloutsos, and Ted Senator for valuable suggestions

More information:

http://www.cs.cornell.edu/projects/kddcup

kddcup2003@cs.cornell.edu

CORNELL Data Mining

Award Talks

Task 1:

Claudia Perlich, Foster Provost, Sofus Kacskassy (New York University)

Task 3:

Janez Brank and Jure Leskovec (Jozef Stefan Institute)

Task 4:

Amy McGovern, Lisa Friedland, Michael Hay, Brian Gallagher, Andrew Fast, Jennifer Neville, and David Jensen (University of Massachusetts Amherst)