

The Case for Anomalous Link Detection

Matthew J. Rattigan, David Jensen
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003-9264 USA
[rattigan,jensen]@cs.umass.edu

ABSTRACT

In this paper, we describe the challenges inherent to the Link Prediction (LP) problem in multirelational data mining, and explore the reasons why many LP models have performed poorly. We present the alternate (and complimentary) task of Anomalous Link Discovery (ALD) and qualitatively demonstrate the effectiveness of simple LP models for the ALD task.

1. INTRODUCTION

Modeling the link structure of relational data is one of the key challenges in multirelational data mining (MRDM). To date, numerous link mining efforts have focused the problem of Link Prediction (LP): given a dynamic graph representing objects and their relationships, we aim to predict which new relationships will appear in the near future.[5][8] A fundamental challenge in link prediction is a highly skewed class distribution --- as networks grow and evolve, the number of negative examples (pairs of objects bearing no direct relationship) increases quadratically in the number of positive examples. Thus, the evaluation of a link predictor is hampered by the computational cost of evaluating all possible pairs of objects for which a link exists. Furthermore, a highly skewed class distribution increases the variance of the link predictor. As a result, most attempts at accurate link prediction have proven unsuccessful. Below, we propose a shifting of focus to the complimentary task of Anomalous Link Discovery (ALD). In this task, we seek to identify the links within a data set that are anomalous (statistically unlikely), which are in many cases the most “interesting” links in the data.

1.1 Link Prediction

Probabilistic models of link structure have become common in the field of statistical relational learning.[2] Most current efforts focus on link prediction. Many algorithms for attacking this problem approach it in the form of a binary classification task in which *pairs* of objects in the data comprise the instances. Given any pair of objects, we can assess the probability of a new link appearing between them at some point in the future.

This paper appears in the Proceedings of the Fourth International Workshop on Multi-Relational Data Mining (MRDM-2005), August 21, 2005, Chicago. The proceedings were edited by Saso Dzeroski and Hendrik Blockeel. The paper is published here with the permission of the authors who retain the copyright of this material.

While this approach seems straightforward, its combinatorics pose some formidable challenges. The number of possible links is quadratic in the number of objects, but in many domains where link prediction is employed, the number of actual links added to the graph in any time period is only a tiny fraction of this number. The result is that the link models face a massive class skew, making learning difficult.[3] Figure 1 illustrates the problem for data from the Digital Bibliography & Library Project (DBLP) over a ten year span. Over this time period, the number of authors increases by an order of magnitude, from 22 thousand to 286 thousand. The number of collaborative papers, though, only increases by a factor of 3.4, while the number of possible collaborations simultaneously increases by a factor of 164.

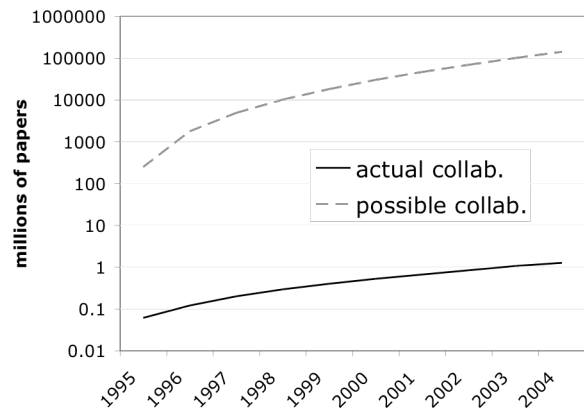


Figure 1. Logarithmic plot of actual and possible collaborations between DBLP authors, 1995-2004.

Figure 2 depicts the proportion of possible collaborations between authors that occur during this same time period. This ratio represents the class distribution for the pairwise LP problem. By 2004, less than one one-thousandth of one percent of DBLP author pairs have written a paper together. In the face of such an uneven class distribution, even the most accurate binary classifiers can be woefully inaccurate at predicting link structure to any reasonable degree.[5] Most of the error can be attributed to variance within the data, as the sheer number of negative instances (pairs of objects who will not share a new link) compared to the positive ones will destroy the model's ability to find meaningful differences between the positive and negative classes.

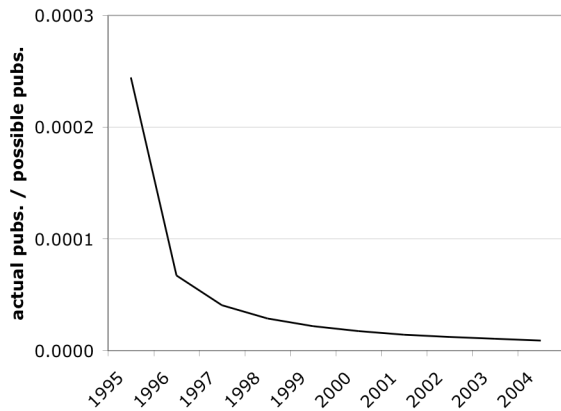


Figure 2. Publications of DBLP authors as a proportion of possible collaborations, 1995-2004.

We illustrate the problem in the abstract in Figure 3 below. Given a hypothetical data set and some predictive statistic s that is measured on each instance pair. Since the statistic is predictive of link structure, we assume that the values of s are drawn from separate distributions for our two classes (linked and non-linked object pairs), illustrated in (a) as Normal distributions with differing means. In the face of massive class skew, however, the entirety of the positive class distribution gets “swallowed” by the tail of the negative class distribution, as seen in (b).

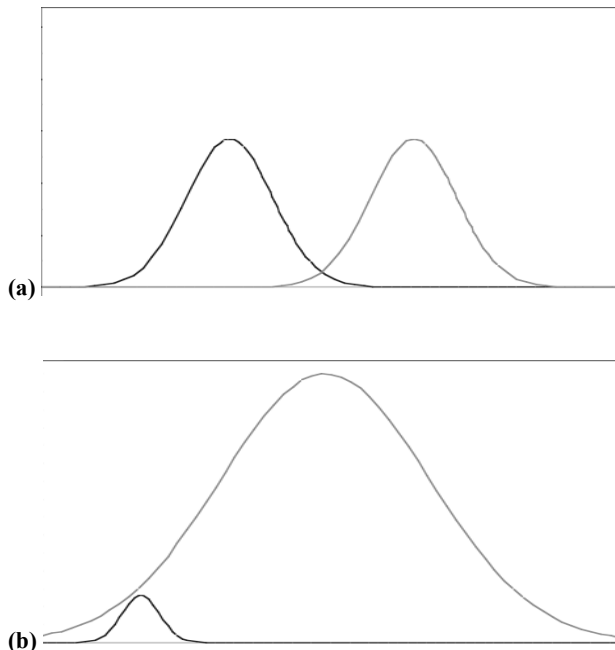


Figure 3. Abstract representation of the effects of massive class skew on a model’s ability to discriminate between classes. In the first case (a), the two classes define two easily distinguished sampling distributions. Below (b), we see the effects of massive class skew on our ability to differentiate.

In the latter case, it is virtually impossible to discriminate between the two classes by examining values of s . Furthermore, any attempts to overcome the variance issues will necessarily introduce additional bias to the classifier.

As an example, we shall examine the concrete example of Professor (and MRDM program committee member) Jiawei Han from UIUC. We took a snapshot of the DBLP database as it existed in 2002, and identified 39 thousand “core authors” (authors who had several publications in the eight preceding years). In the standard LP task, our goal is to predict which authors Prof. Han will collaborate with in the future (2003-2004 in our example). To discriminate between positive coauthor pairs (those that Prof. Han does write a paper with) and negative pairs, we shall use the Katz measure, which has been shown to be (relatively) effective in predicting links.[5] The relative frequency distributions for each class can be seen in Figure 4. Clearly, the values of the Katz measure for Prof. Han’s actual coauthors are drawn from a different distribution than the author population at large. This would lead us to believe that predicting links is not difficult.

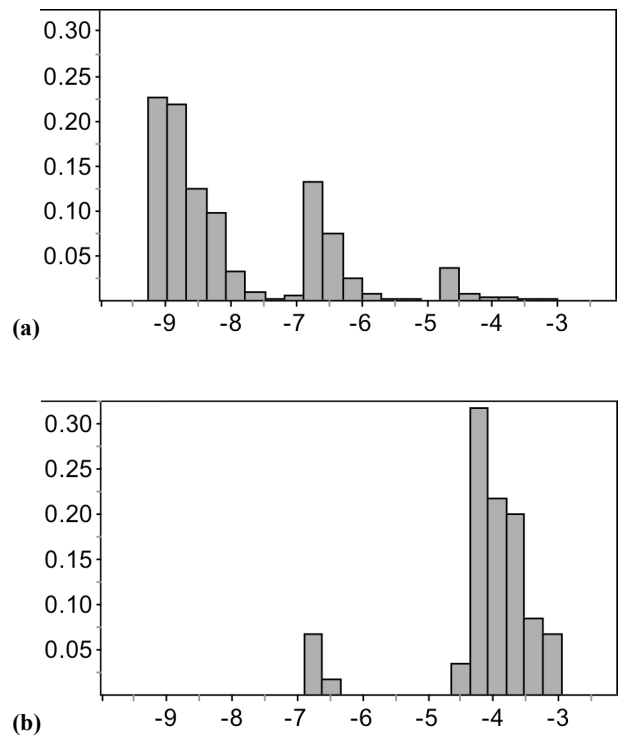


Figure 4. Relative frequency distributions of the log-Katz statistic for author pairs that include Jiawei Han. The distribution for all possible coauthors is shown in (a), the distribution for actual coauthors is in (b).

The histograms in Figure 5 tell a different story, however. Here, we see the effects of the massive class skew. The sampling distributions depicted are the same as those shown in Figure 4, but plotted with actual frequency counts rather than relative ones. As Prof. Han collaborates with 63 different core authors in 2003-

2004 (out of a possible 39 thousand!), it becomes impossible to differentiate the positive examples from the pairs in the tail end of the “random” distribution. Moreover, the problem is actually four times worse than depicted, as computational constraints allowed us to only calculate Katz measures for ten thousand possible coauthors. Thus a Katz-based ranking classifier that draws a classification threshold generous enough to capture even half of the actual positives will misclassify hundreds of negative pairs.

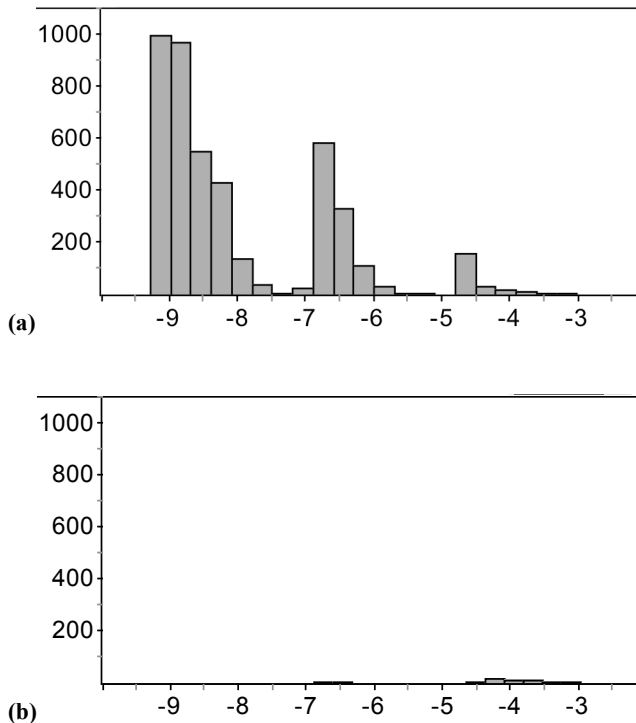


Figure 5. Actual frequency distributions of the log-Katz statistic for negative Jiawei Han author pairs (a) and positive author pairs (b).

The results of these effects are models of questionable utility. The vast majority of new links are not correctly predicted, and the number of false negatives is quite substantial.[5] In addition, the links that are predicted correctly tend to be the “obvious” ones, those that are out of the “reach” of the tail of the massive negative sampling distribution.

Finally, this sort of pairwise link prediction is often computationally infeasible.[7] Given new objects in the graph, it may not be possible to efficiently consider their link probabilities with all other objects. Again, attempts to restrict the search introduce a bias that may eliminate the possibility of predicting “interesting” links.

Given the issues described above, the LP task is a daunting undertaking. Furthermore, due to the challenging combinatorics, it may very well be *impossible* to satisfactorily address LP as currently specified. In the face of these facts, what is the aspiring link modeler to do? A potential avenue may be to refocus efforts structure learning away from existential link prediction and toward examining models of existing link properties —

specifically, identifying anomalous links that do appear in the data. As we will see below, by respecifying the task being addressed from LP to anomalous link discovery (ALD), we can leverage existing algorithms to gain insight into the structure of graphs.

1.2 Anomaly Detection

In some ways, anomaly detection can be seen as the most basic of tasks knowledge discovery. The goal is to identify the “outliers” in a data set. In many contexts, we can often find the thing that “doesn’t belong” without having to know what we are looking for a priori, thus we do not have to specify anything about it. In traditional IID data mining, anomaly detection algorithms are most often employed in domains that deal with security issues: can we pick out the suspicious login session at a computer, a strange financial transaction, or a person who appears “suspicious”?

In the case of multirelational links, anomaly detection involves examining the links that do appear in the data and modeling their likelihood. MRDM techniques seem especially suited to handle the anomaly detection problem, as the use of structured data lends itself to a of host possible methods for finding interesting instances in a data set. As we have seen, the interesting facets of our data may be defined by the relations between entities as well as the intrinsic properties of the entities themselves. Yet despite this seemingly natural match of tasks and techniques, anomaly detection tasks have been all but ignored in the MRDM literature.[6]

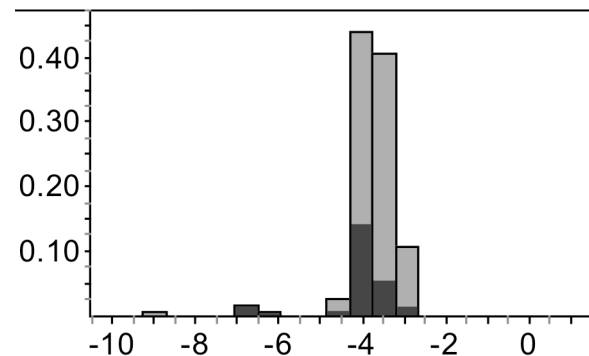


Figure 6. Relative frequency distribution of the log-Katz measure for Jiawei Han's collaborations, 1995-2004. The shaded region represent the collaborations from the final two years.

Furthermore, in many domains, identifying anomalous links may actually be preferable to predicting links. For instance, bibliometric data are often the target of LP.[2] However, is it more useful to “predict” coauthorships and citation, or to realize when an interesting (or unlikely) collaboration occurs? In the case of internet applications, is predicting which web pages will potentially link to each other very helpful? Or would we rather be able to comb the links that do exist in reality and be alerted when surprising connections are made? Obviously, the utility of ALD algorithms is entirely domain (and representation) dependent, but even a cursory examination of the literature opens up several possible applications. Lastly, in the ALD task, we don’t waste

computation time reasoning through the multitudes of negative examples, so our models scale nicely with the size of the data, unaffected by an ever-widening class distribution gap.

Returning to the example of Prof. Han, in the ALD context we examine the properties of the links that actually do appear in 2003-2004, and see how they measure up statistically with past collaborations. Figure 6 shows the distribution of the Katz measure for all of Prof. Han’s collaborations with core authors dating back to 1995. Using this distribution, we can calculate the likelihood of each the papers written.

On one extreme, we have a collaboration between Jiawei Han and Jian Pei of SUNY Buffalo, which was measured to be the most *likely* collaboration involving not just Prof. Han, but any MRDM program committee member, with a Katz score of 6.96×10^{-4} . An examination of the authors’ home pages bears this out --- the two have written dozens of papers together over the years, thus their recent joint work is not surprising to anyone.

Not all of Dr. Han’s collaborations are that predictable, however. During the same year, he also authored a paper with Martin Ester of Simon Fraser University, one of the (according to the Katz measure) more unlikely collaborations with a score of 4.3×10^{-7} . Both people are accomplished scientists, yet until last year they were only connected in the authorship graph by paths of at least length four.

2. PRELIMINARY RESULTS

To test the validity of our intuitions concerning the use of LP techniques for anomaly detection, we examined the author graph drawn from DBLP data. In our representation, we have a single object type, representing authors. Journal and conference papers are represented as links that connect the authors together through coauthorships. Thus a single paper with four authors will spawn ten different items in the graph: a node for each author, and six links expressing the pairwise coauthorships between them. A visual representation of the relational schema can be seen in Figure 7.

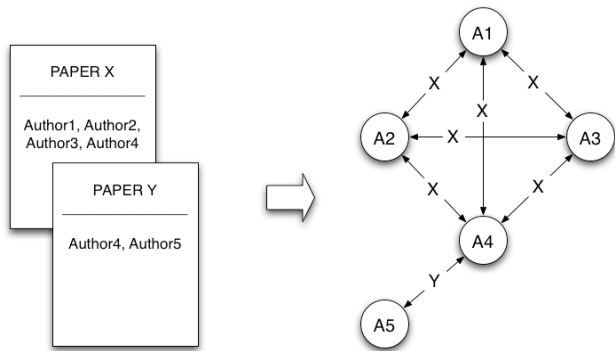


Figure 7. Relational schema DBLP data: authors are represented by author that are linked through co-authored papers.

We tested several metrics identified in the LP literature as being useful at predicting links. The histograms in Figure 8 illustrate the “discriminative ability” for three of these measures. In each

case, the sampling distribution for negative (unlinked) pairs is shown directly above the distribution of pairs that contain links.

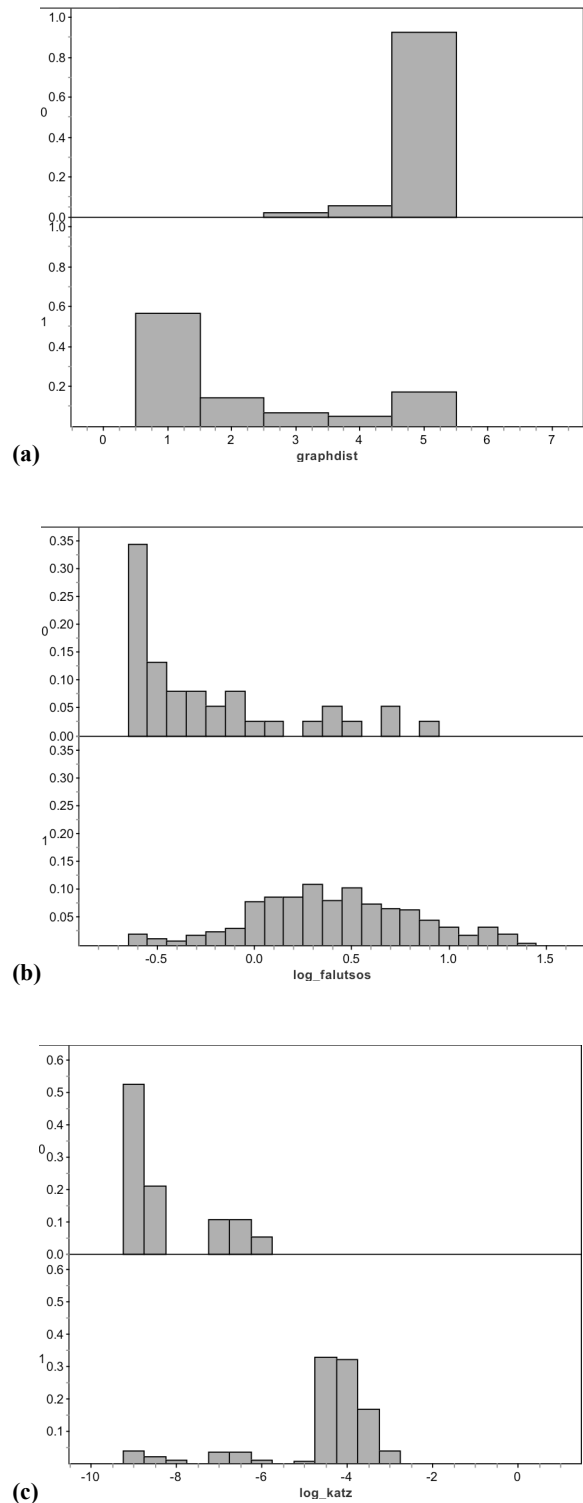


Figure 8. Sampling distributions for linked (bottom) and unlinked (top) author pairs for three measures of link likelihood: graph distance (a), Falutsos (b), and Katz (c).

The first statistic shown is simple graph distance between the objects in the pair. While the linked and unlinked values are clearly drawn from different distributions, there is still a great deal of overlap between the two. The second measure depicted, called Falutsos, is a measure of the maximum deliverable current when the nodes in the subgraph connecting the start and end nodes are treated like wired resistors in an electric circuit.[1] A quick comparison judges this measure to be more useful than simple graph distance. Lastly, we show the positive and negative sampling distributions for the Katz measure as defined in [5] and used in the examples in the previous section. The Katz measure is a weighted sum of the number of paths in the graph that connect two nodes, with shorter paths being given the most weight. Here we have clear separation between the bulk of each distribution. Regardless, for the reasons described previously, even the Katz measure is mostly ineffective at predicting links.

Of great interest, however, is the leftward tail of the Katz distribution for positive pairs. This tells us that a small number of new links in the graph appear in structural contexts that appear to be random. It is these links (the statistically unlikely ones) that deserve our interest in the DBLP domain --- they represent collaborations between authors from vastly different spheres of the academy.

To verify the ability of the Katz measure to help us identify links that are non-qualitatively anomalous (i.e., entirely random), we conducted a synthetic experiment. We inserted several “artificial” links into the data set, and calculated the Katz score for each one. Figure 9 shows where these scores lie in the distribution of actual links (shaded portions of the bars in the histogram represent the synthetic links). This result is about what we expected. Certainly, the Katz values for the artificial links are grouped in the tail of the overall distribution. However, given the presence of the “interesting” links as described above, it is understandable that these actual anomalies could be confused with the apparent ones.

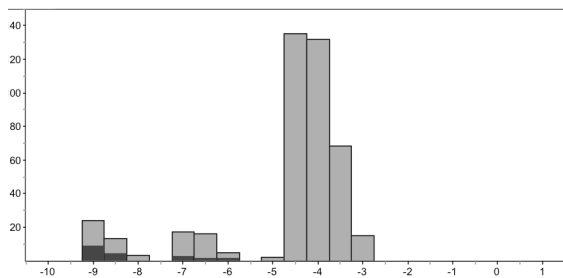


Figure 9. Distribution of log-Katz scores for pairs in DBLP. Artificial links are represented by the shaded regions.

Finally, we conducted a qualitative examination of a subset of the dataset to test our intuition about the usefulness of the Katz measure to find “interesting” coauthor pairs. From the DBLP data set, we considered the 203 papers (representing 450 pairwise collaborations) written by members of the MRDM program committee in 2003 and 2004 with another author, and calculated the Katz score for each pairwise collaboration. The overall distribution of Katz scores for the group is shown in Figure 10. While the bulk of the distribution of pairs has a score between 0.001 and 0.0001 (note that the histogram is plotted on a log scale), the distribution has a significant tail. It is these

collaborations that are often the most interesting, as they are statistically indistinguishable from random pairings in the data. Likewise, pairs with high Katz scores represent links that are not unexpected given the existing structure in the data.

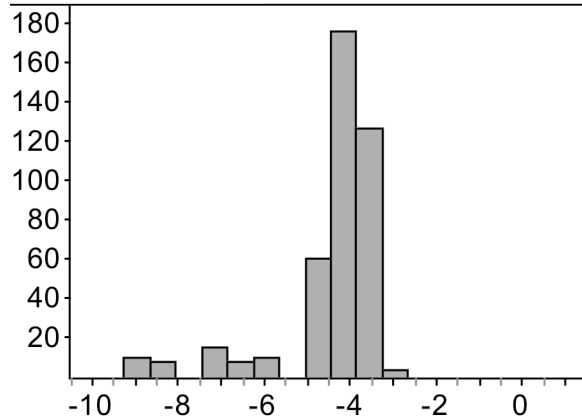


Figure 10. Frequency distribution of log-Katz scores for collaborations involving MRDM program committee members in 2003-2004.

Below we present lists of the ten most “likely” and “unlikely” papers, as ranked by the mean pairwise Katz score for all pairs of authors of the paper. Note that some authors may be omitted, as the data were drawn from the arbitrarily designated pool of “core” authors. In addition, multiple papers written by the same group were disregarded for the sake of interest.

Top 10 Most Unlikely Papers of 2003-2004

1. Jean-Francois Boulicaut, Celine Robardet. Constraint-Based Mining Of Formal Concepts In Transactional Data. 2004
2. Thomas Gartner, Stefan Eicker. Einsatz Virtueller Computerpools Im E-Learning. 2003.
3. Jiawei Han, Michael Welge, David Clutter. MAIDS: Mining Alarming Incidents From Data Streams. 2004.
4. Jiawei Han, Michael Garland. Mining Scale-Free Networks Using Geodesic Clustering. 2004.
5. Peter A Flach, Annalisa Appice, Michelangelo Ceci. Redundant Feature Elimination For Multi-Class Problems. 2004.
6. Kristian Kersting, Jorg Fischer. Scaled CGEM: A Fast Accelerated EM. 2003.
7. Jean-Francois Boulicaut, Celine Robardet. Using Classification And Visualization On Pattern Databases For Gene Expression Data Analysis. 2004.
8. Jean-Francois Boulicaut, Bruno Cremilleux. Using Transposition For Pattern Discovery From Microarray Data. 2003.
9. Foster J Provost, Raymond J Mooney, Prem Melville. Active Feature-Value Acquisition For Classifier Induction. 2004.
10. Hiroshi Motoda, Kouzou Ohara, Noboru Babaguchi. Constructive Inductive Learning Based On Meta-Attributes. 2004.

Top 10 Most Likely Papers of 2003-2004

1. Jiawei Han, Jian Pei, Jianyong Wang. CLOSET+: Searching For The Best Strategies For Mining Frequent Closed Itemsets. 2003.

2. Takashi Washio, Hiroshi Motoda. State Of The Art Of Graph-Based Data Mining. 2003.
3. Jiawei Han, Joyce M W Lam, Guozhu Dong, Ke Wang, Jian Pei. Mining Constrained Gradients In Large Databases. 2004.
4. Donato Malerba, Michelangelo Cec, Floriana Esposito. Learning Logic Programs For Layout Analysis Correction. 2003.
5. Hendrik Blockeel, Saso Dzeroski. First Order Random Forests With Complex Aggregates. 2004.
6. Saso Dzeroski, Ljupco Todorovski. Using Domain Specific Knowledge For Automated Modeling. 2003.
7. Donato Malerba, Oronzo Altamura, Teresa Maria Altomare Basile, Nicola Di Mauro, Stefano Ferilli, Michelangelo Ceci, Giovanni Semeraro, Floriana Esposito. Machine Learning Methods For Automatically Processing Historical Documents: From Paper Acquisition To XML Transformation. 2004.
8. Jiawei Han, Ling Feng, Anthony K H Tung, Hongjun Lu. Efficient Mining Of Intertransaction Association Rules. 2003.
9. Raghu Ramakrishnan, Raghav Kaushik, Rajasekar Krishnamurthy, Jeffrey F Naughton. On The Integration Of Structure Indexes And Inverted Lists. 2004.
10. Ashwin Srinivasan, Ross D King. An Empirical Study Of The Use Of Relevance Information In Inductive Logic Programming. 2003.

A cursory inspection of these results confirms our intuitions about the value of the Katz score in determining link likelihood. The papers on the “unlikely” list have author lists similar in character to the Han-Pei collaboration discussed in the previous section, made up of established researchers from different “relational neighborhoods” within the graph. Omitted from this list is an extremely unlikely paper on “voltage scheduling” written by one of the authors of the paper you are reading now --- so unlikely, in fact, that said author does not recall writing it. As it turns out, this lack of recollection is entirely warranted, as there are two computer scientists with the name “David Jensen”, and our version of the DBLP database mistakenly combined them.

Conversely, the papers on the second list are written by prolific authors connected by shared work and coauthors. Note, for instance, the presence on the list of the two MRDM 2005 chairs, as well as teams of authors who have collaborated dozens of times in the literature.

3. RELATED AND FUTURE WORK

The vast majority of work dealing with probabilistic models of link structure addresses the problem of link prediction or detection ([5], [7], [8] for example). In fact, ALD work grew out of an effort to replicate the work of Liben-Nowell and Kleinberg on the DBLP dataset. To our knowledge, though, the simpler problem of anomaly detection has gone largely ignored. Certainly, ALD could be cast as one of several challenges in structure learning identified by Getoor [2], such as link type prediction (if we consider existential status as a type), link cardinality prediction, or existence and reference uncertainty. The work on “rarity analysis” in [6] seems to be closest in spirit to ALD in terms of motivation and approach, though it focuses on identifying statistically unlikely paths through the data graph rather than individual links. Certainly, an obvious extension to ALD would involve assessing the likelihood of paths as a function of the likelihoods of the individual links that comprise it.

While our work on using LP models to perform ALD is very preliminary, the results so far are encouraging. Further examination and testing on additional data sets (both real and synthetic) is necessary before we can gain a real understanding of how these models work. Further down the road, we would like to learn link models that combine a number of statistics (Katz measure, Falutsos measure, etc.) in order to identify interesting links rather than rely on a single measure.

Regardless of method, though, it is clear that more effort needs to be spent on the anomaly detection problem in multirelational data mining in general, as the problem gets at the heart of what defines MRDM a unique field --- the relationships encoded in the structure of the data.

4. ACKNOWLEDGEMENTS

The authors wish acknowledge the helpful input of Ross M. Fairgrieve.

This research is supported by DARPA and LLNL/DOE Lawrence Livermore National Laboratory and the Department of Energy under contract numbers HR0011-04-1-0013 and W7405-ENG-48. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements either expressed or implied, of DARPA, LLNL/DOE Lawrence Livermore National Laboratory and the Department of Energy, or the U.S. Government.

5. REFERENCES

- [1] C. Faloutsos, K. McCurley and A. Tomkins. Fast Discovery of Connection Subgraphs. *Proc. 10th ACM Conference on Knowledge Discovery and Data Mining*, 2004.
- [2] L. Getoor. Link Mining: A New Data Mining Challenge. *SIGKDD Explorations*, volume 5, issue 1, 2003.
- [3] D. Jensen, M. Rattigan, and H. Blau. Information awareness: A prospective technical assessment. *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2003.
- [4] W. Lee and S. Stolfo. Data Mining Approaches for Intrusion Detection. *Proceedings of the Seventh USENIX Security Symposium (SECURITY '98)*, 1998.
- [5] D. Liben-Nowell, J. Kleinberg. The Link Prediction Problem for Social Networks. *Proc. 12th International Conference on Information and Knowledge Management (CIKM)*, 2003.
- [6] S. Lin and H. Chalupsky. Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis. *Proceedings of the Third IEEE International Conference on Data Mining*, 2003.
- [7] R. Mooney, P. Mellville, L. Tang, J. Shavlik, I. Dutra, D. Page, V. Costa. Relational Data Mining with Inductive Logic Programming for Link Discovery. *Proceedings of the National Science Foundation Workshop on Next Generation Data Mining*, 2002.
- [8] A. Popescul and L. Ungar. Statistical Relational Learning for Link Prediction. Workshop on Learning Statistical Models from Relational Data at IJCAI 2003.