AFRL-IF-RS-TR-2006-222
Final Technical Report
June 2006

SCALABLE DETECTION AND OPTIMIZATION OF N-ARY LINKAGES

Carnegie Mellon University

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

STINFO FINAL REPORT

This report has been reviewed by the Air Force Research Laboratory, Information Directorate, Public Affairs Office (IFOIPA) and is releasable to the National Technical Information Service (NTIS). At NTIS it will be releasable to the general public, including foreign nations.

AFRL-IF-RS-TR-2006-222 has been reviewed and is approved for publication.

APPROVED: /s/

ROBERT L. HAWKINS
Project Engineer

FOR THE DIRECTOR: /s/

JOSEPH CAMERA
Chief, Information & Intelligence Exploitation Division
Information Directorate

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JUN 06	Final	Sep 01 – Feb 06

4. TITLE AND SUBTITLE
SCALABLE DETECTION AND OPTIMIZATION OF N-ARY LINKAGES

5a. CONTRACT NUMBER
F30602-01-2-0569

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
31011G

6. AUTHOR(S)
Andrew Moore, Jeff Schneider, Jeremy Kubica, Anna Goldenberg, Artur Dubrawski, John Ostlund, Patrick Choi, Jeanie Komarek, Adam Goode, Purna Sarkar

5d. PROJECT NUMBER
EELD

5e. TASK NUMBER
01

5f. WORK UNIT NUMBER
15

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh Pennsylvania 15213-3890

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/IFED
525 Brooks Rd
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-222

12. DISTRIBUTION AVAILABILITY STATEMENT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA #06-483

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Link detection and analysis has long been important in the social sciences where a single link can be the key evidence that leads an intelligence analyst to additional clues to a threat event. A significant effort is focused on the structural and functional analysis of "known" networks. Similarly, the detection of individual links is important but is usually done with techniques that result in "known" links. More recently, the internet and other sources have led to a flood of circumstantial data that provide probabilistic evidence of links. Co-occurrence in news articles and simultaneous travels to the same location are two examples. We propose a probabilistic model of link generation based on membership in groups. The model considers both observed link evidence and demographic information about the entities. The parameters of the model are learned via a maximum likelihood search. In this paper, we describe the model and then show several heuristics that make the search tractable. We test our model and optimization methods on synthetic data sets with a known ground truth and a database of news articles.

15. SUBJECT TERMS
Group Detection, Bayes classifier, probabilistic model, search algorithms, link generation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Robert L. Hawkins
U	U	U	UL	12	19b. TELEPONE NUMBER (*Include area code)*

TABLE OF CONTENTS

1 Algorithms invented and evaluated

1.1 Group Detection Algorithm (GDA)

Link detection and analysis has long been important in the social sciences and in the government intelligence community. A significant effort is focused on the structural and functional analysis of "known" networks. Similarly, the detection of individual links is important but is usually done with techniques that result in "known" links. More recently the internet and other sources have led to a flood of circumstantial data that provide probabilistic evidence of links. Co-occurrence in news articles and simultaneous travel to the same location are two examples. We designed and implemented a probabilistic model (GDA) of link generation based on membership in groups. The model considers both observed link evidence and demographic information about the entities. The parameters of the model are learned via a maximum likelihood search. We described the model and investigated several heuristics that make the search tractable. We tested our model and optimization methods on synthetic data sets with a known ground truth and many databases of links, including links derived from automatic entity extraction from news articles and manual extraction of 5000 links from public web pages about terrorism.

1.2 Link Completion Methods

Link data, consisting of a collection of subsets of entities, can be an important source of information for a variety of fields including the social sciences, biology, criminology, and business intelligence. However, these links may be incomplete, containing one or more unknown members. We worked on the problem of link completion, identifying which entities are the most likely missing members of a link given the previously observed links. We concentrated on the case of one missing entity. We compared a variety of recently developed along with standard machine learning and strawman algorithms adjusted to suit the task. These included GDA, Logistic Regression, k-nearest-neighbor, cGraph, Bayesian Networks along with several graph-theoretic methods. The algorithms were tested extensively on a simulated and a range of real-world data sets.

1.3 K-groups (Computationally tractable, approximate GDA)

Discovering underlying structure from co-occurrence data is an important task in a variety of fields, including: insurance, intelligence, criminal investigation, epidemiology, human resources, and marketing. We extended our Group Detection Algorithm (GDA) - an algorithm for finding underlying groupings of entities from co-occurrence data. GDA is based on a probabilistic generative model and produces coherent groups that are consistent with prior knowledge. Unfortunately, the optimization used in GDA is slow, potentially making it infeasible for many large data sets. To this end, we created k-groups - an algorithm that uses an approach similar to that of k-means to significantly accelerate the discovery of groups while retaining GDA's probabilistic model. We compared the performance of GDA and k-groups on a variety of data, showing that k-groups' sacrifice in solution quality is significantly offset by its increase in speed.

1.4 cGraph (fast connection-finding)

Many techniques in the social sciences and graph theory deal with the problem of examining and analyzing patterns found in the underlying structure and associations of a group of entities. However, much of this work assumes that this underlying structure is known or can easily be inferred from data, which may often be an unrealistic assumption for many real-world problems. In this work we considered the problem of learning and querying a graph-based model of this underlying structure. The model is learned from noisy observations linking sets of entities. We explicitly allow different types of links (representing different types of relations) and temporal information indicating when a link was observed. We quantitatively compare this representation and learning method against other algorithms on the task of predicting future links and new "friendships" in a variety of real world data sets.

1.5 Alias Detection

The problem of detecting aliases - multiple text string identifiers corresponding to the same entity - is increasingly important in the intelligence domain. We investigated the extent to which probabilistic methods can help. Aliases arise from entities who are trying to hide their identities, from a person with multiple names, or from words which are unintentionally or even intentionally misspelled. While purely orthographic methods (e.g. string similarity) can help solve unintentional spelling cases, many types of alias (including those adopted with malicious intent) can fool these methods. However, if an entity has a changed name in some context, several or all of the set of other entities with which it has relationships can remain stable. Thus, the local social network can be exploited by using the relationships as semantic information. The active learning algorithm we proposed takes advantage of both orthographic and contextual information to detect aliases. By applying the best combination of both types of information, the algorithm outperforms the ones built solely on one type of information or the other. Two entities have high context similarity if they tend to have similar groups of friends. The algorithm learns from training data how best to combine many sources of similarity information Empirical results on three real world data sets support this claim. Further details can be found in the paper “Alias Detection in Link Data Sets” downloadable from www.autonlab.org.

1.6 Sparse Bayesian Network Search

Probabilistic, Bayesian modeling of social networks is a very attractive, but technically difficult problem, especially if the underlying networks are large. SBNS approach addresses three questions. Is it useful to attempt to learn a Bayesian network structure with hundreds of thousands of nodes? How should such structure search proceed practically? The third question arises out of our approach to the second: how can Frequent Sets, which are extremely popular in the area of descriptive data mining, be turned into a probabilistic model? Large sparse datasets with hundreds of thousands of records and attributes appear in social networks, warehousing, supermarket transactions, web logs as well as intelligence applications. Traditionally, the complexity of structural search makes learning of factored probabilistic models on such datasets unfeasible. To overcome that, we successfully used Frequent Sets paradigm to significantly speed up the

structural learning. Unlike previous approaches, we not only cache n-way sufficient statistics, but also exploit their local structure. Empirical evaluation of our algorithm applied to several massive datasets reveals practical utility of the approach to tackle real-world sized problems. SBNS takes noisy links, involving two or more entities per link, and searches for significant dependency structure that explains the observed n-ary links. It has been applied to datasets with 100 times more entities than previously published Bayesian Network search. Model allows nonlinear interactions between entities, even including negative interactions (e.g. models "X and Y are related indirectly, but to a significant extent X showing up makes it less likely that Y will show up"). Model is Scalable to O (10^5) entities and links. More details can be found in a paper titled "Tractable Learning of Large Bayes Net Structures from Sparse Data", available at www.autonlab.org.

1.7 Activity from Demographics and Links

AFDL algorithm takes into account three sets of data: (1) Noisy links, involving two or more entities per link, (2) Available entity-specific "demographics" information and (3) Information of which entities are known threats. The first step involves using the link data to derive numerical graph-based characteristics of each entity, such as the likelihoods of reaching a known threat entity in a given number of steps. These features are joined with demographic data of the entities, and the combined data is used to build a probabilistic classification model which then could be used to predict the likelihood of each unlabeled entity to be a threat entity. The algorithm returns a rank-ordered list of the originally unlabeled entities according to their predicted threat level. AFDL uses massive, scalable log-linear approach to construct the classification model. It has been tested on all the Challenge 2004 datasets and it appears to get greater accuracy than any other EAGLE participant in the classification of threats.

1.8 XGDA

The need for time-critical analysis and understanding of the underlying group structure from transactional data has been growing in domains such as law enforcement and intelligence. In the context of such applications we proposed to use GDA and then k-groups algorithms described above. Even though k-groups is reported to be significantly faster than its predecessor GDA, k-groups was too slow and memory-intensive for large data in practice. In order to resolve the issue we proposed XGDA, a framework for scalable and robust group discovery. XGDA combines ideas of k-groups and a novel clustering method called X-means. X-means draws from the popular K-means. K-means is a well-used and effective method for finding clusters in clouds of data. It is not directly applicable to the link grouping problem, but in earlier work we developed data structures and algorithms for making it extremely fast. Building on prior work for algorithmic acceleration that is not based on approximation, X-means efficiently searches the space of cluster locations and number of clusters to optimize the Bayesian Information Criterion (BIC) or the Akaike Information Criterion (AIC) measure. Using the X-means idea we extended k-groups towards XGDA, lifting the assumption that the number of k groups is known in advance. Evaluation of the performances of XGDA and k-groups shows that XGDA can handle extremely large datasets in a reasonable time (10-50 fold

acceleration has been reported w.r.t. the GDA) and it yields more robust solutions than k-groups (while automatically selecting the optimal number of groups k). The complete reference is available in the technical report titled “Scalable and robust group discovery on large transactional data” available for download from the Auton Lab web site.

1.9 Latent Space Modeling of Link Data

We developed an extension of group detection approach that projects entities into a 2 or 3 dimensional space instead of a group. This allows for a more natural way to decide who is working with whom without needing to artificially define a notion of a group. It also allows for easy visualization of “where” a person is relative to their associates. Additionally, our approach generalizes a static model of relationships into a dynamic model that accounts for relationships drifting over time. We managed to make it tractable to learn such models from data, even as the number of entities gets large. The generalized model associates each entity with a point in p-dimensional Euclidean latent space (where p can be small). The points can move as time progresses but large moves in latent space are improbable. Observed links between entities are more likely if the entities are close in latent space. Such models are tractable (sub-quadratic in the number of entities) by the use of appropriate kernel functions for similarity in latent space; the use of low dimensional KD-trees; a new efficient dynamic adaptation of multidimensional scaling for a first pass of approximate projection of entities into latent space; and an efficient conjugate gradient update rule for non-linear local optimization in which amortized time per entity during an update is O (log n). We use evaluated this method using both synthetic and real-world data on up to 11,000 entities. The results indicate near-linear scaling in computation time and improved performance over four alternative approaches. More details can be found in the paper titled “Dynamic Social Network Analysis using Latent Space Models” available for download on the Auton Lab web page.

2 Software produced

All software produced can be downloaded from www.autonlab.org . Upon request we can additionally send the software on a CD-ROM.

2.1 Group Detection Algorithm (GDA)

GDA has been implemented, tested, documented, and distributed to a variety of users (listed in the transition section). It runs under Linux and Windows and takes input and output in the form of comma-separated ASCII files, though it also has an API for developers who wish to link it directly with database queries. It is available for other researchers to evaluate (we are currently only allowing use by academic institutions for research purposes).

2.2 K-groups (Computationally tractable, approximate GDA)

This has been implemented and will be delivered to DARPA with this final report. More usefully, it is part of the k-groups/GDA combo described below.

2.3 cGraph (fast connection-finding)

This has been implemented and is also available on our website.

2.4 K-Groups/GDA combination

It is available for other researchers to evaluate (we are currently only allowing use by academic institutions for research purposes).

2.5 GDA output browser

This is an independent PERL script that can run on the output of GDA to allow users the chance to point-and-click their way around the groups detected by GDA. This runs under Linux and required the Posgres database and the Mozilla browser.

2.6 AFDL

This program takes data about entities. The entities are linked together according to a table of links. Some entities are known to be active: these are recorded in the table of active entities. Optionally, the user may specify a table of properties of the entities called the demographics. The program then asks the question: which additional entities are most likely to be active? It rank-orders the entities by likelihood of being active and returns the answer in a table called ranking. It is downloadable from the Auton Lab web site.

2.7 SBNS

Screen-based Bayes Net Structure search. As described above, **Bayes net is** a computationally efficient algorithm that performs Bayes Net structural learning from a very large binary dataset Available for download from the Auton Lab web site.

2.8 XGDA

The XGDA software takes link information as input, and learns groups, subgroups and friends (i.e., most likely collaborators) from that link information. Downloadable from the Auton Lab web page (we are currently only allowing use by academic institutions for research purposes).

3 Reports produced

All these reports are available from http://www.autonlab.org/autonweb/paperSearch.jsp.

- Stochastic Link and Group Detection, Jeremy Kubica Andrew Moore Jeff Schneider Yiming Yang, *Proceedings of the Eighteenth National Conference on Artificial Intelligence, 2002.*

- A Comparison of Statistical and Machine Learning Algorithms on the Task of Link Completion, Anna Goldenberg, Jeremy Kubica, Paul Komarek, Andrew Moore, Jeff Schneider, *KDD Workshop on Link Analysis for Detecting Complex Behavior, 2003.*

- cGraph: A Fast Graph-Based Method for Link Analysis and Queries, Jeremy Kubica Andrew Moore, David Cohn, Jeff Schneider, *Proceedings of the 2003 IJCAI Text-Mining & Link-Analysis Workshop, 2003.*

- Finding Underlying Connections: A Fast Graph-Based Method for Link Analysis and Collaboration Queries, Jeremy Kubica, Andrew Moore, David Cohn, Jeff Schneider, *Proceedings of the International Conference on Machine Learning, 2003.*

- Tractable Group Detection on Large Link Data Sets, Jeremy Kubica, Andrew Moore and Jeff Schneider, *The Third IEEE International Conference on Data Mining, 2003.*

- Empirical Bayes Screening for Link Analysis, Anna Goldenberg and Andrew Moore, *Workshop on Text Analysis and Link Detection*, IJCAI 2003.

- Tractable Learning of Large Bayes Net Structures from Sparse Data, Anna Goldenberg and Andrew Moore, *International Conference on Machine Learning*, 2004.

- Alias Detection in Link Data Sets, Paul Hsiung Andrew Moore Daniel Neill and Jeff Schneider, *International Conference on Intelligence Analysis*, 2005.

- Dynamic Social Network Analysis using Latent Space Models, Purnamrita Sarkar and Andrew Moore, *SIGKDD Explorations: Special Edition on Link Mining, 2005.*

4 Transitions

4.1 Transition to a government customer

We have received funding from another government agency to extend and apply our algorithms for their intelligence needs. Their application areas are classified so we do not have direct access to them. We are funded jointly with Alphatech and they have the required clearances. Our GDA, k-groups and XGDA software has been transferred to Alphatech in source code form. They have successfully applied the algorithms.

4.2 Participation in TIA

Also in collaboration with Alphatech, the GDA and kGroups algorithms have been used in two of the TIA wind experiments: Rafale, and Nor'easter. These demonstrations produced very positive feedback and the plans were to continue with them in the following experiments pending decisions on the future, if any, of TIA.

4.3 Software available for other academic researchers

Most of the software listed above is currently available on the web (with a human in the loop to ensure distribution is to academic researchers). The software comes with manuals and sample data sets.

4.4 SAIC/RDEC Work

We continue working with SAIC in testing of Auton EAGLE components in the RDEC context. This has involved a great deal of discussion of how to use XGDA as a visualization tool, and more recently led to the identification of a few improvement possibilities of the AFDL algorithm.

5 Presentations

- *Eighteenth National Conference on Artificial Intelligence, 2002*
- *2002 DIMACS Summer tutorial on Homeland Security Data Mining*
- *2002 Federal CIO tutorial on Homeland Defense Data Mining (Arlington VA)*
- *KDD Workshop on Link Analysis for Detecting Complex Behavior, 2003*
- *2003 IJCAI Text-Mining & Link-Analysis Workshop, 2003*
- *International Conference on Machine Learning, 2003*
- *Third IEEE International Conference on Data Mining, 2003*
- *Scientific Advisory Board Meeting for Rome Labs (assisting R. Hawkins), 2003*
- *NASCIO Homeland Security Technologies Panel, 2003*
- *SAMSI Workshop on Data Mining and Machine Learning, 2003*
- *Center for Democracy and Technology, 2003*
- *International Conference on Machine Learning, 2004*
- *International Conference on Intelligence Analysis, 2005*

6 Participation in Program-wide evaluations

6.1 Challenge 2002

CMU participated in the TIE2 team led by Metron for the 2002 challenge problem. The primary contribution was the GDA algorithm and the API that allowed Metron to compile it directly with their software. The final results for the challenge problem were submitted without GDA due to time constraints with the integration. A subsequent evaluation demonstrated how GDA significantly reduced the number of false positive hypotheses Metron's system would have to evaluate. GDA was also used to anlayze the Leninist Terror Cell data and demonstrated superior performance at reconstructing the network compared to straw-man algorithms.

6.2 Challenge 2003

CMU participated in the EvenTIE for the 2003 challenge problem. Near the end of the evaluation, EvenTIE was separated into an "East" and a "West" team. CMU participated on both of these teams, but made a larger, more integrated contribution on the East team. For the East team CMU built a query engine that assessed the likelihood of: 1) assets comprising a valid vulnerability mode, 2) a vulnerability mode existing for a target, 3) a

set of individuals comprising a team from the same group, and 4) the team being dangerous. Queries of type number 3 were answered by kGroups. The other queries were answered by a Bayesian classifier that was learned from the historical data. The result of the challenge problem was that the EvenTIE East team received significantly better scores for total cost and several of the other measures.

6.3 Challenge 2004

Schneider participated in the April 2004 Challenge meeting. Moore participated in the June 2004 Challenge meeting. Moore, Schneider and Dan Pelleg met with Metron on bJuly 14th, 2004 to coordinate challenge work.

The 2004 challenge work involved:

- PBE experiments for MNOP (presented at June 2004 workshop)
- Comparing applicability of GDA vs ISI's clusterer on different datasets
- Assessing the extent to which cluster detection helps event detection (with Metron and ISI)
- Assessing the extent to which event detection helps cluster detection (with Metron)
- We have also put time into preparing for the Y3 challenge new data format and participating in discussion of Y3 evaluation.