Link Mining Applications: Progress and Challenges

Ted E. Senator*
DARPA/IPTO
3701 N. Fairfax Drive
Arlington, VA 22203
ted.senator@darpa.mil

ABSTRACT

This article reviews a decade of progress in the area of link mining, focusing on application requirements and how they have and have not yet been addressed, especially in the area of complex event detection. It discusses some ongoing challenges and suggests ideas that could be opportunities for solutions. The most important conclusion of this article is that while there are many link mining techniques that work well for individual link mining tasks, there is not yet a comprehensive framework that can support a combination of link mining tasks as needed for many real applications.

Keywords

Link mining, link analysis, link discovery, pattern discovery, pattern analysis, pattern matching, structured data, complex event detection, data mining applications.

1. INTRODUCTION

Link mining is a fairly new research area that lies at the intersection of link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining [10]. However, and perhaps more important, it also represents an important and essential set of techniques for constructing useful applications of data mining in a wide variety of real and important domains, especially those involving complex event detection from highly structured data. This article provides examples of problems and domains that require link mining techniques and discusses the technical requirements of these problems and domains. It reviews progress in developing link mining techniques that can meet some of these requirements and outlines some needs that are not yet met and, therefore, are both open research challenges and potential opportunities for new application construction.

A central claim of this article is that link mining presents both challenges and opportunities. It presents challenges because data mining techniques for non-linked data are inadequate for similar problems with linked data and because the combinatorics of linked domains typically far exceed those of domains characterized by non-linked data. It presents opportunities because the structure of linked data provides both constraints on what can be inferred and additional information for inference than can be obtained from non-linked data.

This article is organized as follows. After this introduction, section 2 describes characteristics of domains and the associated tasks that require link mining techniques. Section 3 reviews the state of techniques for analyzing linked data as of about 10 years ago. Section 4 overviews progress in the past decade. Section 5 discusses some unmet needs and suggests some possible approaches for research to address them. Section 6 concludes.

2. PROBLEM DOMAIN & TASK CHARACTERISTICS

Many domains of interest are inherently relational. In fact, they are not just relational, but highly structured. By highly structured, we mean not only that there are semantically meaningful connections[†] between entities of the same and different types, but also 1) that there are grouped entities that have as members other entities (of single or multiple types) and 2) that there are multiple abstraction hierarchies on these entities and their relationships. (Note that these two types of hierarchies correspond to the usual *part-of* and *is-a* predicates in many knowledge representation systems; the distinction between these two types of hierarchies is sometimes ignored in many link mining techniques, but is as important for learning and inference in link mining as it is in knowledge representation and reasoning.) At a high level of abstraction the application needs may be to correct inaccuracies and resolve uncertainties in the data, infer the existence of missing entities and links, identify higher-order entities that are not explicitly represented in the data, classify or score entities according to some attributes of interest, detect interesting subgraphs, detect changes or anomalies, and/or learn significant patterns.

These rich domain structures provide a wide choice of representations; selecting an appropriate representation that enables all aspects of an application to be solved is a major challenge because the distinct techniques needed for particular aspects of an application often depend on different representational choices. Typically a database or set of databases is available. The databases may use incompatible data models, any or none of which may be appropriate for the link mining techniques. The combined databases may be thought of as a large heterogeneous graph; however, translating from a particular schema to a graphical data model is a one-to-many process. Often the data models are sparsely populated, contain uncertain and incorrect information, require inference to specify entities of interest – which may occur very rarely – from those that are

* The views and conclusions expressed in this paper are solely those of the author and should not be interpreted as representing the official policies, either expressed or implied, of DARPA, the Department of Defense, or the US Government.

† Sometimes link analysis techniques are applied to the far weaker type of connections resulting from co-occurrence of terms in text or to concepts such as "topics," with correspondingly weaker results.

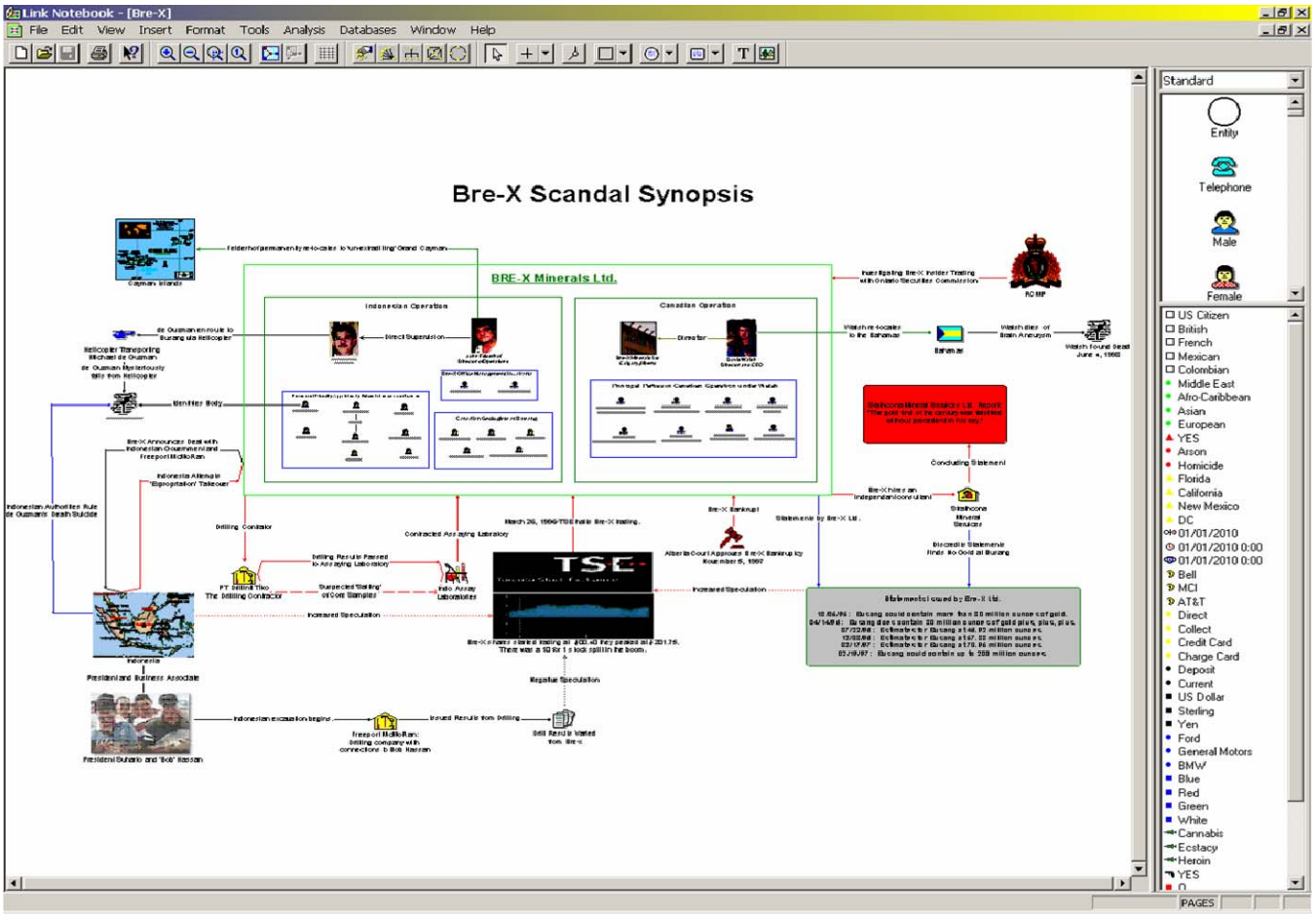

Figure 1: Example of Stock Fraud Case Showing Characteristics of Highly Structured Domains

directly observed, contain heterogeneous observations from multiple sources, are not clearly and unambiguously identifiable, depend on consolidation of individual transactions, are not amenable to sampling, and exhibit high degrees of relational autocorrelation.

2.1 Domain Examples

Examples of complex structured application domains in which link mining techniques are needed include counterterrorism detection, counterdrug and other law enforcement, money laundering, stock fraud, and many others. Figure 1 is an example of a portion of a stock fraud domain that illustrates many of the key features of real domains that must be addressed by link mining techniques. (It depicts a well-known stock fraud case that was discovered in 1997, involving the falsification of core samples by a geologist and an executive, that resulted in a loss to investors of over $3B.) The boxes around the sets of people show the group hierarchy, with individuals being part of geographically distinct operations of the same company. The labels on the links show semantically meaningful connections between entities and between events. The rich iconography and color is used to indicate many is-a relationships. The mixing of entities (people, organizations, countries, etc.), relationships, and events on a single diagram is typical of many domains, and represents a key challenge for successful applications of link mining techniques, which often are focused at less heterogeneous data. There are quasi-static role relationships (e.g., drilling contractor) and transactional events (e.g., bankruptcy occurs). This example is, of course, the end result of a complex analysis. In reality, it should be thought of as an interesting and relatively complete, but also very small portion of a large database, that was determined to represent an interesting set of entities, relationships, and events.

Figure 2 provides another example. It depicts a nuclear material theft that occurred in 1992. A key aspect of figure 2 is the depiction of both a data model and a domain model, clearly showing how these two concepts are distinct. The circled numbers indicate the correspondence between specific items in the data and domain representations. Many of the other characteristics of figure 1 are also present in figure 2.

2.2 Task Examples

The application goal of link mining in highly structured domains may generally be described as learning or inferring as much as possible about the domain. Applications can include learning models of the domain as well as inferring information about the entities, relationships, their attributes, and the connectivity structure of the domain. One important application task in such domain typically consists of classifying inferred entities over time. Another key task is detecting activities of interest. These two tasks are, of course, closely related. Typically the former task is based

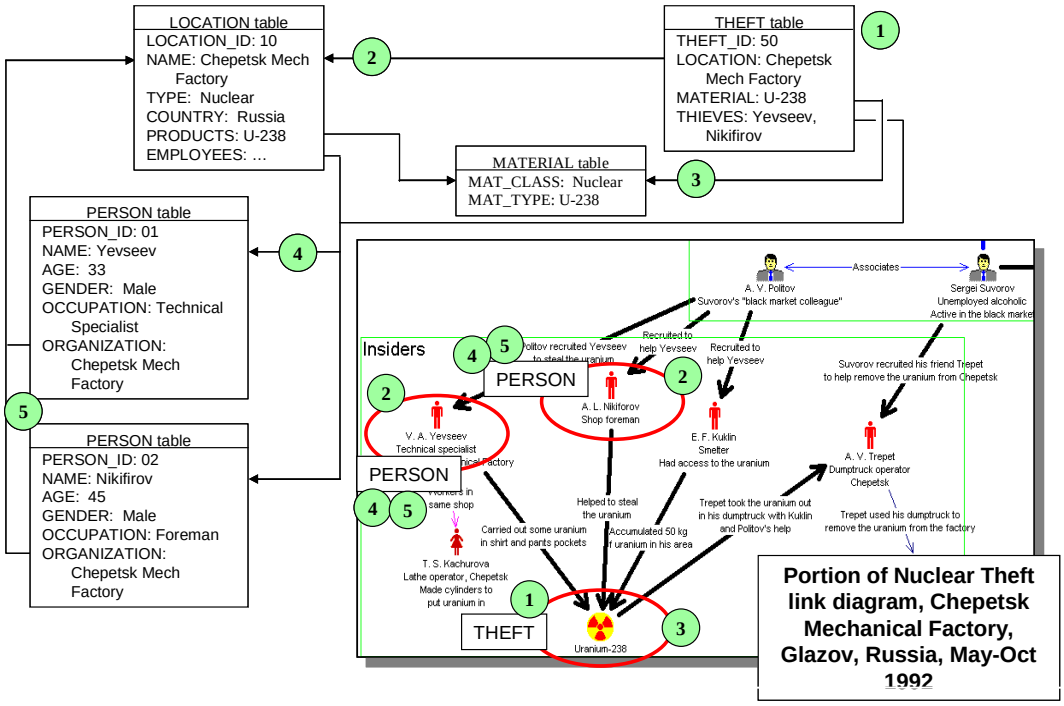

Figure 2: Nuclear Smuggling Example, showing data and domain models

on long-term role or trust type relationships while the latter is based on events. For example, an intelligence organization would want to keep track of certain people, places, events, and materials, infer the existence of organizations or (planned) activities, and classify (or score) these entities as threatening or not. A regulator might want to track activities by organizations and identify individuals who are at a high-risk for committing fraud. A transaction approval authority might want to track accounts over time and disapprove (or ask for additional verification for) transactions that might be indicative of improper account use.

Accomplishing these goals requires many subsidiary problems to be solved first. In particular, the challenges of integrating data from different databases, resolving identities and consolidating transactions, inferring and acquiring missing data (entities and especially links), clustering entities into groups, inferring link strength and importance, and constructing unspecified features must all be solved to some degree.

For such application tasks, the basic method for detection of entities or organizations of interest is pattern matching. However, patterns are rarely known exactly or specified fully. Even if a pattern of activity that is indicative of improper activity is fully specified, the data to support a full match may not all be available. And if the data are available, it may reflect a deviation from the specified pattern because the activity of interest is performed in different ways by different people. Supporting this pattern matching task is the task of discovering patterns. While pattern matching may be thought of as inference, pattern discovery is more properly thought of as learning. Pattern matching and pattern discovery are really complementary – a pattern is discovered because of the existence of many instances, and a pattern instance is matched from the pattern template.

2.3 Terminology

Knowledge discovery in databases (KDD) has been defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [9]. Data mining is considered to be a particular step in this process, albeit the significant automated one in which algorithms are applied to extract the patterns. However, there is an inherent ambiguity in the use of the term "pattern" that is particularly important when one considers link mining applications. The patterns that are discovered are typically templates, or class-based. They contain variables that, when instantiated, provide examples of the pattern. The "pattern" is then used as the basis for inference to detect other instances by pattern matching. These other matches may also be referred to by the same term. Unfortunately, this dual meaning of the word pattern leads to confusion about what is data mining – i.e., pattern discovery – and what is the application – i.e., pattern detection.* Finally, in some domains, a third use of the term pattern is with respect to a frequently occurring (i.e., repeated) behavior of an individual entity. In this case, many instances of a pattern template are observed with the same instantiation of at least one of the variables.

There is also an inherent ambiguity with respect to the term "link" that occurs in many circumstances, but especially in discussions with people whose background and research interests are in the database community. In the database community, especially the subcommunity that uses the well-known entity-relationship (ER) model, a "link" is a connection between two records in two different tables. This is important because in this model a link is not identical to a real-world relationship. The real world relationship is modeled by having a corresponding table, or relation. Links are abstract connections between tables that refer to the co-occurrence of the same value in fields of two different relations. Links can be one-to-one, one-to-many, or many-to-many. So, for example, a link might connect a "person" record to the person's account or to that person's address.

This usage of the term "link" in the database community differs from that in the intelligence community and in the AI research community. In the intelligence community a "link" typically refers to some real world connection between two entities. The difference between these two communities' use of the term "link"

* This confusion may have contributed somewhat to recent public controversies regarding the use of data mining.

is seen in figure 2; the database use of the term describes the lines in the top and left of the figure, while the domain use of the term describes the lines in the bottom right. That real world connection might be that one of them called the other one's telephone number or that they both called the same telephone number. It may refer to a common attribute value, a common transaction, or a common group membership. Figure 3 illustrates the different ways that a semantically meaningful link might be represented in a database. The link between *person a* and *person b* from having the same telephone number is illustrated on the left. Taken together with the phone calls table, there would be a link between *person a* and *person c* (and another between *person b* and *person c*) based on the telephone calls. And there would be a common group membership between *person a* and *person b* based on the table depicted in the bottom right.

Converting between the information specified in a relational database and maintaining the appropriate link semantics is currently done ad-hoc, if at all. Link mining methods that rely on graphical algorithms such as those that measure match quality by graphical edit distance will give vastly different results depending on how this mapping is performed. Figure 4 illustrates three alternative graphical representations of a telephone call between *person a* and *person b*.

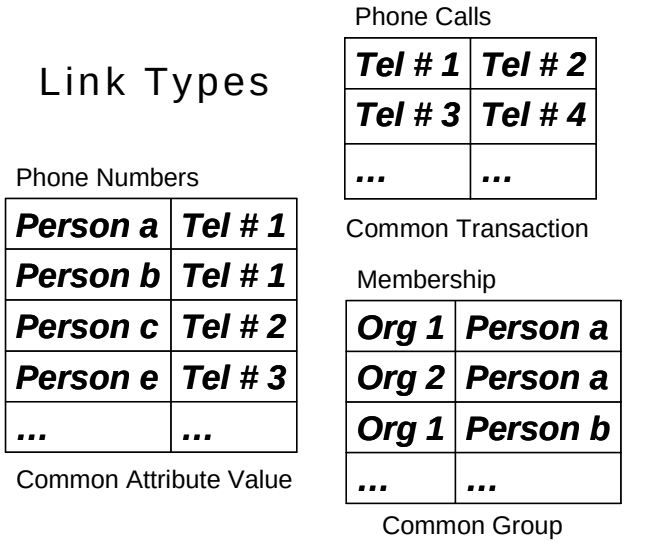

Figure 3: Three DB Link Representations

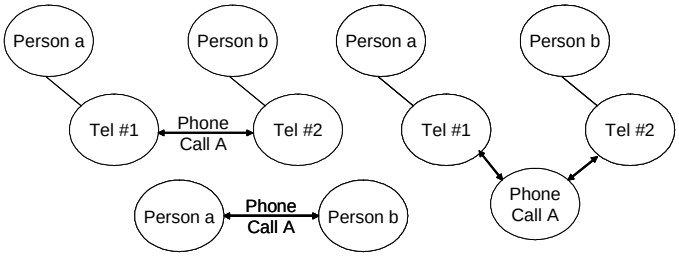

Figure 4: Alternative Graphical Representations of a Transaction

Entity_ID	Entity_Name	Entity_Type	Attribute1	Attribute2	...

Link_ID	Link_Name	Link_Type	Entity1_ID	Entity2_ID	Attribute1	...

Figure 5: Typical Link Analysis Tool 2-Table Structure

2.4 Data Structures

Databases are frequently represented in third-normal form. This representation is efficient for typical database query operations. However, it is inefficient for the types of queries typically performed in link analysis, such as finding all paths between two entities or finding all entities within a certain distance of a specified entity. It is also inflexible with respect to the addition of new entity or link types. Link analysis applications typically require the ability to refocus on different link types or different sets of entities, as they concentrate on detecting, analyzing, and displaying different subsets of data with their full connectivity structure. Link analysis representations often are based on two tables, one for entities and one for links, with the semantics of the data types determined by the entries in these tables rather than being specified in the data model. This type of representation is inefficient for aggregate queries but extremely flexible for interactive queries on reasonably sized datasets and difficult to update. Figure 5 illustrates the two-table data structure typically used in link analysis tools.

3. HISTORICAL BACKGROUND

In 1995 link mining had not yet been born. Despite the almost complete lack of understanding of the need for techniques to analyze structural data, there had been some relevant prior work. This section briefly reviews such work in the areas of structural learning, data mining, and link analysis.

3.1 Research in Structural Learning

Machine learning had addressed the problem of learning structural descriptions as early as the 1970's, beginning with the work by Winston on "Learning Structural Descriptions from Examples" [28]. Constructive induction had addressed the issues of feature construction in complex domains. This work was, of course, not based on structured data in the sense that we use the term, but on a carefully constructed set of features that represented the possible relationships between parts of structured objects. Some applications had begun to address structural representations. But this work was applied to very small sets of examples, rather than to massive data sets of the sizes considered in link mining. And, more important, there was no connection between components of different objects; i.e., each example was independent of the others, rather than part of a large structured database from which the objects had to be inferred or detected.

3.2 Early Work in Data Mining

In 1995 most people in the data mining community were unaware that linked data presented new and different challenges that could not be addressed with known techniques. In 1991 Manago and Kodratoff [21] had applied an ID-3 style algorithm to analyse frame-based data structures. KDD-95 had three papers that addressed aspects of linked or structured data, [4] [13] and [24]. The FinCEN AI System [25] was an early KDD application that explicitly dealt with linked data. In fact, it noted explicitly that "flat feature vector[s] … are unable to describe the more complex data structures that are required to represent money-laundering schemes." Dzeroski [5]

proposed inductive logic programming as a KDD technique that could handle structured data. One key overview paper from KDD-96 [9] did include in its list of research and application challenges the item "complex relationships between fields" which it explained as "hierarchically structured attributes or values, relations between attributes, and more sophisticated means for representing knowledge about the contents of a database will require algorithms that can effectively utilize such information. Historically, data mining algorithms have been developed for simple attribute-value records, although new techniques for deriving relations between variables are being developed"; however, the other key overview paper [23] did not. The techniques referred to in [9] are those described in [4] and [5].

In fact, even several years later many people confused link mining with techniques that created structural models of non-linked data. For example, in the 30 October 1997 the "what's new" section of the kdnuggets.com web site stated that it "renamed Dependency Analysis section to Link Analysis." An examination of this section showed that the contents consisted of work in graphical models, rather than models of relational or structured data. Others used the term "link analysis" to refer (incorrectly) to association rules, which, while they suggest links between categories of items as an output, do not operate on linked data. A search of the full text of the Proceedings of the KDD conferences from 1995 to 1998 for the term "link analysis" matches only one paper, reference [13] herein. Similarly, the only reference to link analysis in [20], which while published in 2002 is based on contributions from [1], is [26], a revised version of [14]. The term "structured data" was used to refer to data stored in databases, in contrast to text or images, for example, but not to refer to the complex objects that are the subject of link mining. Some papers began to analyze semi-structured data; however, by this they meant mostly things like web documents. Reference [22] concluded that "there is a need to apply data mining to real databases which are characterized by complex data structures."

3.3 Link Analysis Tools

Link analysis is a technique used in law enforcement, intelligence analysis, fraud detection, and related domains. [27] It is sometimes described using the metaphor of "connecting the dots." Link diagrams, showing the connections between people, places, events, and things, are invaluable tools in these domains. Link diagrams are often constructed painstakingly by a manual process, based on information that is collected incrementally as the focus of a particular investigation. What link mining can add to link analysis is the automatic construction of link diagrams from a large database, in which the link diagram represents a particular set of evidence that is somehow related to a particular higher-order concept of interest. In particular, link mining can enable not only the visual representation of the structured results of a series of directed queries, but also the automated selection of the information that an analyst would have obtained had he or she been aware that there was a common thread. Metaphorically, link mining offers the potential not only for connecting the dots, but for determining which dots to connect, a far more difficult task.

Link analysis tools in 1995 were essentially filing and visualization systems for analysts. They facilitated construction of linked diagrams from user entered information. Each instantiation of a tool was a "case" in law enforcement or intelligence terms. The case had to be opened (and there had to

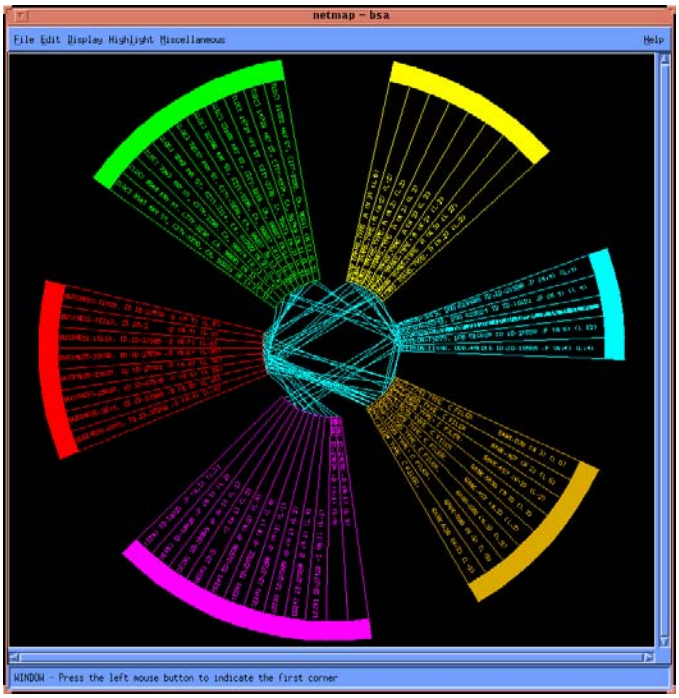

Figure 6: NetMap "Wagon-Wheel" Display

be a crime committed or an investigation initiated before a case would come into existence) and then populated with information that was obtained at user direction, by some combination of investigative work and database queries. These tools were a great aid in organizing and presenting the results of investigations; however, they did not aid in identifying new investigations. Figures 6 & 7 are examples of screen shots from a link analysis tool called NetMap [25]. Figure 6 shows a large-scale view of a dataset; each sector of the display is a different entity type (e.g., person, address, account) and the lines in the middle show the connections. The diagram in figure 6 would be used as part of the

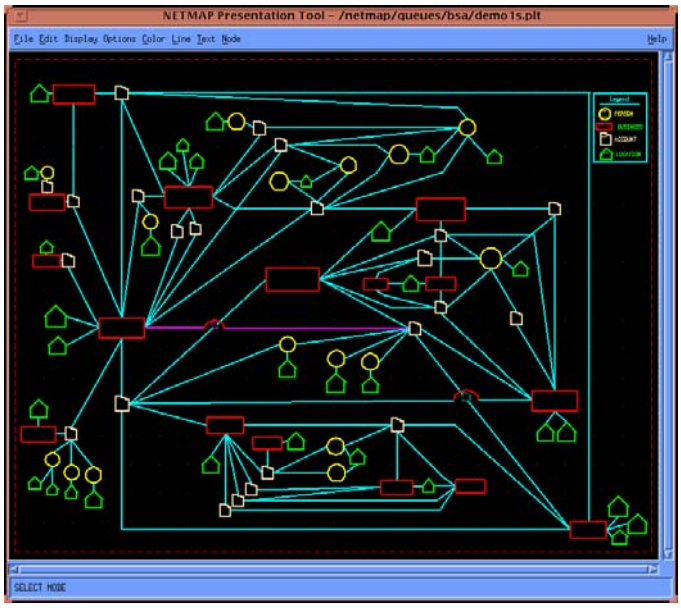

Figure 7: NetMap "Anacapa" Chart Display

investigative process as an analyst is attempting to determine which entities and relationships are relevant; the corresponding diagram in figure 7 shows an actual case that resulted from such an investigation.

4. PROGRESS

The past 10 years have seen an explosion of interest and associated progress in techniques for learning from relational and structured data. Several factors contributing to this explosion are the awareness of the importance of "connecting the dots" that developed subsequent to September 11, 2001, the ubiquity of the world-wide web, and the newly emerging area of network science, all of which have made researchers (and the population at large) generally aware of the importance of links and networks as first-class objects for analysis.

4.1 Research, Applications and Tools

Various workshops were held in the late 1990's to bring together the AI and link analysis communities. In 1997 there was a workshop on AI Approaches to Fraud Detection and Risk Management [1], [6] at which several papers discussed the ideas of link analysis. This was followed by a 1998 AAAI Fall Symposium on Artificial Intelligence and Link Analysis [17], which is the first time that the two communities came together with a direct focus on extending AI techniques to linked data.

In 1998 Carnegie Mellon University hosted a workshop sponsored by DARPA on Knowledge Discovery, Data Mining and Machine Learning. [15] This workshop focused on new technologies that could be developed and applied to assist Counter Transnational Threat (C-XNT) activities within the Federal government, particularly within DoD and the US intelligence community. The report of this workshop recognized the following as key technical "challenges met poorly by [then] current technology": relational and probabilistic patterns, constructive inference, vast data volume with few positive instances, knowledge-based learning and inference, temporal and spatial relationships, heterogeneous data sources and data items, fragmentary data and active learning, adversarial conduct, high costs of failure, and dynamic patterns. Its recommendations were grouped into the areas of learning using prior knowledge, active learning, and incremental and cumulative learning, all areas which can and have been studied in the context of propositional data as well as relational data.

In the second half of the decade from 1995-2005, there were a multitude of workshops on topics such as learning statistical models from relational data (at AAAI-2000 and IJCAI-2003), on the relationship of statistical relational learning to other fields (at ICML-2004), on link analysis for detecting complex behavior (at KDD-2003), on link analysis and group detection (at KDD-2004), on link discovery (at KDD-2005) and on multi-relational data mining (at several conferences).

Concurrently with the increased research interest in the late 1990's, several applications that combined ideas from data mining and link analysis began to appear. The FinCEN AI System (FAIS) was an early attempt to combine link analysis techniques with automated detection. [25] In this system the patterns that indicated potential money laundering were derived from expert consultation rather than from any automated data mining, although they were verified by an examination by these same experts of the results that were produced on real data. Most of the computational load of the system was devoted toward the consolidation of reported transactions by person, organization, and account. Patterns of suspiciousness were then applied to these consolidated persons, organizations, and accounts to generate potential leads that were reviewed by analysts.

Other notable applications include the work on combating cellular telephone fraud by combining ideas from machine learning and data mining [7], [8] and work on detecting improper activity in the Nasdaq Stock Market [19]. Research such as the development of techniques such as probabilistic relational models [11], the continued development and scaling-up of inductive logic programming techniques, an increasing awareness of the statistical issues in dealing with networked data [16], and techniques for graph-based data mining [2] are but some of the key technical advances in this period.

An overall view of the technical state of the art in link mining techniques as of 2003 is presented in [10]. The KDD-cup competition in 2003 was specifically focused at mining large relational datasets. [18]

Link analysis tools have also become more capable.[*] In particular, sophisticated searching for nodes, links and groups of nodes and links based not only on attribute values but also on the connectivity structure, on aggregate values, and on graph-theoretic and on social network metrics is typically supported. A greater variety of graphical displays, including temporal evolution views, are available. Larger datasets, up to hundreds of thousands of nodes and links, are supported, and tighter database integration enables more effective searches. Non-Obvious Relationship Awareness (NORA), recently acquired by IBM, detects multi-link paths between records based on common data values in massive databases and data streams.

4.2 Evidence Extraction and Link Discovery (EELD)

In the spring of 2001, DARPA released a Broad Agency Announcement soliciting ideas to develop, demonstrate, and evaluate technology to extend technical capabilities in the three technical areas of interest to structured, or relational, data. [3] These relationships could be transactional, social, temporal, or geographical. Of particular interest was the extraction and discovery of related temporal events that may be components of scenarios of interest and the learning of patterns that comprise such scenarios. Evidence extraction technology would be extended from the ability to extract accurately named entities and attributes to the ability to extract relationships between entities and attributes of these relationships. Link discovery technology would be developed to enable the ability to discover related entities, additional attributes, and other relevant relationships from a starting point of a set of entities, relationships, and attributes potentially relevant to a scenario of interest. Pattern learning technology will be developed to enable learning from examples of instances of scenarios of interest patterns or models which will facilitate the extraction and discovery of additional instances of scenarios of interest.

The resulting program was the first large-scale effort to focus research on the problem of linked and structured data. Contracts

[*] See http://www.i2inc.com/ and http://www.visualanalytics.com/ for two examples.

were awarded in late summer of 2001, and interest in this area increased drastically shortly thereafter following the attacks of September 11, 2001. EELD's goals were to increase the size of datasets that could be analyzed by several orders of magnitude over the life of the program on an increasing set of link mining tasks. EELD was responsible for much of the progress in link mining between 2000 and 2005; for example, 6 of the 11 papers presented in [12] were from work funded as precursors to EELD as was much work reported in other workshops and conferences during this time. All three winners of the KDD-Cup 2003 task 4 (open task) competition were EELD funded researchers. [18]

5. OPEN ISSUES AND IDEAS

Real applications typically have many requirements, with pattern discovery and pattern detection at their core. Other requirements arise from characteristics of the environment; e.g., the need to combine data from multiple sources, the need to compute certain derived attributes for use by the pattern discovery and pattern detection algorithms, the need to support multiple analysts, the need to record and audit system activities, the need to sort data into distinct threads of interest, the need to support detection of patterns that may occur over long time periods, the need to visualize patterns (in both senses of the term), the need to both discover and detect patterns that are continually changing, the need to protect the identity of entities until a particular level of belief in their interestingness is supported and, perhaps, proper approvals are obtained, the need to allow for multiple competing hypotheses, the need to allow for refutation of previously asserted evidence, the need to allow for frequent tuning between false positives and false negatives, the need to support both comprehensive review of all data based on approved patterns and ad-hoc review based on particular external indicators, the need to support what-if analyses by individuals, the need to support organizational workflow processes, the need to operate continuously and perhaps autonomously on incrementally arriving massive data streams, and perhaps others.

Given the progress that has occurred in the past decade and the challenges outlined in section 2, what remains to be done? Rather than focus on the research issues that would enable more effective solutions to particular link mining tasks, we discuss the overall needs to construct effective link mining applications. To address this question, it is helpful to imagine a complete "link mining toolkit." What would such a toolkit look like? It would have to solve all the application issues discussed earlier and it would have to do so in a manner that enabled it to be operated as part of an integrated application. Such a link-mining toolkit would have to, at a minimum, enable the exchange of data and models between different link mining techniques. To build integrated applications, it would require not only a common representation for linked data but also meta-level descriptors of the inputs, outputs, transformations, and assumptions, for each of the included link mining techniques so they could be used in various combinations to solve the different aspects needed by an application. For real-time applications it might also require a control architecture that enabled automatic dynamic algorithm selection and application, as well as an ability to evaluate the quality of results from any particular technique. And it would have to be tightly integrated with underlying databases management systems.

What would be needed to construct such a toolkit? Most important, it would require a language that enabled the natural representation of entities and links. Such a language would also allow for the representation of pattern templates and for specifying matches between the templates and their instantiations. The language would have to accept an arbitrary database schema as input, with a specified mapping between relations in the database and fundamental link types in the language. It would have to compile into efficient and rapidly executable database queries. It would need to be able to represent grouped entities and multiple abstraction hierarchies and reason at all levels. It would have to enable the creation of new schema elements in the database to represent newly discovered concepts. It would need to represent both pattern templates and pattern instances, and to have a mechanism for tracking matches between the two. It would have to have constructs for representing fundamental relationships such as *part-of*, *is-a*, and *connected-to* (the most generic link relationship), as well as perhaps other high-level link types such as temporal relationships (e.g., before, after, during, overlapping, etc.), geo-spatial relationships, organizational relationships, trust relationships, and activities and events. The toolkit would include at least one and possibly many pattern matchers. It would require tools for creating and editing patterns. It would have to include visualizations for many different types of structured data. It would need mechanisms for handling uncertainty and confidence. It would have to track the dependence of any conclusion (e.g., pattern match or discovered pattern) back to the underlying data, and perhaps incorporate backtracking so the impact of data corrections could be detected. It would need configuration management tools to track the history of discovered and matched patterns. It would need workflow mechanisms to support multiple users in an organizational structure. It would need mechanisms for ingesting domain-specific knowledge. It would have to be able to deal with multiple data types including text and imagery. And it would have to be able to rapidly incorporate new link mining techniques as they are developed. Finally, it would need to include mechanisms for maximum privacy protection.

6. CONCLUSIONS

The most important conclusion of this article is that while there are many link mining techniques that work well for individual link mining tasks, there is not yet a comprehensive framework that can support a combination of link mining tasks as needed for many real applications. The construction of successful and useful link mining applications is still very much an ad-hoc enterprise. Choosing the right representation for the underlying data and concepts is still the key, much more important than the particular choice of algorithms for scoring or classification. Designing an effective architecture to support all necessary functions of an integrated application is also a key to success.

7. ACKNOWLEDGMENTS

Some of the ideas and techniques described in this paper were developed as part of DARPA's EELD program between 2001 and 2003. All the researchers who were part of this program, especially David Jensen and Foster Provost, contributed greatly. Many of the ideas pre-dating EELD were developed at FinCEN and NASD Regulation, Inc. I thank all my colleagues from those organizations, especially Henry Goldberg, who helped develop and implement them. Finally, I thank Willem Van Der Westhuizen, the Internet Services Manager aboard the M/V Norwegian Jewel, without whose assistance I would not have had

available a computer to prepare and submit this paper.

8. REFERENCES

[1] *AI Approaches to Fraud Detection and Risk Management*, Technical Report WS-97-07, AAAI Press, 1997.

[2] Cook, D. and Holder, L., "Graph-Based Data Mining," *IEEE Intelligence Systems*, 15(2):32-41, 2000.

[3] DARPA, BAA 01-27 Evidence Extraction and Link Discovery. Available at http://www.darpa.mil/baa/BAA01-27.htm

[4] Djoko, S., Cook, D. J., and Holder, L.B., "Analyzing the Benefits of Domain Knowledge in Substructure Discovery," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pp. 75-80, AAAI Press, 1995.

[5] Dzeroski, S., "Inductive Logic Programming for Knowledge Discovery in Databases," in *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.

[6] Fawcett, T., Haimovitz, I., Provost, F., and Stolfo. S. "AI Approaches to Fraud Detection and Risk Management," *AI Magazine* 19(2): Summer 1998, 1998, 107-108

[7] Fawcett, T. and Provost, F., "Combining Data Mining and Machine Learning for Effective User Profiling", in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*. AAAI Press, 1996, pp. 8-13

[8] Fawcett, T. and Provost, F., "Activity Monitoring: Noticing Interesting Changes in Behavior," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99)*, pp. 53-62, ACM, 1999.

[9] Fayyad, Usama et. al., "Knowledge Discovery and Data Mining: Towards a Unifying Framework," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 82-88, AAAI Press, 1996.

[10] Getoor, L. "Link Mining: A New Data Mining Challenge" *SIGKDD Explorations* 4(2), 2003.

[11] Getoor, L., Friedman, N., Koller, D., and Pfeffer, A., "Learning Probabilistic Relational Models," in *Relational Data Mining*, Dzeroski, S., and Lavrac, N., (Eds.), pp. 307-335, Kluwer, 2001.

[12] Getoor, L. and Jensen, D. (Eds.). *Learning Statistical Models from Relational Data: Papers from the AAAI Workshop.* AAAI Press, Menlo Park, CA, 2000.

[13] Goldberg, H. and Senator, T. "Restructuring Databases for Consolidation and Link Analysis" in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pp. 136-141, AAAI Press, 1995.

[14] Goldberg, H. and Senator, T., "Break Detection Systems" in *AI Approaches to Fraud Detection and Risk Management*, Technical Report WS-97-07, AAAI Press, 1997.

[15] Goldszmidt, M. and Jensen, D. (Eds.). *DARPA Workshop on Knowledge Discovery, Data Mining, and Machine Learning (KDD-ML): Recommendations Report*, June 13-14, 1998, Carnegie Mellon University.

[16] Jensen, D. "Statistical challenges to inductive inference in linked data," in *Seventh International Workshop on Artificial Intelligence and Statistics*, 1999

[17] Jensen, D., and Goldberg, H. (eds.). *Artificial Intelligence and Link Analysis: Papers from the 1998 AAAI Fall Symposium*, AAAI Press, Menlo Park, CA 1998.

[18] KDD Cup 2003 homepage, Results and Sldes links: http://www.cs.cornell.edu/projects/kddcup/

[19] Kirkland, James D., Senator, Ted E., et. al., "The NASD Regulation Advanced Detection System (ADS)," *AI Magazine* 20(1):55-67, 1999.

[20] Kloesgen, W. and Zytkow, J. (eds.). *Handbook of Knowledge Discovery and Data Mining*. Oxford University Press, 2002.

[21] Manago, M. and Kodratoff, Y. "Induction of Decistion Trees from Complex Structured Data" in *Knowledge Discovery in Databases*, G. Piatesky-Shapiro and W.J. Frawley (Eds.), AAAI Press, 289-308.

[22] McKearney, Stephen and Roberts, Huw., "Reverse Engineering Databases for Knowledge Discovery," *in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 375-378, AAAI Press, 1996.

[23] Piatesky-Shapiro, Gregory, et. al., "An Overview of Issues in Developing Data Mining and Knowledge Discovery Applications," *in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp. 89-95, AAAI Press, 1996.

[24] Ribeiro, J.S., Kaufman, K.A., and Kerschberg, L., "Knoweldge Discovery from Multiple Databases," in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*, pp. 240-245, AAAI Press, 1995.

[25] Senator, T., Goldberg, H., et. al., "The Financial Crimes Enforcement Network AI System (FAIS)." *AI Magazine* 16(4):21-39, 1995.

[26] Senator, T. and Goldberg, H., "Break Detection Systems," in *Handbook of Knowledge Discovery and Data Mining*, W. Kloesgen and J. Zytkow (eds.), pp. 863-873. Oxford University Press, 2002.

[27] Sparrow, M. The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects. *Social Networks* 13, 251-274, 1991

[28] Winston, P.H., "Learning Structural Descriptions from Examples," *The Psychology of Computer Vision*, Winston, P.H. (Ed.), McGraw Hill, New York, ch. 5., 1975.