Network Data Mining: Methods and Techniques for Discovering Deep Linkage between Attributes

John Galloway [1,2] and Simeon J. Simoff [3,4]

[1]Complex Systems Research Centre, University of Technology Sydney
PO Box 123 Broadway NSW 2007 Australia
john.galloway@uts.edu.au
[2] Chief Scientist, NetMap Analytics Pty Ltd,
52 Atchison Street, St Leonards NSW 2065 Australia
[3]Faculty of Information Technology, University of Technology Sydney
PO Box 123 Broadway NSW 2007 Australia
simeon@it.uts.edu.au
[4]Electronic Markets Group, Institute for Information and Communication Technologies, University of Technology Sydney - PO Box 123 Broadway NSW 2007 Australia
http://research.it.uts.edu.au/emarkets

Abstract. Network Data Mining identifies emergent networks between myriads of individual data items and utilises special algorithms that aid visualisation of 'emergent' patterns and trends in the linkage. It complements conventional data mining methods, which assume the independence between the attributes and the independence between the values of these attributes. These techniques typically flag, alert or alarm instances or events that could represent anomalous behaviour or irregularities because of a match with pre-defined patterns or rules. They serve as 'exception detection' methods where the rules or definitions of what might constitute an exception are able to be known and specified ahead of time. Many problems are suited to this approach. Many problems however, especially those of a more complex nature, are not well suited. The rules or definitions simply cannot be specified. For example, in the analysis of transaction data there are no known suspicious transactions. This chapter presents a human-centred network data mining methodology that addresses the issues of depicting implicit relationships between data attributes and/or specific values of these attributes. A case study from the area of security illustrates the application of the methodology and corresponding data mining techniques. The chapter argues that for many problems, a 'discovery' phase in the investigative process based on visualisation and human cognition is a logical precedent to, and complement of, more automated 'exception detection' phases.

Introduction

The proliferation of data is both an opportunity and a challenge. It provides the details that businesses need to solve problems and gain market advantage, that organisations need to improve their operations, that banks and financial institutions need to fight fraud and that governments need to uncover criminal and terrorists activities. At the same time, a large volume of data with different storage systems, multiple formats and all manner of internal complexity can often hide more than it reveals. Data mining – "the process of secondary analysis of large databases aimed at finding unsuspected relationships that are of interest or value to the database owners" (Klösgen and Zytkow 2002) (p. 637) – emerged as an "eclectic discipline" (Klösgen and Zytkow 2002) that addresses these large volumes of data. Earlier data mining technologies have been primarily focused on the analysis of structured data (Fayyad, Piatetsky-Shapiro and Smyth 1996; Han and Kamber 2001; Hand, Mannila and Smyth 2001). Although the data mining researchers have developed methods and techniques that support a variety of tasks, the main interest of analytics practitioners has been focused on predictive modelling. The dominant scenario in predictive modelling is the "black box" approach, where we have a collection of inputs and one or more outputs, and we try to build an algorithmic model that estimates the value of the outputs as a function of the values of the inputs. There are several measures of model quality (Weiss and Zhang 2003), with the accuracy of predictions remaining as a key measure of model quality, rather than the theory that may explain the phenomena.

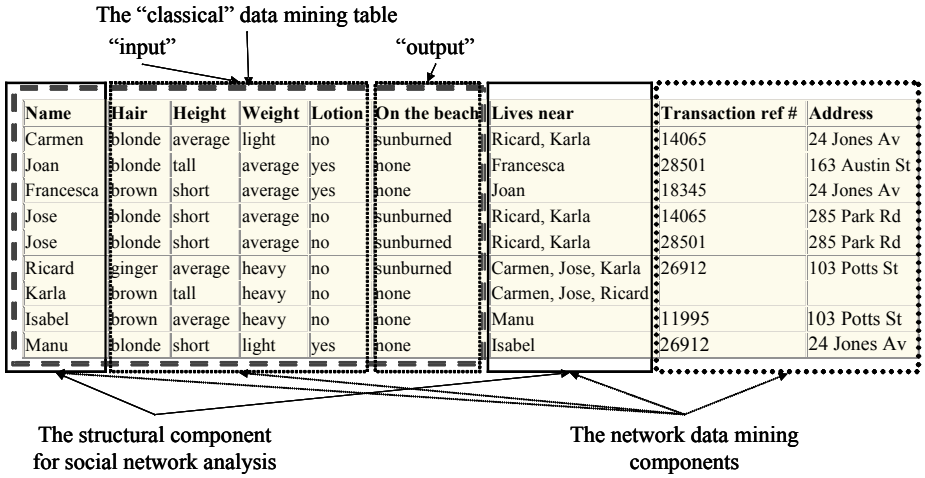
Fig. 1 illustrates the concepts in terms of a simple example of data about a group of college friends. The data table includes the following columns: name of the student; colour of hair; height and weight; a record of whether the student has been using lotion when exposed to sun; a record of whether the student gets sunburned when on the beach; a record about the proximity of the living locations of the students; transaction reference numbers; and student address. The double dotted line contours the portion of the data table which will be considered in the predictive modelling approach. As the "Name" column contains unique identifiers, it will be ignored, and the data mining task will be to develop a model of the student from this college with respect to the attributes "hair", "height", "weight", "lotion" and "on the beach". In an unsupervised approach, the students will be clustered into groups and the analyst ends up with the description of the different groups. In this case, the analyst is interested in predicting whether a new student will get sunburned when visiting the beach. The attribute "on the beach" is selected as the "output" (or "target") and the

attributes “hair” to “lotion” form the input vector. Given the values for the attributes “hair” to “lotion” for a new student, the resultant classifier should be able to predict whether the student will get sunburned or not. The key measure of the quality of the model is the accuracy of predictions, rather than the theory that may explain the phenomena through the relations between the values of the attributes.

In practice, the focus on predictive accuracy in the “black box” approach (inherited from regression analysis and automatic control systems) makes perfect sense. For example, the more accurate a classifier of tumours is based on the characteristics of mammogram data, the better the aid it can provide to young practitioners (Antonie, Zaiane and Coman 2003). In business areas like marketing, such an approach makes sense (for example, determining which few books Amazon should also offer to someone who searches for a particular book). Numerous examples from different areas of human endeavour, including science and engineering, are presented in (Nong 2003).

The theory explaining the “black box”, i.e. how and why inputs are related to outputs, is often of secondary importance to predictive accuracy. However, data mining applications in many areas, including businesses and security (Nong 2003) require deeper understanding of the phenomena. Such complex phenomena can be modeled with techniques from the area of complex systems analysis.

The network perspective with respect to a data set is illustrated in Fig. 1. The structural component of the data set describes explicitly some relationships between the individual entities in the data (in our example in Fig. 1, the column “Lives near” *explicitly* represents the relation of physical proximity between the areas where each student lives). Social network analysis (Wasserman and Faust 1994; Scott 2000) deals with this type of data analysis.



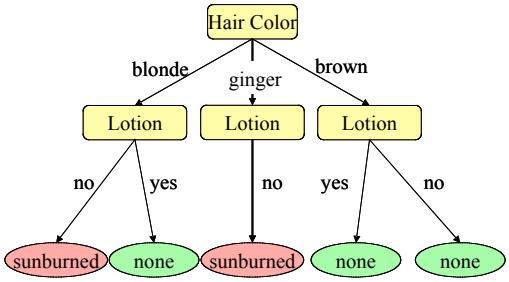
Name	Hair	Height	Weight	Lotion	On the beach	Lives near	Transaction ref #	Address
Carmen	blonde	average	light	no	sunburned	Ricard, Karla	14065	24 Jones Av
Joan	blonde	tall	average	yes	none	Francesca	28501	163 Austin St
Francesca	brown	short	average	yes	none	Joan	18345	24 Jones Av
Jose	blonde	short	average	no	sunburned	Ricard, Karla	14065	285 Park Rd
Jose	blonde	short	average	no	sunburned	Ricard, Karla	28501	285 Park Rd
Ricard	ginger	average	heavy	no	sunburned	Carmen, Jose, Karla	26912	103 Potts St
Karla	brown	tall	heavy	no	none	Carmen, Jose, Ricard		
Isabel	brown	average	heavy	no	none	Manu	11995	103 Potts St
Manu	blonde	short	light	yes	none	Isabel	26912	24 Jones Av

Fig. 1. Different views at a data collection: the “classical” data mining view, the social network analysis and the network data mining view.

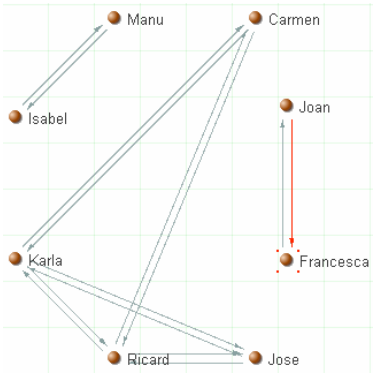
In addition to the explicitly coded relationships, there often are *implicit* relationships between the entities described by the data set, especially in the realm of transactions data. Any attribute can come into play for establishing such relations depending on the point of investigation. In our example in Fig. 1, the two attributes “Transaction reference #” and “Address” have been used to look for possible links between the college students. The revealing of such implicit relationships between entities and the discovery of ‘buried’ patterns in them is the focus of network data mining methods.

These different perspectives infer different sets of models. Fig. 2 uses the simple example in Fig. 1 to illustrate these differences. The classifier model in Fig. 2a is the well-known decision tree model (Dunham 2002). The tree shows a generalisation of the concept “student that gets sunburned” described in terms of the four input attributes in the table in Fig. 1, i.e. in terms of student height and weight, hair colour and whether they are in the habit of using lotion. The knowledge that we have discovered from the sample is that brunettes do not get sunburned regardless of whether they use lotion or not. The height and weight do not affect the result of whether you get sunburned or not. In some sense these facts are equivalent to a statistical summary of the data. As the aim is to identify general trends, the analysis and the model do not take in account the relationships between the entities described at the level of individual data points.

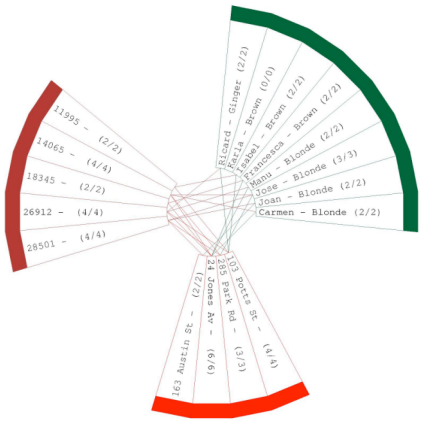
The structure of the relationships between the individual entities is revealed by the *network models*. The model in Fig. 2b is a social network based on the values of two columns in Fig. 1: “Name” and “Lives near”. It reveals possible relationships between Carmen, Ricard, Jose and Karla. Deeper analysis of these relationships may reveal why Carmen, Jose and Ricard are the only ones to get sunburned (e.g. it may turn out that it was Ricard’s influence to go to the beach at noon and stay there longer). On the other hand, it also reveals that there are two isolated groups of possible interest – Joan and Francesca, and Manu and Isabel (none of them got sunburned).



a. "Black-box" input-output generalization



b. Structure of a network model



c. Implicit relations between the entities linked through the values of some of their attributes. The links between students and the values of some attributes of interest ("Transaction reference #" and "Address" from Fig. 1) can reveal the existence of implicit relations.

Fig. 2. Illustration of the differences between predictive models based on generalization of the relation between the input and output attributes, and network models that take in account the relations between individual instances.

The model in Fig. 2c illustrates the implicit relations encoded through the transactions and street address. Note that this view reveals a heterogeneous network of relations between values of attributes.

Network models, which include the topology of the network and the characteristics of its nodes, links, and clusters of nodes and links, attempt to explain observed phenomena through the interactions of individual nodes or node clusters. During recent years there has been an increasing interest in these types of models in a number of research communities (see (Albert and Barabási 2002; Newman 2003) for extensive surveys of the research work on network models in different areas). Historically, sociologists have been exploring the social networks between people in different social settings. A typical social network analysis research scenario involves data collection through questionnaires or tables, where individuals describe their interactions with other individuals in the same social setting (for example, a club, school, organization, across organizations, etc.). Collected data is then used to infer a social network model in which nodes represent individuals and edges represent the interactions between these individuals. Classical social network analysis studies deal with relatively small data sets and look at the structure of individuals in the network, measured by such indices as centrality (which individuals have most links, can reach many others, are in a position to exert most influence, etc.) and connectivity (paths between individuals or clusters of individuals through the network). The body of literature is well covered by the following complementary books (Wasserman and Faust 1994; Scott 2000).

Works looking at the *discovery of network models* beyond "classical" social network analysis, date back to the early 1990s (for example, the discovery of shared interests based on the history of email communications (Schwartz and Wood 1993)). Recent years have witnessed the substantial input of physicists (Albert and Barabási 2002), mathematicians (Newman 2003) and organizational scientists (Borgatti 2003) to network research, with the focus shifting to large scale networks, their statistical properties and explanatory power, the discovery of such models and their use in explaining different phenomena in complex systems. The areas include a wide spectrum of social, biological, information, technological and other heterogeneous networks (see Table II in (Newman 2003) and Table I in (Albert and Barabási 2002), and other works of interest, (Krebs 2005) and (Batagelj and Mrvar 2003). Recent researches in the data mining community are looking at network models for predictive tasks (for example, predicting links in the model (Liben-Nowell and Kleinberg 2003), or the spread of influence through a network model (Domingos and Richardson 2001; Richardson and Domingos 2002; Kempe, Kleinberg and Tardos 2003). The interest towards link analysis and corresponding network models has increased during recent years, evidenced by the number of workshops

devoted to these topics, with a major focus on algorithms that work on graphs (for example, for example, see the presentations at recent workshops on link analysis at ACM KDD conference series[1] and SIAM Data Mining Conference[2]).

Such interest towards network models and new approaches to derive them has been driven largely by the availability of computing power and communication networks make possible the collection and analysis of these large amounts of data. Early social network research investigated networks of tens, at most hundreds of nodes. The networks that are investigated in different studies in present days tend to include millions (and more) links and nodes. This change of scale of the network models required change in the analytics approach (Newman 2003). Network Data Mining addresses this challenge.

However, the above mentioned efforts do not look at approaching the integrated data set and the process of facilitating discoveries from such data set. The network data mining approach is looking at this area. We define *network data mining as the process of discovering emergent network patterns and models in large and complex data sets.* The term denotes the methods and techniques in the following contexts:

- mining network models out of data sets
- mining network data (i.e. data generated by the interaction of the entities in a network, for example, in communications networks that can be the network traffic data).

The original data may not necessarily have been collected with the idea of building network models. The network patterns are derived from the integrated data set, which includes the interaction data and the descriptive attributes. Further in the chapter we briefly discuss the "loss of detail" problem and the "independency of attributes" assumption in knowledge discovery, present a human-centered knowledge discovery methodology that addresses these issues, and present a case study that illustrates the solutions that network data mining approach offers.

An illustrative comparison between predictive and network data mining is presented in Table 1.

Table 1. Comparison between predictive and network data mining

	Predictive data mining	**Network data mining**
Models and data	Predictive models derived from attributes data	Implicit network models derived from attributes data
Primary function	Prediction of outcomes	Discovery of irregularities
Level	Summarised data - often	Detailed elemental data
Perspective	Generalisation	Digging the details of interlinkage
Assumptions	a. Independence of attributes b. Independence of records	a. Linkage between attributes b. Linkage between records
Role of cognition	Little or none	Central and integral

The "loss of detail" problem in data mining

Data mining has been described as the art and science of teasing meaningful information and patterns out of large quantities of data – turning 'dusty' data that organisations have already collected into valuable information, operationally and strategically. Most data mining and analysis tools work by statistically summarising and homogenising data (Fayyad et al. 1996; Nong 2003), observing the trends and then looking for exceptions to normal behaviour. In addition, as pointed in (Fayyad 2003) "data mining algorithms are "knowledge-free", meaning they are brittle, and in real applications lack even the very basic "common sense reasoning" needed to recover even from simple situations. This process results in a *loss of detail* which, for intelligence and detection work, can defeat the whole purpose as it is often in the detail where the most valuable information is hidden. More generally, the identifying of exceptions to the 'norm' requires a top down-approach in which a series of correct assumptions needs to be made about what is normal and abnormal, and what will be applied to constitute a query. For many complex problems it can be difficult to even start this process since it is impossible to be specific about normal behaviour and what could or should constitute an exception.

The "independency of attributes" assumption in data mining

The "independency of attributes" assumption is accepted in some forms of data mining (for example, classifiers building algorithms like Naïve Bayes (Dunham 2002)). Under this assumption, the distributions of the values of the attributes are independent of each other. However, real-world data rarely satisfies the attribute value independence assumption. In fact, some data mining techniques like association and correlation analysis (Han and Kamber 2001), techniques that look at causality and the discovery causal networks (for example, Bayesian network models (Ramoni

[1] http://www-2.cs.cmu.edu/~dunja/LinkKDD2004/

[2] http://www-users.cs.umn.edu/~aleks/sdm04w/

and Sebastiani 2003)), make exactly the opposite assumption. Moreover, there are situations where the assumption sounds counterintuitive as well. For example, it is natural for the salary value to be correlated with the values of the age in a sample.

The logic is clear: by missing detail or making the wrong assumptions or simply by being unable to define what is normal, an organisation that relies solely upon predictive data mining may fail to discover critical information buried in its data.

Network data mining – the methodology

When there are few leads and only an open-ended specification on how to proceed with an analysis, *discovery* is all important. The discovery phase in the analysis process is too often overlooked or only implemented via exception based detection methods that are constrained to domain knowledge already known. Increased needs for automated detection have been accompanied by an increased reliance upon exception based forms of detection and rules based querying. However, along with the proliferation of data and increasingly advanced concealment tactics employed by the parties that want to avoid detection, there is a dilemma. We cannot write the rules ahead of time to specify a query or to write an exception (e.g. for an outlier, a threshold, an alert or an alarm) if the nature of what constitutes an exception and therefore the ability to specify relevant rules is unknown.

Network data mining is concerned with *discovering* relationships and patterns in linked data, i.e. the inter-dependencies between data items at the lowest elemental level. These patterns can be revealing in and of themselves, whereas statistically summarised data patterns are informative in different but complementary ways.

Similar to aspects of visual data mining (Wong 1999), network data mining integrates the exploration and pattern spotting abilities of the human mind with the processing power of computers to form a powerful knowledge discovery environment that is supposed to capitalise on the best of both worlds. This human-centred approach creates a powerful solution. However, to realise its full value the discovery phase needs to be repeated at regular intervals so that new irregularities that arise and variations on old patterns can be identified and fed into the *exception detection phase*.

The overall network data mining process is illustrated in Fig. 3. In network data mining, the data miner is in a role similar to Donald Schön's "reflective practitioner" (Schön 1983; Schön 1991), originally developed in the analysis of design processes. In his later work (Schön, 1991), Schön has presented a number of examples of disciplines that fit a process model with characteristics similar to the design process. What is relevant to our claim is Schön's view that designers put things together (in our case, the miner replaces the designer, and the things s/he puts together are the visual pieces of information) and create new things (in our case the miner creates new chain of inquiries interacting with the views) in a process involving large amount of variables (in our case these are the attributes and the linkages between them). Almost anything a designer does, involves consequences that far exceed those expected. In network data mining approach the inquiry techniques may lead to results that far exceed those expected and that in most cases may change the line of the analysis of the data into an emerging path. Design process is a process which has no unique concrete solution. In network data mining we operate with numerous network slices of the data set, assisting in revealing the different aspects of the phenomena. Schön also states that he sees a designer as someone who changes an indefinite situation into a definite one through a reflexive conversation with the material of the situation. By analogy, the network data miner serves to change the complexity of the problem through the reflexive investigative iterations over the integrated data set. The reflective step is a revision of the reference framework taken in the previous step in terms of attributes selection, set of visual models and corresponding guiding techniques, and the set of validation techniques.

The main methodological steps and accompanying assumptions of network data mining (NDM) approach include:

Sources of data and modelling: NDM provides an opportunity to integrate data so as to obtain a single address space, a common view, for disparately sourced data. Hence, a decision as to which sources are accessible and most relevant to a problem is an initial consideration. Having arranged and decided the sources, the next question relates to modelling. Which fields of data should serve as entities/nodes (and attributes on the entities) and which should serve as links (and attributes on the links)[3]. Multiple data models can and often are created to address a particular problem.

Visualization: The entities and a myriad of linkages between them must be presented to screen in meaningful and color coded ways so as to simplify and facilitate the discovery of underlying patterns of data items that are linked to other items and on-linked to still other items. This is especially so with large volumes of data, e.g. many hundreds of thousands or millions of links and entities which the user must be able to easily address and then readily make sense of from an interpretative and analytical point of view.

[3] A tool that interfaces to relational databases and any original sources of data (e.g. XML files) is basic to NDM, and is a capability provided in the NetMap software suite which is a premier technology in this space and was used in this case presented later in this chapter.

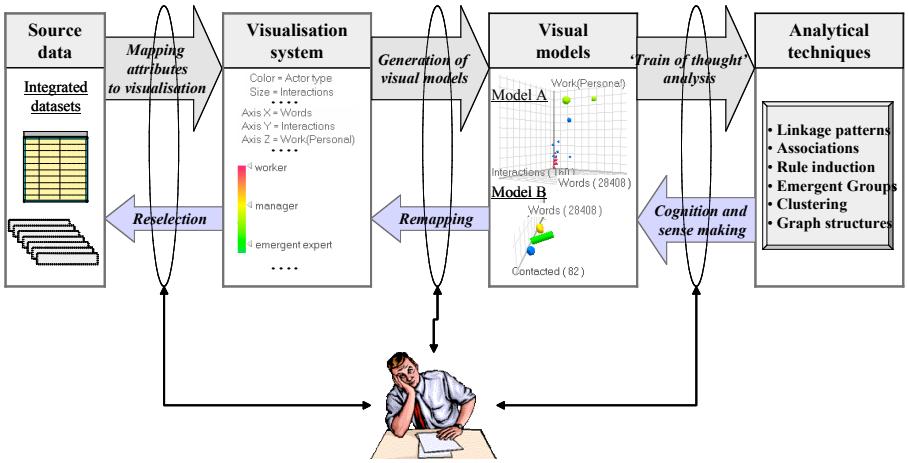


Fig. 3. Network data mining as a human-centered knowledge discovery process.

The main methodological steps and accompanying assumptions of network data mining (NDM) approach include:

Sources of data and modelling: NDM provides an opportunity to integrate data so as to obtain a single address space, a common view, for disparately sourced data. Hence, a decision as to which sources are accessible and most relevant to a problem is an initial consideration. Having arranged and decided the sources, the next question relates to modelling. Which fields of data should serve as entities/nodes (and attributes on the entities) and which should serve as links (and attributes on the links)[4]. Multiple data models can and often are created to address a particular problem.

Visualization: The entities and a myriad of linkages between them must be presented to screen in meaningful and color coded ways so as to simplify and facilitate the discovery of underlying patterns of data items that are linked to other items and on-linked to still other items. This is especially so with large volumes of data, e.g. many hundreds of thousands or millions of links and entities which the user must be able to easily address and then readily make sense of from an interpretative and analytical point of view.

'Train of thought' analysis: Linkage between data items means that the discovery of patterns can be a process whereby the analyst uses the reflective practitioner approach mentioned earlier. Explicit querying is less often the case; rather the analyst may let the intuition guide him or her. For example, "Why are all those red links going over there?", "What are they attached to, and in turn what are they attached to?" Such 'train of thought' processes invariably lead the analyst to discover patterns or trends that would not be possible via more explicit querying or exception based approaches – for the specification for the queries is not known.

Cognition and sense-making: An integral assumption in NDM is that the computer in the analyst's mind is more powerful by orders of magnitude that the one on the desktop. Hence, intuition and cognition are integral, and need to be harnessed in the analytical process especially at the discovery phase in those cases where there is only limited domain knowledge as a guide to analysis and understanding.

Discovery: An emergent process, not prescriptive one. It is not possible to prescribe ahead of time all the query rules and exception criteria that should apply to a problem, if domain knowledge is less than perfect. And of course in many if not most cases it is, otherwise the problem would already have been solved. By taking an emergent or bottom up approach to what the data are 'saying', patterns and linkages can be discovered in a way that is not too different from 'good old fashioned' policing, where curiosity and intuition have always been integral in the ability to discover the facts, then qualifying them and solving the crime.

Finding patterns that can be re-discovered: Any linkage *pattern* observed on screen is simply that, an observation of potential interest. In the context of retail NDM for example, any sales assistant with a high ratio of refunds to sales (statistically flagged) might attract attention. In a case in point known to the authors, the perpetrators of a scam knew about such exception criteria. As longer term employees "in the know", they could easily duck under them. They had taken it in turns to report levels of refunds always just under the limits no matter what the limits were varied to over an extensive period. NDM was able to show collusive and periodic reporting linkages to supervisors – patterns discovered through visualization and algorithms that facilitate the intuition. Such patterns are of particular interest, and in fact often an objective of the network mining approach. They are termed *scenarios* and characterised as definable and re-usable patterns. Their value is that they are patterns that have now become 'known'. Hence they can be defined, stored in a knowledge base, and applied at the front end of a process as, for example, in a predictive modelling environment. The important methodological step is that the definitions need to be discovered in the first place before they can be further applied.

Network data mining is complementary to statistical summarizing and exception detection data mining. Network data mining is particularly useful in the *discovery phase*, of finding things previously unknown. Once discovered, particular patterns and abnormal behaviors and exceptions are of course able to be better defined, and these scenarios

[4] A tool that interfaces to relational databases and any original sources of data (e.g. XML files) is basic to NDM, and is a capability provided in the NetMap software suite which is a premier technology in this space and was used in this case presented later in this chapter.

can then be saved and applied automatically across new volumes of data, including by the feeding of discovered scenarios back into traditional data mining tools. The section below illustrates the network data mining approach using a real-world case study

Example of network data mining approach in fraud detection

Many cases could be used to illustrate a network data mining approach and each case would be instructive of different features. Nonetheless, this particular case is not atypical of the methods involved in knowledge discovery in network data mining. The case is presented as a reflection on the steps that the analyst did. The presentation is structured along the main methodological steps that we have identified in the previous section.

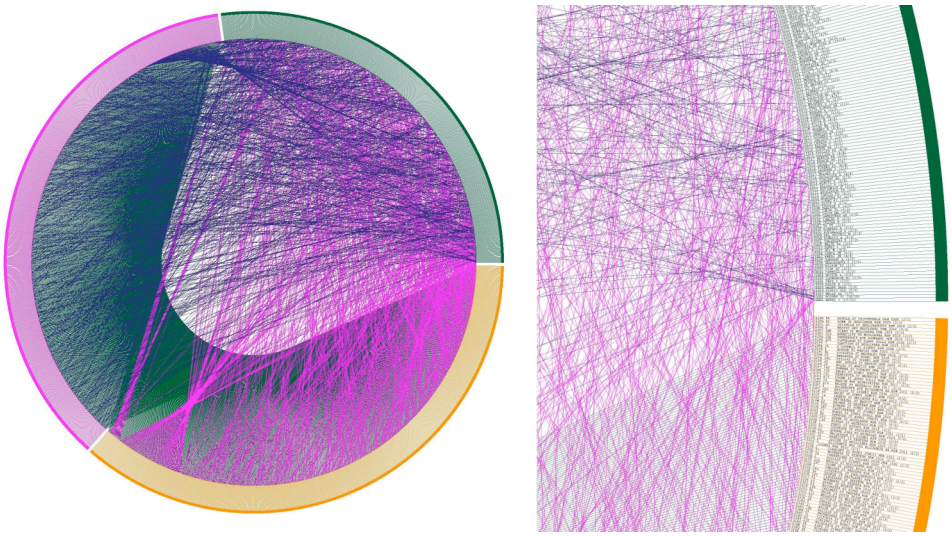
Sources of data and modelling: The case involved analysis of approximately twelve months of motor vehicle insurance claims from one company. A small set of five thousand records were available as a pilot project from one state of Australia. Note that all the information presented here has been de-identified.

The brief from the company was essentially open-ended and without specification. A concern was expressed that there *could* be suspicious transactions and perhaps fraudulent activity but there were no known persons or transactions of interest. The case was clearly in the realm of 'discovery'.

Visualization: The study used the NetMap software from NetMap Analytics. The analyst first built a set of linkages from the available fields. These were between persons, addresses, claim numbers, telephone numbers, and bank accounts into which claims monies had been paid.

Initially the analyst only looked at three fields of data and the linkages identified between them, as shown in Fig. 4. To start to make sense of the data overviewed in Fig. 4, the analyst then processed the data through an algorithm in NetMap to produce the display of potential irregularities shown in Fig. 5 (also close-ups in Fig. 6).

Cognition and sense-making (and 'Train of thought' analysis): What appeared to be 'regular' patterns were observed. These were seen to be small triangles of data comprised of a person, a claim number and an address, all fully inter-linked. The explanation as to why the little triangles appeared to be 'regular' patterns in the data was simple after the analyst had stopped to think about it: most people just had one claim and one address. In comparison, the 'bumps' looked as though they were 'irregularities'. The 'bumps' comprised persons linked to multiple claims and/or addresses. By taking the seemingly regular patterns out of the picture (the triangles), the analyst was quickly able to produce a short list of potentially suspicious transactions and inter-related behaviours. That is, from a larger amount of data she was able to quickly focus *where* to look in the myriad of linkages to drill down for more details.



a. Overview of links between persons, addresses and claim numbers

b. Close up at approximately 3 o'clock of the linked data shown in Fig. 4a.

Fig. 4. An illustration of an initial macro view of linkages in the data with a subsequent step in digging deeper in the details

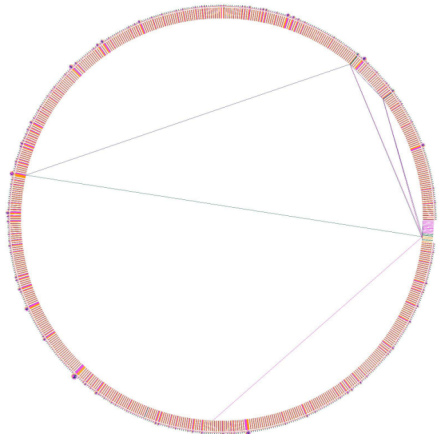
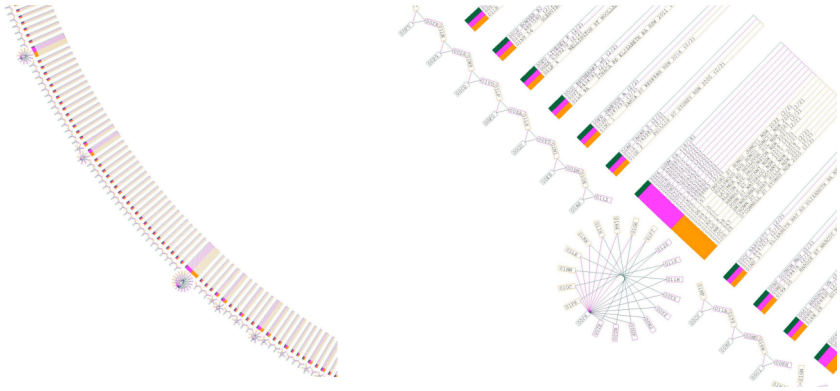


Fig. 5. Data from Fig. 4 processed to show certain patterns of irregularities (see detail in Fig. 6)



a. Overview of links between persons, addresses and claim numbers

b. Overview of links between persons, addresses and claim numbers

Fig. 6. Close up segments from approximately 4 o'clock in Fig. 5 showing what appeared to be regular patterns (the small triangles of linkage) and also irregularities (the larger 'bumps')

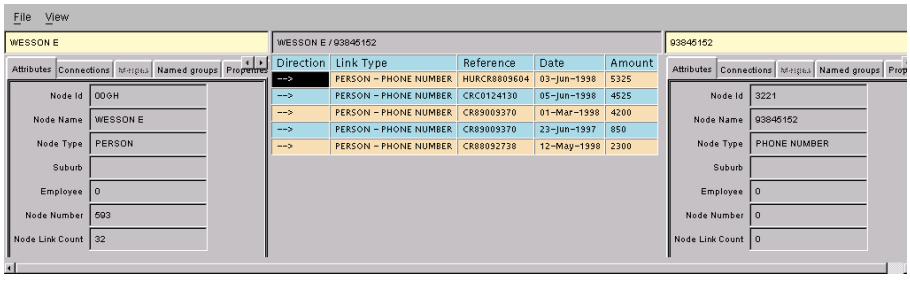


Fig. 7. More detailed information underlying any 'entity' or 'link' was able to be accessed by clicking on that data element. The analyst could then quickly qualify observed patterns.

Discovery (and 'Train of thought' analysis): Other and more interesting types of 'irregular' patterns in this case were those seen to be extending across the middle of the main circle in Fig. 5. It was clear to the analyst that most of the data items did not have links across the middle and, hence, these were 'irregularities' of some sort. They seemed to emanate from about the 3 o'clock position in Fig. 5. Accordingly, the analyst zoomed into the display at about 3 o'clock and selected one of the inter-linked data items for 'step-link' purposes. She selected Simons JL, but it could have been any of these data items since several were linked to each other. The beginning of this step is shown below in Fig. 8.

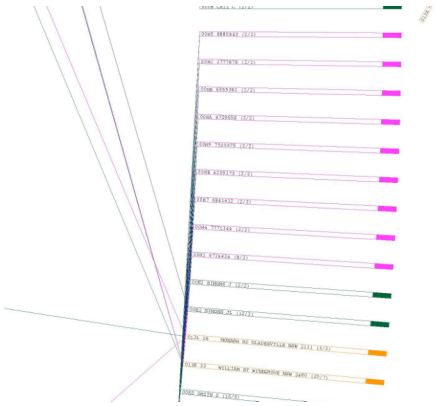


Fig. 8. The analyst chose one of these inter-linked data items to 'step out' from (Simons JL)

Finding patterns that can be re-discovered: Step-linking from a selected data item follows the network links out through any number of degrees of separation. As shown in Fig. 9 and Fig. 10, the analyst stepped out from JL Simons to 'infinity' degrees of separation, the net effect being to bring to screen all of that party's indirect linkage. Obviously, if further sources of data could have been added (which is often done but additional sources were unavailable in this case) richer indirect linkages would have most likely further assisted the inquiry. Also, although the analyst did not do so in this case, destinations of data items may be specified and then stepped to from chosen source data items, thereby determining whether certain parties are linked at all and if so who or what are the intermediaries.

Fig. 9. Step-linking from the selected entity to infinity degrees of separation

Fig. 10. Close up of portion of the data shown in Fig. 9.

The particular extract of linked data in Fig. 9 and Fig. 10 was then processed through an algorithm unique to NetMap called *emergent groups*. The resulting pattern is shown in Fig. 11.

Discovery: From the extract of data in Fig. 11, four emergent groups were identified as shown in Fig. 12. An emergent group comprises closely inter-related data items; they have more links within the group than outside to any other group. Thus, they are defined 'out of' the data in a bottom-up way ('what are the data telling us?'), rather than prescriptively by rules or queries which in this case would have been impossible - there was simply no prior knowledge about how to frame such rules or program any queries. The analyst was squarely in 'discovery' mode.

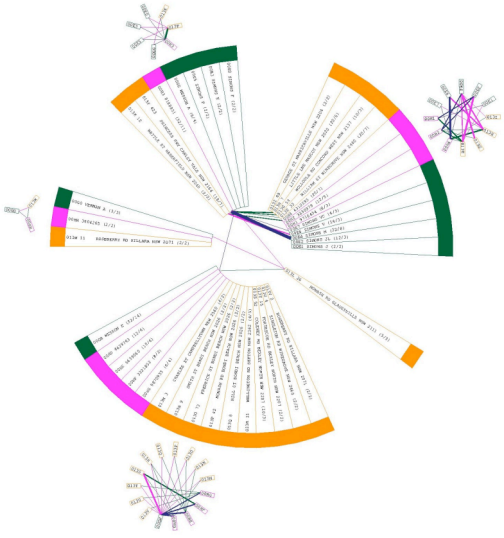


Fig. 11. Emergent groups discovered within the extract of data shown in Fig. 9

Fig. 12. Close-up views of the emergent groups shown in Fig. 11

'Train of thought' analysis: The emergent group on the right in Fig. 12 comprised five people called Simons, three claims and four addresses, all closely inter-related (relationships between id codes are shown in the satellite, and the id codes plus full names are shown in an the arc on the main circle). They were linked across to another group at 11 o'clock comprised of more people called Simons and somebody called Wesson. That Wesson (initial A) was linked down to the group at 5 o'clock to E Wesson via a common claim. That in turn took the analyst over to the address at 4 o'clock and then to an 'A Verman' at 9 o'clock.

This 'train of thought' analysis led the analyst to Verman. Her intuition indicated she wanted to look at Verman more closely although she could not have been specific as to why. To cut a long story short, when Verman was investigated and went to jail it was learned that he had originally had a 'regular' pattern (comprising one claim and one address – see Fig. 6). He reckoned this was a good way to make money. So, he recruited the Simons and the Wessons as the active parties in a scam of staged vehicle accidents while he tried to lie as low as he could. He was careless however, since he had left a link to another address (the one at 4 o'clock in Fig. 11 – note, he must have been a witness or a passenger). This link indirectly implicated him back into the activity that involved the Simons and Wessons.

Finding patterns that can be re-discovered: The analyst did not know this of course at the time, but could sense that Verman was a few steps removed from the activity of the Simons and thought she would like to quickly qualify this person. She firstly took Verman and stepped out to obtain all his indirect linkage (see Fig. 13).

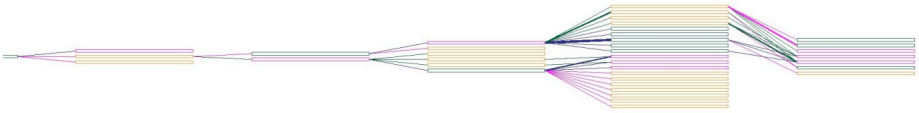


Fig. 13. A potential suspect on the left (Verman) with all his indirect linkage to other data items

The analyst then sought to 'enrich' the linkage of Verman, i.e. to regard him as a potential case that she would like to qualify on the spot if possible, by adding extra linkage (as shown in Fig. 14). In this case, she only had two extra fields available: bank account information and telephone numbers. Nonetheless, she quickly discovered one extra and crucial link that helped her to qualify Verman – one of his two telephone numbers was also linked to A Wesson (see Fig. 15). That additional link provided the 'tipping point', the extra knowledge that gave her sufficient confidence to recommend that Verman be investigated. This subsequently led to his arrest and conviction on fraud charges.

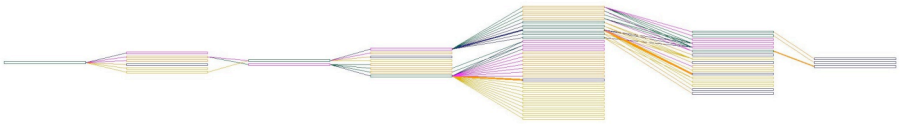


Fig. 14. Enrichment of the linkage in Fig. 13 by adding in extra fields of data.

It transpired that Verman had been the mastermind behind a claims scam totalling approximately $150,000. He could not have been discovered by traditional data mining methods since no domain knowledge existed that could help define any exception rules or serve as inputs to SQL queries. If red flags or alerts had been applied as the only investigative approach, he would have escaped detection since he was essentially 'regular' and not unusual in anyway. The use of detailed linkage and the easy ability to navigate and make sense of it made the difference.

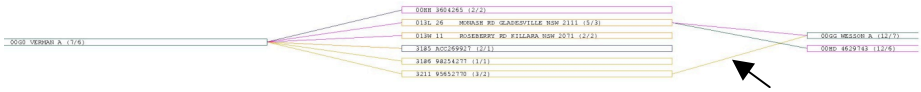


Fig. 15. Close up of portion of Fig. 14 showing the extra 'tell tale' link (arrowed: a telephone number in common). This extra information led to an investigation and then the arrest and successful prosecution of Verman, the person shown on the left.

Conclusion

This chapter has described the concept of network data mining and presented a case study by way of illustrating its real-world implementation and its distinction from more traditional approaches to data mining, and also its distinction from social network analysis.

Network data mining focuses upon knowledge discovery. It involves a human-centred process which harnesses the intuitive powers of the human intellect in conjunction with color-coded linkage patterns and unique algorithms to facilitate the intuition, a process referred to as 'train of thought' analysis.

We can summarize the main steps in the knowledge discovery process (Fayyad et al. 1996) as follows:

1. Define scenarios in terms of query specifications and exception rules
2. Process the data
3. Interpret or initiate action

The discovery phase, which network data mining gives emphasis to, logically precedes and feeds into step 1. This prior step we refer to as step 0:

0. Discover patterns and qualify them as scenarios.

Note also that discovery (step 0) and exception detection (step 1) are inter-related. Once patterns of interest have been discovered then they can defined and this information incorporated in more conventional exception detection and querying. However, many situations require that discovery takes place first. It then needs to be applied to the same sets of data and new sets since endless variations and inventive ways of concealing nefarious activity and avoiding detection are put into play that need to be discovered in order to keep ahead of the game.

We presented a case which illustrated the cornerstones of the network data mining approach. The party in the case who was eventually convicted (Verman) would have escaped detection if a traditional data mining approach had been used. He had only had one claim, no 'red flag' information was involved, and nothing particularly anomalous occurred with respect to him. By all accounts he would have slipped under any exception detection processes.

Successful discovery did occur through the use of the network data mining methods outlined. Train of thought analysis enabled a discovery to be made that would have been highly unlikely otherwise. Querying could have not have been successful since there is no way that a relevant query could have been framed. Exception rules could not have been specified since there was no information as to what could constitute an exception that would have discovered Verman. He was essentially below the radar.

Masterminds concealing their behavior, tend to be as unobtrusive as possible. They also often know the rules and the exceptions and know what they need to do if the rules and exceptions change so as to avoid being caught. Hence, any discovery tool must go beyond programmatic and prescriptive exception based detection.

A complementary usage of traditional data mining could have in principle been used in this case as follows. Several of the underlings recruited by Verman had the same family name, Simons. Hence, the name Simons appeared more would have been the case in terms of its occurrence, say, in the telephone directory. Therefore, a rule could have been to display all names occurring more often than expected in the telephone population. The result being that the name Simons would be flagged. In a complementary approach, this flagging would be a helpful initial task. The next task would be within an NDM linked data environment to 'step out' several steps from the Simons and so commence the discovery process from that point.

This twin approach has been used to great effect. It essentially uses statistical approaches to flag where the detailed linkage should be examined by discovery-oriented 'train of thought' analysis. In numerous cases we are aware of this combined approach has leveraged the best of both network and non-network data mining.

References

Albert, R. and A.-L. Barabási (2002): "Statistical mechanics of complex networks." *Reviews of Modern Physics* **74**(January 2002), 47-97.

Antonie, M.-L., O. R. Zaiane, et al. (2003): Associative classifiers for medical images. *Mining Multimedia and Complex Data*. O. R. Zaiane, S. J. Simoff and C. Djeraba. Heidelberg, Springer, 68-83.

Batagelj V. and A. Mrvar (2003): *Pajek - Analysis and Visualization of Large Networks*. In: Jnger M and Mutzel P, eds. Graph Drawing Software. Berlin: Springer, 2003.

Borgatti, S. P. (2003): "The network paradigm in organizational research:

A review and typology." *Journal of Management* **29**(6), 991-1013.

Domingos, P. and M. Richardson (2001): Mining the network value of customers. *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, ACM Press, 57-66.

Dunham, M. H. (2002): *Data Mining: Introductory and Advanced Topics*, Prentice Hall.

Fayyad, U. M. (2003): "Editorial." *ACM SIGKDD Explorations* **5**(2), 1-3.

Fayyad, U. M., G. Piatetsky-Shapiro, et al. (1996): From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining*. U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy. Cambridge, Massachusetts, AAAI Press/The MIT Press, 1-34.

Han, J. and M. Kamber (2001): *Data Mining: Concepts and Techniques*. San Francisco, CA, Morgan Kaufmann Publishers.

Hand, D., H. Mannila, et al. (2001): *Principles of Data Mining*. Cambridge, Massachusetts, The MIT Press.

Kempe, D., J. Kleinberg, et al. (2003): Maximizing the spread of influence through a social network. *Proceedings ACM KDD2003*, Washington, DC, ACM Press

Krebs, V. (2005): http://www.orgnet.com

Klösgen, W. and J. M. Zytkow, Eds. (2002). *Handbook of Data Mining and Knowledge Discovery*, Oxford University Press.

Liben-Nowell, D. and J. Kleinberg (2003): The link prediction problem for social networks. *Proceedings CIKM'03*, November 3–8, 2003, New Orleans, Louisiana, USA., ACM Press

Newman, M. E. J. (2003): "The structure and function of complex networks." *SIAM Review* **45**, 167-256.

Nong, Y., Ed. (2003). *The Handbook of Data Mining*. Mahwah, New Jersey, Lawrence Erlbaum Associates.

Nong, Y. (2003): Mining computer and network security data. *The Handbook of Data Mining*. Y. Nong. Mahwah, New Jersey, Lawrence Erlbaum Associates, 617-636.

Ramoni, M. F. and P. Sebastiani (2003): Bayesian methods for intelligent data analysis. *Intelligent Data Analysis: An Introduction*. M. Berthold and D. J. Hand. New York, NY, Springer, 131-168.

Richardson, M. and P. Domingos (2002): Mining knowledge-sharing sites for viral marketing. *Proceedings of the Eighth International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, ACM Press, 61-70.

Schön, D. (1983): *The Reflective Practitioner*. New York, Basic Books.

Schön, D. (1991): *Educating The Reflective Practitioner*. San Francisco, Jossey Bass.

Schwartz, M. E. and D. C. M. Wood (1993): "Discovering shared interests using graph analysis." *Communications of ACM* **36**(8), 78-89.

Scott, J. (2000): *Social Network Analysis: A Handbook*. London, Sage Publications.

Scott, J. (2000): *Social Network Analysis: A Handbook*. London, Sage Publications.

Wasserman, S. and K. Faust (1994): *Social Network Analysis: Methods and Applications*. Cambridge, Cambridge University Press.

Weiss, S. M. and T. Zhang (2003): Performance analysis and evaluation. *The Handbook of Data Mining*. Y. Nong. Mahwah, New Jersey, Lawrence Erlbaum Associates.

Wong, P. C. (1999): "Visual Data Mining." *IEEE Computer Graphics and Applications* **September/October**, 1-3.