# Data Mining Case Studies

## Proceedings of the Second International Workshop on Data Mining Case Studies

held at the Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining in San Jose CA


Edited by

**Brendan Kitts**, Microsoft

**Gabor Melli,** Simon Fraser University


Elder Research, Inc.
*Data Mining & Pattern Discovery*


**ISBN 0-9738918-2-3**

**The Association for Computing Machinery, Inc.**

**1515 Broadway**

**New York, New York 10036**



Additional copies may be ordered prepaid from:

ACM Order Department
PO Box 11405
Church Street Station
New York, NY 10286-1405

Phone 1-800-342-6626 (USA and Canada)
+1-212-626-0500
(all other countries)
Fax: +1-212-944-1318
Email: acmhelp@acm.org

# Contents

# Organizers

**Chairs**

Brendan Kitts, Microsoft
Gabor Melli, Simon Fraser University

**Prize Committee**

John Elder, PhD., Elder Research
Brendan Kitts, Microsoft
Gabor Melli, Simon Frasier University

**Program Committee**

Gregory Piatetsky-Shapiro, PhD., KDNuggets
Richard Bolton, PhD., KnowledgeBase Marketing, Inc.
Pip Courbois, PhD., Amazon
Simeon J. Simoff, PhD., University of Technology Sydney
Gang Wu, PhD., Microsoft
Jing Ying Zhang, PhD., Microsoft
Teresa Mah, PhD., Microsoft
Tom Osborn, PhD., Verism Inc.
Ed Freeman, Washington Mutual
Brendan Kitts, Microsoft
Gabor Melli, Simon Fraser University
Luis Adarve-Martin, Microsoft
Karl Rexer, PhD., Rexer Analytics
John Elder, PhD., Elder Research

**Sponsors**

Elder Research Inc. (ERI)
Association for Computing Machinery (ACM)

# Participants

| | |
|---|---|
| *Abdul Qadeer Khan Rajput* | Mehran University of Engineering and Technology |
| *Brendan Kitts* | Microsoft |
| *C. Perlich* | IBM T.J. Watson Research Center |
| *Chris Curry* | Open Data Group |
| *Claus Neubauer* | Siemens Corporate Research |
| *David L. Hicks* | Aalborg Universitet Esbjerg |
| *David Locke* | Open Data Group |
| *Doug Bryan* | KXEN |
| *Ed Freeman* | Washington Mutual |
| *Edward "Ned" Porter* | US Census Bureau |
| *Fabian Moerchen* | Siemens Corporate Research |
| *Françoise Fogelman Soulié* | KXEN |
| *Gabor Melli* | Simon Fraser University |
| *Gang Wu* | Microsoft |
| *Henrik Legind Larsen* | Aalborg Universitet Esbjerg |
| *I. Khabibrakhmanov* | IBM T.J. Watson Research Center |
| *John Elder* | Elder Research Inc. |
| *Joseph Bugajski* | Visa International |
| *Karl Rexer* | Rexer Analytics |
| *Ken Reed* | LowerMyBills.com |
| *Klaus Brinker* | Siemens Corporate Research |
| *Michael Ikeda* | US Census Bureau |
| *Nasrullah Memon* | Aalborg Universitet Esbjerg |
| *Patricia B. Cerrito* | University of Louisville |
| *R. Lawrence* | IBM T.J. Watson Research Center |
| *Robert Grossman* | Open Data Group |
| *S. Mahatma* | IBM T.J. Watson Research Center |
| *S. Rosset* | IBM T.J. Watson Research Center |
| *S. Weiss* | IBM T.J. Watson Research Center |
| *Steve Vejcik* | Open Data Group |
| *Tom Osborn* | Thought Experiments |

# Corporate Contributors

Aalborg Universitet Esbjerg
Elder Research Inc.
IBM T.J. Watson Research Center
KXEN
LowerMyBills.com
Mehran University of Engineering and Technology
Microsoft Corporation
Open Data Group
Rexer Analytics
SAS Corporation
Siemens Corporate Research
Simon Fraser University
Thought Experiments
University of Louisville
US Census Bureau
Visa International
Washington Mutual

# Sponsors

Elder Research Inc.

Elder Research is a leader in the practice of Data Mining -- discovering useful patterns in data and successfully harnessing the information gained. The principals are active researchers in Data Mining, contributing to the literature of this emerging field in books, conferences, and through highly-regarded short courses and training seminars. Further details can be found at http://www.datamininglab.com

ACM – The Association for Computing Machinery

The Association for Computing Machinery is an international scientific and educational organization dedicated to advancing the arts, sciences, and applications of information technology. With a world-wide membership ACM is a leading resource for computing professionals and students working in the various fields of Information Technology, and for interpreting the impact of information technology on society.

ACM is the world's oldest and largest educational and scientific computing society. Since 1947 ACM has provided a vital forum for the exchange of information, ideas, and discoveries. Today, ACM serves a membership of computing professionals and students in more than 100 countries in all areas of industry, academia, and government. ACM's 34 Special Interest Groups (SIGs) address the varied needs of today's IT and computing professionals, including computer graphics, human interfaces, artificial intelligence, data mining, mobile communications, computer education, software engineering, and programming language. Each SIG is organized around specific activities that best serve its practitioner and research-based constituencies. Many SIGs sponsor leading conferences and workshops, produce newsletters and publications, and support email forums for information exchange. ACM can be found on the web at http://www.acm.org


# The Data Mining Case Studies Workshop

From its inception the field of Data Mining has been guided by the need to solve practical problems. Yet few articles describing working, end-to-end, real-world case studies exist in our literature. Success stories can capture the imagination and inspire researchers to do great things. The benefits of good case studies include:

1. Education: Success stories help to build understanding.
2. Inspiration: Success stories inspire future data mining research.
3. Public Relations: Applications that are socially beneficial, and even those that are just interesting, help to raise awareness of the positive role that data mining can play in science and society.
4. Problem Solving: Success stories demonstrate how whole problems can be solved. Often 90% of the effort is spent solving non-prediction algorithm related problems.
5. Connections to Other Scientific Fields: Completed data mining systems often exploit methods and principles from a wide range of scientific areas. Fostering connections to these fields will benefit data mining academically, and will assist practitioners to learn how to harness these fields to develop successful applications.

The *Data Mining Case Studies Workshop* was established in 2005 to showcase the very best in data mining case studies. We also established that *Data Mining Practice Prize* to attract the best submissions, and to provide an incentive for commercial companies to come into the spotlight.

This first workshop was followed up by a *SIGKDD Explorations Special Issue on Real-world Applications of Data Mining* edited by Osmar Zaiane in 2006. It is our pleasure to continue the work of highlighting significant industrial deployments with the *Second Data Mining Case Studies Workshop* in 2007.

Like its predecessors, Data Mining Case Studies 2007 has highlighted data mining implementations that have been responsible for a significant and measurable improvement in business operations, or an equally important scientific discovery, or some other benefit to humanity. Data Mining Case Studies papers were allowed greater latitude in (a) range of topics - authors may touch upon areas such as optimization, operations research, inventory control, and so on, (b) page length - longer submissions are allowed, (c) scope - more complete context, problem and solution descriptions will be encouraged, (d) prior publication - if the paper was published in part elsewhere, it may still be considered if the new article is substantially more detailed, (e) novelty - often successful data mining practitioners utilize well established techniques to achieve successful implementations and allowance for this will be given.

# The Data Mining Practice Prize

## Introduction

The Data Mining Practice Prize is awarded to work that has had a significant and quantitative impact in the application in which it was applied, or has significantly benefited humanity. All papers submitted to Data Mining Case Studies will be eligible for the Data Mining Practice Prize, with the exception of members of the Prize Committee. Eligible authors consent to allowing the Practice Prize Committee to contact third parties and their deployment client in order to independently validate their claims.

## Award

Winners and runners up receive an impressive array of honors including

a. Plaque awarded at the KDD conference General Session on August 12th 2007.
b. Prize money comprising $500 for first place, $300 for second place, $200 for third place, donated by Elder Research.
c. Winner announcements to be published in the journal SIGKDD Explorations
d. Awards Dinner with organizers and prize winners.

We wish to thank Elder Research, their generous donation of prize money, incidental costs, time and support, and the ACM for making our competition and workshop possible.

# Analytics-driven solutions for customer targeting and sales force allocation[1]

R. Lawrence, C. Perlich, S. Rosset, I. Khabibrakhmanov, S. Mahatma, S. Weiss
IBM T.J. Watson Research Center
Route 134 / PO Box 218
Yorktown Heights, NY 10598
{ricklawr,perlich,rosset,ildar,mahatma,sholom}@us.ibm.com

## ABSTRACT
Improving sales force productivity is a key strategic priority to drive corporate revenue growth. Sales professionals need to be able to easily identify new sales prospects, and sales executives need to ensure that the overall sales force is deployed against the best future revenue-generating sales accounts. In this paper, we describe two analytics-based solutions developed within IBM to address these related issues. The first initiative, OnTARGET, provides a set of analytical models designed to identify new sales opportunities at existing client accounts as well as non-customer ("whitespace") companies. The OnTARGET models estimate the probability of purchase at the product-brand level, and use training examples drawn from historical transactions, with explanatory features extracted from transactional data joined with company firmographic data. The objective of the second initiative, the Market Alignment Program (MAP), is to drive the sales allocation process based on field-validated analytical estimates of future revenue opportunity in each operational market segment. The estimates of revenue opportunity are generated by defining the opportunity as a high percentile of a conditional distribution of the customer's spending, i.e., what we could "realistically hope" to sell to this customer. We describe the development of both sets of analytical models, as well as the underlying data models and web sites used to deliver the overall solution. We conclude with a discussion of the business impact of both initiatives.

## Categories and Subject Descriptors
H.2.8 [**Database Management**]: Database applications – *data mining*

## General Terms
Algorithms, Management, Performance



## Keywords
Propensity models, Quantile estimation, Customer wallet estimation, Sales efficiency

## 1. INTRODUCTION
Improving sales productivity is an essential component of driving organic growth for many major companies today. While hiring the best sales representatives is an obvious first step, it is increasingly recognized [1] that the realization of the true potential of any sales force requires that sales reps and executives be equipped with relevant IT-based tools and solutions. The past decade has seen the development of a number of customer relationship management (CRM) systems [2, 3] that provide integration and management of data relevant to the complete marketing and sales process. Sales force automation (SFA) systems [4] enable sales executives to better balance sales resources against identified sales opportunities. While it is generally (but not uniformly [5]) accepted that such tools improve the overall efficiency of the sales process, major advances in sales force productivity require not only access to relevant data, but informative, predictive analytics derived from this data.

In this paper, we develop analytical approaches to address two issues relevant to sales force productivity, and describe the deployment of the resulting solutions within IBM. The first solution addresses the problem faced by sales representatives in identifying new sales opportunities at existing client accounts as well as at non-customer ("whitespace") companies. The analytical challenge is to develop models to predict the likelihood (or propensity) that a company will purchase an IBM product, based on analysis of previous transactions and other available third-party data. These modeling results, along with the underlying data, have been integrated in a web-based tool called OnTARGET. A second, but related business challenge is to provide quantitative insight into the process of allocating sales reps to the best potential revenue-generation opportunities. In particular, we are interested in the allocation of resources to existing IBM client accounts. Here, the analytics challenge is to develop models to estimate the true revenue potential (or opportunity) at each account within IBM product groups. These models were developed as part of an internal initiative called the Market Alignment Program (MAP), in which the model-estimated revenue opportunities were validated via extensive interviews

[1] An earlier version of this paper appeared online in July, 2007 in the IBM Systems Journal – see http://www.research.ibm.com/journal/sj/

with front-line sales teams. We describe this process, and the web-based MAP tool, later in this paper.

Although they address different business problems, the OnTARGET and MAP tools share a common architecture. Both employ a data model that effectively joins historical IBM transaction data with external third-party data, thereby presenting a holistic view of each client in terms of their past history with IBM as well as their external "firmographic" information like sales, number of employees, and so on. Both systems exploit this linked data to build the models described above and in the sections below. Given the different business objectives, the tools employ different web-based user interfaces; however, both interfaces are designed to facilitate easy navigation and location of the relevant analytical insights and underlying data.

In the following section, we describe the OnTARGET project, its data model, and overall system design motivated by the business requirements. We then describe the propensity models at the heart of OnTARGET. Turning to the MAP project in Section 4, we discuss the MAP business process, and describe the differences in the MAP tool design relative to OnTARGET, again motivated by the different business objectives of MAP. Section 5 describes the MAP revenue-opportunity models. Finally, we describe the deployment of these systems, and discuss the operational impact against their respective business objectives.

# 2. OnTARGET: A Customer Targeting Solution

In this section, we begin with a discussion of the OnTARGET business objectives, describe the overall system design and data model developed to meet these objectives, and conclude with a description of the web-based user interface to the OnTARGET tool.

## 2.1 OnTARGET Business Objectives

After the collapse of the Internet bubble, corporate growth has returned to what are likely to be normal growth rates for the next few years. Since the broad market is likely to grow in aggregate at rates only slightly higher than GDP, companies will need to generate organic revenue growth at rates greater than the market overall to remain competitive. One approach is to pursue growth opportunities in emerging markets. But it is also necessary to generate significant growth in a company's core businesses and markets. This requires a renewed focus on identifying and closing new sales opportunities with existing clients, as well as finding new companies that will be receptive to the company's core offerings. Improving sales force productivity is essential to both objectives.

Early in the OnTARGET project, we spoke to a number of leading sales professionals and sales leaders about potential IT-enabled tools that they believed could enhance sales productivity. One common sentiment is that sales people are often forced to use multiple tools and processes that not only fail to provide the relevant information needed to do their jobs better, but also take valuable time away from actual sales activities. While some cross-sell models were available for use by the sales teams, these analytics were often delivered via spreadsheets and lacked integration with important underlying data needed to understand the client sales history and potential IT requirements. The sales professionals with whom we spoke were open to using a new tool, provided that such a tool

1. References a large universe of existing clients and potential new clients,
2. Incorporates relevant data that may require multiple existing tools to access,
3. Includes analytical models to help identify the best sales opportunities, and
4. Integrates all such data for each company under a single user interface designed by end users to facilitate easy navigation.

As discussed further in Section 6.1.1, OnTARGET is now used by 7,000 IBM sales representatives, and has largely replaced the previous set disparate sales tools and spreadsheets. The success of OnTARGET is due in large measure to our ability to deliver these key capabilities directly to the front-line sales force.

In the rest of this section, we discuss specific design decisions and implementations in light of these requirements. In particular, we discuss the types of data selected for inclusion in the tool, the integration of this data in the overall OnTARGET system, and the design of the user interface. The OnTARGET analytical models are described in the following section.

## 2.2 Architecture and Data

### *2.2.1 Design Objectives and System Overview*

As noted above, the broad business objective for OnTARGET is to provide sales professionals with relevant, actionable data and analytics under one user interface (UI). From a design perspective, this requirement drove decisions on the specific data and linkages to be incorporated, as well as the criteria to specify the universe of companies to be made available within the tool. After discussions with sales professionals, the following sources of data were selected for inclusion:

1. All transactions executed by IBM with its clients over the past 5 years
2. Dun & Bradstreet (D&B) firmographic data [1], e.g. company revenue, number of employees, corporate organizational hierarchy, etc.
3. Information on installed hardware and software at IBM client sites
4. Contact information for both customers and non-customers
5. Competitive information from external vendors
6. Assignments of companies to sales territories.

In terms of defining the universe of companies, it was required that OnTARGET include all significant IBM clients, as well as potential new customers drawn from the universe of companies available in D&B. Using the historical IBM transactional data, we select client companies for inclusion based on a minimum threshold of their spending with IBM over the past five years. Non-customer prospects are selected via minimum thresholds on company sales and number of employees, based on the D&B data.

Using these criteria, OnTARGET currently contains well over 1 million D&B company sites for the United States alone. Over 2 million sites are included worldwide.

From the beginning, OnTARGET was developed with a web-based front end that would be flexible enough to allow end users to execute complex queries directly from the user interface. OnTARGET is implemented as a Java application running on a Websphere application server, with DB2 as the relational database. It has a performance requirement of less than seven-second response time for all transactions executed.

Figure 1 provides a high-level overview of the OnTARGET system. The OnTARGET architecture can be viewed in terms of three key elements: the data store, the analytical models, and the user interface. The architecture somewhat isolates these elements to provide flexibility during the development and deployment process. It also allows the transformation and refresh of data to occur in a staging area, with subsequent deployment to the production database. These operations are quite resource intensive, so executing them outside of the production environment eliminates any impact to the production application. The analytical models are developed outside the OnTARGET system, and are imported onto the staging server and integrated with the other data sources. Separate cross-sell rules are specified by sales people, and are integrated in much the same way as the analytical models.

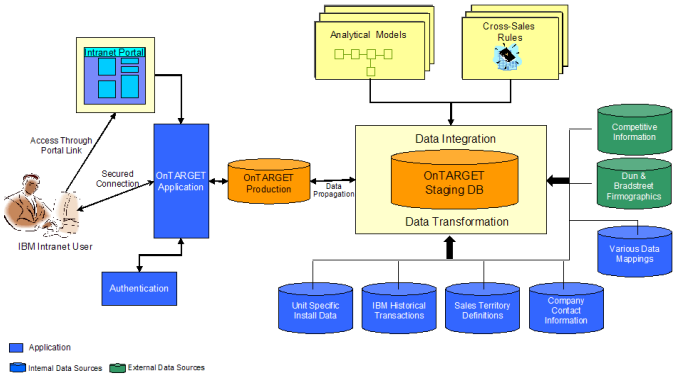



**Figure 1: Overview of the OnTARGET architecture and data.**

### *2.2.2 Data Model*

The principal design objective of the OnTARGET data model was to facilitate the support and maintenance of key data drawn from multiple data sources across all major geographic regions. Some of the data from each geographic region came from disparate data sources, so commonality of data elements had to be designed into the model. For example, the source contact entity from one region may have different fields and lengths from another or an element might have a common field name with different domains. An analysis of the domain and length of each data element was done to ensure that a common data model could be created to allow the user interface to work more efficiently, standardize queries, and have a standard code base worldwide. All relevant pieces of data from each of the required entities were gleaned and assembled in a Computer Aided Software Engineering (CASE) tool from which a logical and physical data model was designed. OnTARGET used IBM's Rational Data Architect [6] for its CASE tool.

OnTARGET was initially deployed to several countries in the Americas, followed by fifteen in the Europe, and three in Asia Pacific. Hence, another key requirement of the common data model was that it readily support integration of new countries as data became available. The standardization of data structures allowed the user interface to remain untouched in many instances even as additional countries were being added.

As noted above, OnTARGET utilizes both internal IBM data (e.g. transactional data) and external reference data (e.g. D&B firmographic information along with competitive data from multiple vendors). IBM uses an internal reference number to identify customers, so it was necessary to introduce a unique database key in order to join internal data with the external reference data for each company. An external reference number (D&B DUNS number [4]) was chosen as the main key primarily because OnTARGET also includes companies that are not currently IBM customers. We developed a flexible process to transform all data to this common key.

Transformation algorithms were developed using a transformation tool, WebSphere DataStage [7], to allow for consistent data presentation within the application. This helped to give the OnTARGET user interface a more consistent look and feel, regardless of the geography in which it was being used. This tool also helped in documenting the data flows within the application and was useful for ongoing maintenance and training.

Major updates to the OnTARGET data are made each quarter. During each update cycle, the historical IBM transactional data are refreshed, and updated populations of non-customers are extracted from the D&B tables. All models are rebuilt using this data, and hence the model scores are always consistent with the latest financial and firmographic data. Updates to the other information, including company contact information and product installation records, are made more frequently.

## 2.3 OnTARGET User Interface

The purpose of the OnTARGET user interface is to help sales personnel quickly identify the best potential revenue opportunities in their sales territory. Figure 2 shows a simplified, conceptual view of the OnTARGET user interface. The basic objective is to allow the user to build a focused customer targeting list composed of companies that meet criteria specified by the user. In the first step, the user defines a broad set of companies based on selections described in Figure 2. For example, one can specify a location (*e.g.* New York state), and an industry (*e.g.* Financial Services) and immediately form a set of companies meeting these criteria. Alternatively, a sales representative interested in a specific sales territory can select the territory identifier(s), and immediately build an initial set of all companies within these sales units.

The second step allows the user to further filter this initial set of companies based on additional criteria. For example, it is possible to filter the list based on upper and lower limits on company size as given by the D&B values for company sales and number of employees. It is possible to select only companies that have purchased in IBM product groups (*e.g. Lotus* software). Furthermore, using the OnTARGET propensity models, a user can select companies that have a high propensity to purchase in one of ten different IBM product groups (*e.g. Tivoli* software, *System x* servers ). The interface allows Boolean operators like AND, OR, and NOT to be applied in specifying the query. Hence, it is

possible, for example, to filter down to companies that have purchased a *Lotus* product, but have not purchased any *System x* product over the past five years. The various selection criteria mentioned here are easily entered in the interface via standard pull-down and selection menus.

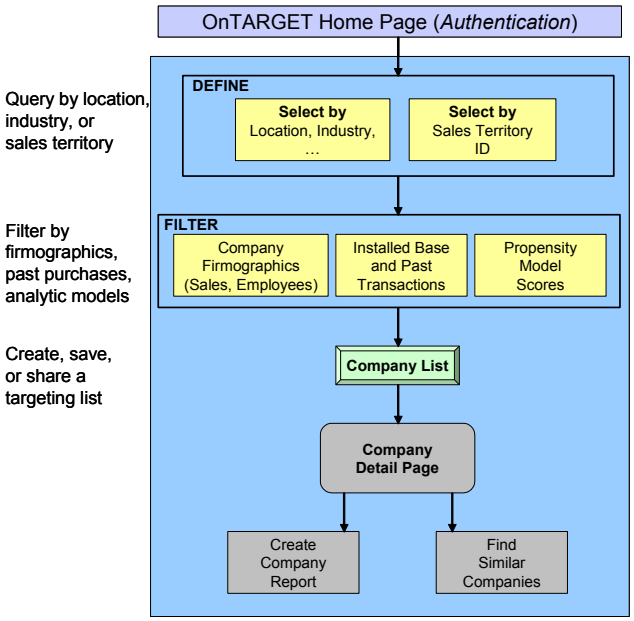



**Figure 2: Conceptual view of the OnTARGET user interface**

The result of the above define-and-filter process is the creation of a query that is executed against the OnTARGET database. All companies that meet the specified criteria are displayed as the resulting targeting list. This list can be further modified by adding and removing companies directly, or by modifying the filter criteria in an iterative process that yields a list of key potential opportunities as a focus for the sales process. Selecting any company in the list takes the user to the company's detail page.

The company detail page includes a comprehensive view of all the information in the OnTARGET database about the company, *e.g.* D&B firmographics, contact information, installed base, competitive information, and propensity scores. This holistic view facilitates the sales process by providing all relevant information in one place, allowing a user to easily generate a downloadable report of this information.

An interesting feature incorporated in OnTARGET is the capability to locate companies that are "similar" to a target company. Similarity is defined by a distance metric constructed using only firmographic information, *e.g.* companies in the same industry with comparable sales and numbers of employees. This feature is useful in further identifying sales prospects, as well as understanding which IBM products have been purchased by other companies of comparable size in the same industry.

The targeting list can be saved for future reference or as a basis for applying other criteria. OnTARGET also provides capability for sellers to collaborate by sharing the targeting lists with others. Users can receive targeting lists and then refine the filters to meet their specific requirements. In many cases, this function enables a sales operations person to define criteria and pass them on to representatives in their region.

An essential feature of the user interface is the enforcement of appropriate security and privacy rules to ensure that all information is protected according to IBM and country-specific policies. This capability is managed from a separate administration interface that allows specification of rules to limit display of sensitive data to users with the appropriate authorization.

OnTARGET also includes the capability to collect usage statistics such as the number of logins by each user, as well as time-stamped user accesses to each company detail page. These data are essential to quantify both the acceptance of the tool, as well as some indication of the extent to which subsequent revenue for a specific client can be linked to usage of the tool. We discuss these metrics in the section on business impact.

## 3. OnTARGET Propensity Models

As mentioned in the Introduction, OnTARGET and MAP employ different predictive models, based on their respective business objectives. For OnTARGET, we develop *propensity* models to predict the probability of purchase within a specific product group, while the MAP models are designed to estimate the potential *revenue opportunity* at each client account. The MAP models are described in a subsequent section.

The goal of the propensity models is to differentiate customers (or potential customers) by their likelihood of purchasing various IBM products. Rather than model at the level of individual products, our models are built to predict purchases within broad product groups or *brands*. Examples of these brands are *Lotus* or *Tivoli* (IBM software brands) and *System p* or *Storage* (IBM server brands). Currently, we develop separate propensity models for ten product brands.

We have at our disposal several major data sources to utilize in this task. The two major ones, which are available for the largest number of companies, are:

1. Historical IBM transactions for all IBM customers
2. Publicly available firmographic data from Dun & Bradstreet and other sources.

Our goal is to make use of this data to build propensity models which are (a) widely applicable, and consider all potential customers, and (b) accurate in terms of differentiating the high-propensity customers from low-propensity ones, on a product-by-product basis. For this purpose, for every product brand Y, we first divide the universe of OnTARGET companies into three distinct groups:

1. Companies that have already purchased Y in the past. These companies are eliminated from the propensity modeling all together.
2. Companies that have a relationship with IBM but have never purchased Y. For these companies we can utilize both data sources 1 and 2 above, in building our *existing customer model*.
3. Companies who have never purchased from IBM. For these companies we only have the firmographic

information (data source 2). The model for these companies is termed the *whitespace model.*

Since we have multiple geographies (Americas, Europe, Asia Pacific), with multiple countries within each geography, and multiple product brands, we end up building a large number of propensity models (currently about 160) in each quarter. In what follows, we summarize our modeling approach and the considerations leading to it, demonstrate its evaluation process during modeling, and show results of actual field testing. Finally, we discuss the modeling automation put in place to handle the overwhelming number of models built each quarter.

## 3.1 Propensity modeling methodology

We begin by specifying a geography, a brand Y and a modeling problem (Existing Customer or Whitespace). Our first step is to identify *positive examples* and *negative examples* to be used for modeling. In each modeling problem, we are trying to understand what drives the first purchase decision for brand Y, and delineate companies by the likelihood of their purchase. Assume the current time period (typically last year, or last two years) is *t*, then this leads us to the formulation of our modeling problem as:

> *Differentiate companies who never bought brand Y until period t, then bought it during period t, from companies who have never bought brand Y.*

Of the companies who never bought brand Y before period *t*, some will have bought other products before *t*. These companies form the basis of the existing customer model for Y. The companies who never bought any brand before *t* are the basis for the whitespace model. Thus, for the whitespace problem, our positive and negative examples are:

> *Positive:* companies who have never bought from IBM before *t*, then bought Y during *t*
>
> *Negative:* companies who have never bought from IBM before or during *t.*

The definitions for the existing customer problem are similar, except that a previous purchase from IBM is required for inclusion. For some combinations of geography, brand and modeling problem, the number of positives may be too small for effective modeling (we typically require at least 50 positive examples to obtain good models). In that case, we often choose to combine several similar modeling tasks (where similarity can be in terms of geography, brand, or both) into one "meta-model" with more positives. In [8], we discuss in detail the tradeoffs involved in this approach and demonstrate its effectiveness.

Next, we define the variables to be used in modeling. For existing customers, we derive multiple variables from historical IBM transactions, describing the history of IBM relationship before period *t*. Examples of these features are

- Total amount spent on software purchases in the two years before *t*
- Total amount spent on software purchases in the two years before *t*, compared to other IBM customers (rank within IBM customer population)
- Total amount spent on storage products purchases in the four years before *t.*

For both existing and whitespace customers, we derive variables from the D&B firmographic data, *e.g.*

- Company size indicators (revenue, employees), both in absolute and relative terms (rank within industry)
- Industry variables – both raw industry classification from D&B and derived sector variable
- Company's location in corporate hierarchy (Corporate HQ, Subsidiary, etc.).

We then build a classification model (more accurately, a probability estimation model), which attempts to use these variables to differentiate the *positive examples* from the *negative examples*. Our most commonly used modeling tool is *logistic regression,* although we have experimented with other approaches, like boosting. For each example, the model estimates the probability of belonging to the positive class. For presentation in the OnTARGET tool, these continuous scores are binned from 1 to 5, with bin distributions specified such that only 15% of existing customer examples receive the highest rating of 5. For the whitespace model, only 5% get a rating of 5, reflecting the observation that it is generally more difficult to sell into a non-customer account.

## 3.2 Example of modeling results

The resulting logistic regression models can be examined and, to some extent, interpreted as "scorecards", describing the effect of different variables on the likelihood of converting a company into a customer for brand Y. We describe here a detailed example from a recent round of Existing Customer models built for North America, and discuss possible interpretations.

The example we give is the existing customer model for the *Rational* software brand. Figure 3 shows the predictive relationships found for this model.

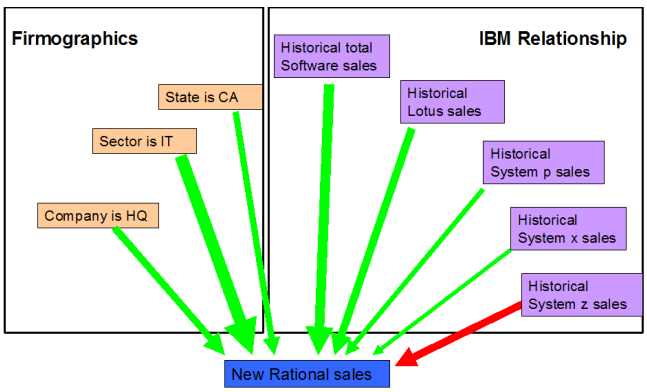



**Figure 3: Predictive relationships between some derived variables and new *Rational* sales to existing customers**

Green arrows signal a positive effect (i.e., increase in propensity from increase in variable's value) while red arrows signal an adverse effect on propensity. The width of the arrows indicates the strength of the effect, as measured by the magnitude of the regression coefficient. We show only statistically significant (measured by p-value) effects in the figure. We see several interesting effects, and most seem to be explainable:

- Industrial sector (IT), geography (California) and company's corporate status (Headquarters) seem to have

a strong predictive effect. This seems consistent with Rational being an advanced software development platform, which medium-sized IT companies in California (and thus likely at the front of the hi-tech industry) might be interested in purchasing.

- The size of total prior software (SWG) relationship with IBM seems to be a strong indicator of propensity to buy. On top of that, having a strong relationship in Lotus seems to afford additional power.
- While the total size of prior non-software relationship does not have a strong effect, some specific non-software brands seem to be important. *System p* (and *System x*, somewhat) seem to encourage Rational sales, while *System z* relationship seems to discourage them. While this last fact may seem puzzling, it may be explainable by the particular nature of the software relationship with *System z* customers, who often manage their software relationship with IBM in conjunction with the *System z* relationship. More analysis would be required to clarify this point.

## 3.3 Evaluating model performance

Statistical model evaluation metrics typically include significance of coefficients, likelihood, and percentage of deviance explained. In all the hundreds of models we build, some of the variables are always highly significant (see Fig. 3 for an example). However, we do not perform further variable selection to save computation and avoid overfitting. In addition, we are not particularly interested in analyzing the "absolute" performance of our models, but care much more about the relative performance to a baseline model along the marketing-relevant performance metrics of Lift and AUC as defined below. We use for our comparisons a *10-fold cross-validation* approach, whereby we divide our data (positive plus negative examples) to 10 equal-sized bins and build 10 models, each time using $9/10^{th}$ of the data (9 of 10 bins) for modeling, and then applying the resulting model to the leave-out $10^{th}$ bin. After repeating this 10 times, we have our whole data scored as "leave-out" by the different models, and we can use it to evaluate the modeling success. For a detailed description of cross validation, see [9].

We then evaluate our model by the *Lift* performance on the holdout data. Lift of the model at percentile $x$ is defined as the ratio of the fraction of positives in the top x% of the data sorted by its scores to the expected fraction, which is of course x itself:

$$\text{Lift}(x) = \frac{\text{fraction of total positive examples in top x\%}}{x}.$$

The lift is a natural measure in the marketing context because it measures how much more successful our model is than a model that simply assigns random scores. The lift also is quite robust to the ratio of positive to negative examples used, which is important since our learning samples are typically biased in favor of positive examples, compared to the full population. (For discussion of these biases and their effect on evaluation, see [10]).

To seriously evaluate a model's success, we should always judge it against a reasonable "baseline" model which a knowledgeable sales person might employ, rather than against the random model. For this we adopt a baseline model that ranks prospects by a measure of company size. We refer to this as the "Willy Sutton" model. (Willy Sutton was the infamous bank robber, who reportedly said that he robbed banks because "That's where the money is".) For existing customers, we thus rank them by the size of their relationship with IBM (largest to smallest, based on total revenue with IBM), implicitly assuming the largest customers are most likely to buy a brand in which they have no current relationship. For whitespace companies, we rank them by their company size (revenue and/or employees) as reported by D&B.

Our cross-validated evaluation indicates that our models almost invariably do significantly better than Willy Sutton. For the most recent Existing Customer problems, our models do significantly better than Willy Sutton on 9 out of 10 problems, with the sole exception being *System z*, for which the performance is only slightly better. Indeed, this may not be surprising, as this is the brand where size of IBM relationship is indeed most likely to be critical for new purchases. A common graphical display, representing the lift of a model at all values of $x$, is the *lift curve*. Figure 4 shows an example lift curve for the whitespace model built for the Rational software brand. The lift at 5% and 10% is calculated explicitly, and the model is compared to the random model and Willy Sutton model. Note that the OnTARGET model significantly out-performs both baseline methods.

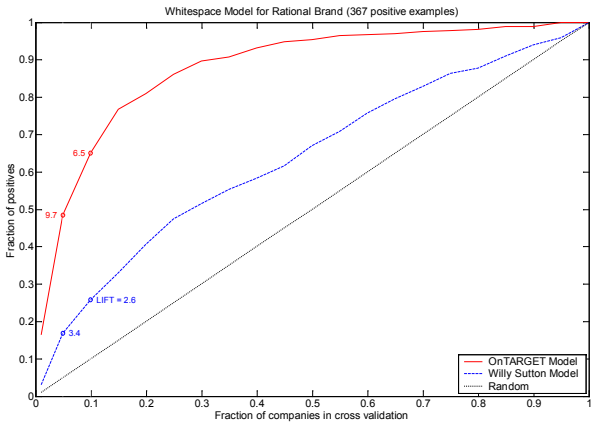



**Figure 4: Example lift curve, comparing OnTARGET model performance to some baseline models**

A more interesting evaluation, however, is to judge the models by their actual success in predicting new sales. We have been able to do this, by considering new sales recorded in 4Q 2006, and investigating the scores which our previously built models in 3Q 2006 assigned to these sales, compared to the whole population of non-customers. These sales were *not* visible in the data at the time these models were built; however they were most likely initiated before the models' results were available, and thus not affected by these results. Hence we are getting a clean evaluation of the models' success in identifying actual sales as high propensity opportunities.

Table 1 shows the evaluation results for the 10 existing customer models. We compare the performance of our model to that of the Willy-Sutton model in terms of AUC (area under the ROC curve, or equivalently, area under the lift curve) and lift in the top 5%. The higher number (better performance) for each modeling problem is presented in bold.

We can observe that our models do better than the baseline model in terms of AUC (which is between 0 and 1 and roughly measures a model's success in ranking all the data successfully) for 9 of the 10 modeling problems. (The tenth, *System z*, has very few positives, and the models end up in a tie.) In terms of lift at 5%, we observe that both modeling approaches do very well (often finding as many as 50% of sales in this top portion), but our models generally outperform the baseline models, sometimes significantly. Overall our models clearly do a better job of identifying sales opportunities, but this advantage is less pronounced at the top of the ranked lists (top 5%), where the Willy-Sutton policy of just taking the biggest customers seems to be a reasonable approximation of the best model we can build.

**Table 1: Model results based on predicting new sales**

| Product Brand | Number positives | Model AUC | WS AUC | Model lift 5% | WS lift 5% |
|---|---|---|---|---|---|
| Information Management | 26 | **0.82** | 0.69 | **9.23** | 6.15 |
| Lotus | 25 | **0.81** | 0.72 | 7.20 | **8.80** |
| Rational | 42 | **0.91** | 0.79 | **12.38** | 10.00 |
| Tivoli | 56 | **0.82** | 0.77 | 8.21 | **8.57** |
| Websphere | 37 | **0.90** | 0.80 | 12.43 | 12.43 |
| System i | 69 | **0.77** | 0.73 | **5.22** | 2.90 |
| System p | 74 | **0.83** | 0.78 | **8.92** | 8.38 |
| System x | 27 | **0.80** | 0.74 | 7.41 | **9.63** |
| System z | 6 | 0.78 | 0.78 | 6.67 | 6.67 |
| Storage | 194 | **0.88** | 0.80 | **10.31** | 6.91 |

## 3.4 Model automation

As we described above, the broad geographical usage of OnTARGET, plus our careful modeling methodology, necessitates the generation of at least 160 new models in each quarter. It is therefore critical to have a reliable, repeatable, and automated modeling methodology implemented. After several initial modeling iterations when all modeling was done manually, to create a deep understanding of the problems and the variables involved, we have created such a system, whose main characteristic is having a large collection of possible predictive variables, and selecting some part of it for every prediction model. The main considerations in choosing variables for each model are:

- The number of positive examples (if there are too few examples, we cannot use too many variables)
- Our experience with the specific modeling problem (some variables are more important than others in specific geographies and/or for specific brands)
- Data availability (some variables are not available in specific geographies).

Taking all of these into account, here is a schematic description of the automated modeling system we have devised:

1. Determine a definition of positive examples (including which length of time period *t* will be used to define the "new" customers), and draw an appropriate negative sample.
2. Based on the number of positive examples, determine how many "degrees of freedom" the model should use.
3. Go down a ranked list of variables (which were ranked a-priori based on domain knowledge) and choose the maximal number of variables so that the degrees of freedom quota is not exceeded.
4. Build a prediction model using this set of variables, and apply the cross validation model evaluation methodology to it.
5. Create a file of results, which the user views to verify that the results are good and that they "make sense", i.e., do not represent data leakage or other unexpected problems.

This almost-fully automated system for model generation has been in place for the two most recent modeling iterations (two quarters). The main manual intervention in this process is currently in examining the final output and the evaluation of the models, and thereby making sure that changes in data, bugs, or other unexpected phenomena have not affected the predictive performance adversely.

# 4. MAP: A Sales Force Allocation Tool

In this section, we discuss the second initiative, the Market Alignment Program (MAP), mentioned in the Introduction. We first describe the project objectives, and then delineate the business process we developed to address these objectives. An integral part of the MAP process is the validation of analytical estimates via an extensive set of workshops conducted with sales leaders. These interviews rely heavily on a web-based tool to convey the relevant information, as well as to capture the expert feedback on the analytical models. We describe the design of the MAP tool infrastructure, its data model, and the key characteristics of the user interface.

## 4.1 The MAP Business Objective

A major challenge faced by many sales organizations is poor alignment of the sales force with market opportunity. The objective of the MAP initiative is to address this challenge by focusing on the three main problem areas in the sales force deployment methodology:

1. **Lack of a uniform disciplined approach to estimating revenue opportunity at a customer level**. This problem leads to alignment of sales resources with past revenue rather than future revenue opportunity. MAP links sales force allocation process to field-validated analytical estimates of future revenue opportunity in each operational market segment
2. **Lack of front-line input into the planning process.** Purely top-down planning process – although easier to manage – does not typically result in an optimal allocation of sales resources and revenue targets. It also often results in disenfranchised sales teams. The MAP business process explicitly requires detailed input from the front-line sales teams.

3. **Lack of easily accessible common fact base and analytical methodology for making resource shift decisions**. This problem limited the scope and impact of the previous deployment optimization efforts because sales leaders in different parts of an organization found it difficult to arrive at rational fact-based trade-off decisions in the absence of a common fact base. MAP solves this problem by delivering the properly aggregated information to decision makers through a web-based tool.

## 4.2 The MAP Business Process

The MAP business process can be broken into four main steps:

1. **Prepare input for front-line sales workshops**. This step includes populating the web-based tool with data on past revenue, model-estimated revenue opportunity, and deployment of sales resources.
2. **Conduct front-line workshops**. This is the most time-consuming phase of the planning process. In this phase we validate the model-estimated future revenue opportunity per customer by product division, validate current coverage, and capture future resource requirements. All workshops are conducted using the MAP tool.
3. **Conduct workshops with regional sales leaders within each product division and each industry sector.** In this phase we prioritize customers within each product division and industry sector and validate coverage requirements.
4. **Conduct regional "summits" with the sales leaders from all product divisions and industry sector teams.** In this phase we develop overall IBM sales coverage and strategy for priority customers.

## 4.3 System Overview

Figure 5 shows a high-level view of the MAP system. As noted above, the MAP web-based tool is used to conduct extensive interviews with IBM sales teams, and this process has motivated several design features that differ from OnTARGET. In addition to many of the data sources used in OnTARGET (see Figure 1), MAP also includes data on assignments of current sales resources. The MAP revenue-opportunity models are built for each major IBM product brand using the combined transactional and D&B data. Revenue opportunities are estimated for each account, and these results are stored in the database for display during the interview process. Unlike OnTARGET, the MAP tool must capture specific feedback on the revenue-opportunity models presented to the sales teams during the interview process. The tool allows the sales team to input their estimates of revenue opportunity, as well as their reasons for recommending a change to the model results. These "validated" opportunities are stored in the MAP database, and then post-processed using separate tools after the interviews have been completed.

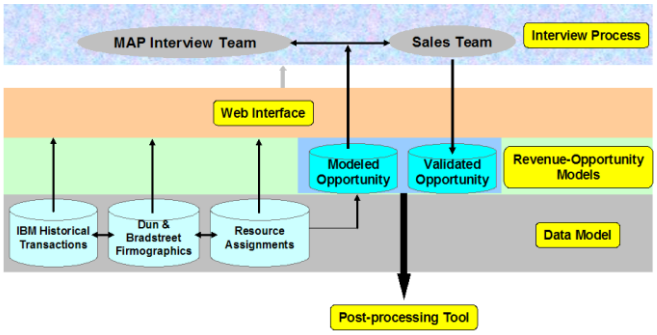



**Figure 5: Overview of the MAP system**

## 4.4 MAP Data Model and User Interface

Figure 6 shows the overall flow of the MAP user interface. An essential requirement in the interface design is that participants in the MAP interview process be able to locate their accounts and sales territories within the tool.[2] For this reason, the query interface shown in Figure 6 returns either a list of accounts or a list of sales territories that satisfy query conditions such as geography and industry sector. In contrast, the analogous queries in OnTARGET (see Figure 2) return lists of companies that are indexed by the D&B DUNS number. Hence, the MAP data model is organized about IBM client accounts, while the OnTARGET data model utilizes the D&B representation of companies. The difference in design is motivated by the different business objectives of the two systems: OnTARGET must provide easy identification of non-customer (whitespace) companies, while MAP must support an IBM-centric view in order to be consistent with the account- and territory-based views of the participants in the MAP interview process.

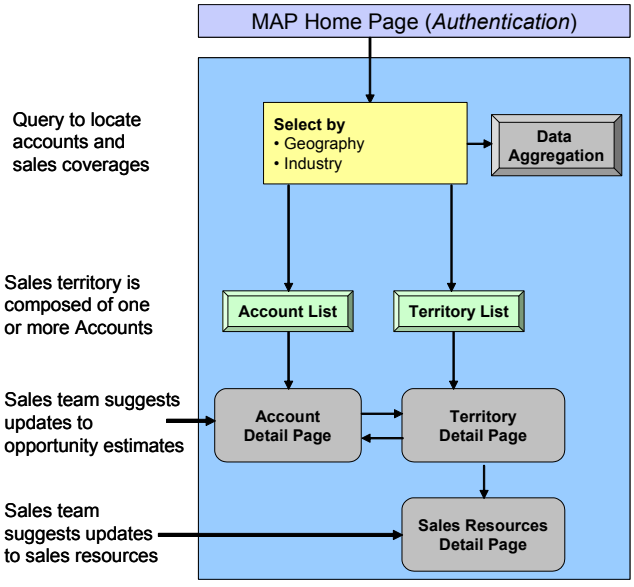



**Figure 6: Conceptual view of the MAP user interface**

[2] In this discussion, we assume that a sales territory is composed of one or more client accounts.

Returning to Figure 6, the generated lists of accounts and territories contain links to pages that summarize all relevant information for each account or sales territory. For example, the account detail page shows a 5-year summary of all IBM revenue for each major IBM product brand and the estimated revenue opportunity for each brand. Feedback on the estimated revenue opportunities is collected via user inputs on this page. The sales team can also enter changes to existing numbers of sales resources that might be required to achieve future revenue levels; note these data are entered at the sales territory level (on the territory detail page) since sales resources are allocated at this level. To facilitate regional and industry-based workshops, the MAP tool also provides a capability to aggregate data such as revenue opportunity within market segments, *e.g.* within a specific geography or industry.

# 5. MAP Revenue-Opportunity Models

The total amount of money a customer can spend on a certain product category is a vital piece of information for planning and managing sales and marketing efforts. This amount is usually referred to as the customer's wallet or *revenue opportunity* for this product category[3]. For the MAP workshops, we needed an unbiased, realistic estimate of the true revenue opportunity at each account for the purpose of driving an informed discussion with each sales team. In this section, we introduce several definitions of revenue opportunity, and then describe several modeling approaches. We describe our approach to model evaluation in which we compare each candidate model with sales-team feedback collected during the initial set of MAP workshops. Although we develop these ideas here in the context of the IBM revenue opportunity with its clients, the methodology is applicable to any company with a large volume of historical transaction data.

## 5.1 Definition of Opportunity

The first question we need to address is what exactly is meant by a customer's wallet or opportunity? We discuss this in the context of IBM as a seller of information technology (IT) products to a large collection of customers for whom we wish to estimate the wallet. We have considered three nested definitions:

1. The total spending by this customer in a particular group of IT products or services. This is simply the total IT spending (by product group) by the customer. We denote this as *TOTAL Opportunity*.
2. The total attainable (or served) opportunity for the customer. In IBM's case, this would correspond to the total spend by the customer in IT areas covered by IBM's products and services. While IBM serves all areas of IT spending (software, hardware and services), its products do not necessarily cover all needs of companies in each of these areas. Thus, the *SERVED Opportunity* is smaller than the total IT spending.
3. The "realistically attainable" opportunity, as defined by what the "best" customers spend. This will be different from SERVED Opportunity, because it is not realistic to expect individual customers spend their entire budget with IBM. We refer to this as *REALISTIC Opportunity*. This is also the definition that we use to define Revenue Opportunity for MAP.

From sources like D&B, the total company revenue (annual sales) is readily available for all companies. We also know the total amount of historical sales (IBM SALES) made by IBM to its customers. In principle, the relation

**IBM SALES < REALISTIC < SERVED < TOTAL < COMPANY REVENUE**

should hold for every company. Note that we expect IBM Sales to approach REALISTIC Opportunity for those companies where IBM is the dominant IT provider.

As noted above, MAP uses the REALISTIC definition, because it is most consistent with IBM sales executives' notion of opportunity. Defining the best customers is essential in estimating REALISTIC opportunity. In what follows, we define best customers in a relative sense, as ones who are spending with IBM as much as we could hope them to, given a stochastic model of spending. Thus, a good customer is one whose spending with IBM is at a high percentile of its "spending distribution". We describe below some approaches that allow us to build models that predict such a high percentile and, perhaps more importantly, allow us to evaluate models with regard to the goal of predicting percentiles of individual spending distributions.

## 5.2 Modeling REALISTIC Opportunity as Quantiles

Under the REALISTIC definition, we are looking for a high percentile (e.g. 80%) of the *conditional* spending distribution of the customer, given all the information we have about this customer. To make this notion more concrete, let us consider a customer X and start by imagining that we have not just one customer but (say) 1,000,000 identical customers exactly like X, except that each customer makes its decision about how much they spend with IBM independently. Then we could just take the $80^{th}$ percentile of this spending distribution (i.e., the quantity O such that 80% of these 1,000,000 identical customers spend $O or less with IBM) as our REALISTIC wallet estimate for X and its 1,000,000 identical twins. In practice we do not, of course, observe multiple copies of each company, and so our challenge is to get a good estimate of this conditional spending percentile for each company from the data we have. In general, the approaches for doing so can be divided into two families:

1. *Local approaches*, which try to take the idea we described above (of having 1,000,000 copies of X) and approximate it by finding companies that are "similar" to X, and estimating X's REALISTIC wallet as the $80^{th}$ percentile of the IBM sales of this "neighborhood".
2. *Global models*, which attempt to describe the $80^{th}$ percentile as a function of all the information we have about our customers. The simpler and most commonly used approach is quantile regression [11], which

[3] We will often use the common marketing term "wallet" interchangeably with "revenue opportunity".

directly models the quantile (or percentile) of a response variable Y (in our case, the IBM spending) as a function of predictors X (in our case, the firmographics from D&B and the IBM historical transaction data).

The standard regression approach estimates the conditional expected value E(y|x) by minimizing sum of squared error $(y - \hat{y})^2$. In the case of quantile regression we have to find a model that minimizes a piecewise linear, asymmetric loss function known as quantile loss

$$L_p(y, \hat{y}) = \begin{cases} p \times (y - \hat{y}) & \text{if } y > \hat{y} \\ (1 - p) \times (\hat{y} - y) & \text{if } \hat{y} > y \end{cases}$$

where p is the particular percentile. This loss function is appropriate for quantile modeling because the loss is minimized in expectation when the desired quantile is being perfectly modeled (for further discussion, see [11]).

## 5.3 Quantile Estimation Techniques

Within the scope of the MAP project, we explored a number of existing approaches for quantile estimation and also developed some novel modeling techniques. We evaluated the different models both in a traditional predictive modeling framework on holdout data, as well as against the expert feedback that was collected in the initial round of MAP sales-team workshops. We will discuss the most relevant approaches to the MAP project in more detail below. But we also note that our research efforts have led to two additional techniques for wallet estimation:

1. *Quanting*, which uses an ensemble of classification models to estimate the conditional quantiles, as described in [12].
2. *Graphical decomposition* models, which assume that the IBM revenue is determined by two independent drivers, 1) the company relationship with IBM and 2) the company IT budget (the opportunity) which itself is determined independently of the IBM relationship by the company's IT needs. This approach is detailed in [13].

Similar to the propensity models from the previous section, we estimate revenue opportunity for each company at the major product-brand level using the firmographics from D&B and the IBM historical transaction data.

### *5.3.1 k-Nearest Neighbor*

The revenue-opportunity estimates provided in the first release of the MAP tool used a k-nearest neighbor (kNN) [16] approach that follows very closely the definition of REALISTIC Opportunity.

The traditional k-nearest neighbor model (kNN) is defined as

$$\hat{y}(x) = 1/k \sum_{x_i \in N_k(x)} y$$

where $N_k(x)$ is the neighborhood of $x$ defined by the $k$ closes points $x_i$ in the training sample for a given distance measure (e.g., Eucledian). From a statistical perspective we can view the set $y_j \in N_k(x)$ as a sample from the approximated conditional distribution of $P(Y|x)$. The standard kNN estimator of $\hat{y}$ is simply the expected value of this conditional distribution approximated by a local neighborhood. For quantile estimation we are not interested in the expected value (i.e., an estimate of $E(Y|x)$) but rather a particular quantile $c_p(x)$ of the conditional distribution $P(Y|x)$ such that $P(Y>c_p(x)|x)=q$. Accordingly we can estimate $c_p(x)$ in a k-nearest neighbor setting as the $q^{th}$ quantile of the empirical distribution of $\{y_j : x_j \in N_k(x)\}$. If we denote that empirical distribution by:

$$\hat{G}_x(c) = 1/k \sum_{x_j \in N_k(x)} 1\{y_j \le c\}$$

then our kNN estimate of the $q^{th}$ quantile of $P(Y|x)$ would be $\hat{G}_x^{-1}(q)$.

The interpretation is similarly that the values of $Y$ in the neighborhood $N_k(x)$ are a sample from the conditional distribution $P(Y|x)$ and we are empirically estimating its $q^{th}$ quantile. An important practical aspect of this estimate is that, in contrast to the standard kNN estimates, it imposes a constraint on $k$. While $k$=1 produces an unbiased (while high variance) estimate of the expected value, the choice of $k$ has to be at least $1/(1-q)$ to provide an upper bound for the estimate of the $q^{th}$ 'high' quantile (more generally we have $k \ge \max(1/q, 1/(1-q))$).

The definition of neighborhood is determined based on the set of variables, the distance function and implicit properties such as scaling of the variables. The performance of a kNN model is very much subject to the suitability of the neighborhood definition to provide a good approximation of the true conditional distribution - this is true for the standard problem of estimating the conditional mean and no less so for estimating conditional quantiles.

For our particular wallet estimation problem, we find for each company a set of 20 similar companies, where similarity is based on the industry and a measure of size (either revenue or employees, depending on the availability of the distribution). From this set of 20 firms, we discard all companies with zero IBM revenue in the particular pillar and report the median of the IBM pillar revenues of the remaining companies. The choice of the median ($50^{th}$ percentile) reflects considerations of both the statistical robustness as well as the total market opportunity (sum over all companies) relative to the total IBM revenue.

### *5.3.2 Linear quantile regression*

A standard technique to estimate the REALISTIC wallet as percentiles of a conditional distribution is linear quantile regression [11]. Similar to standard linear regression models, quantile regression models aim to find a coefficient vector β such that Xβ is close to Y. The main difference between traditional linear regression and quantile linear regression is the loss function. While linear regression models the conditional expected value by minimizing the sum of squared error, quantile regression minimizes quantile loss as defined earlier. Figure 7 shows conceptually the difference between the linear regression line in black and the quantile regression line for the $90^{th}$ percentile in red.


**Figure7: Comparison of quantile regression and linear regression**

### *5.3.3 Quantile Regression Tree*

Tree-induction algorithms are very popular in predictive modeling and are known for their simplicity and efficiency when dealing with domains with large number of variables and cases. Regression trees are obtained using a fast divide and conquer greedy algorithm that recursively partitions the training data into subsets. Therefore, the definition of the neighborhood that is used to approximate the conditional distribution is not predetermined as in the case of the kNN model but optimized locally by the choice of the subsets. Work on tree-based regression models traces back to Morgan and Sonquist, but the major reference is the book on classification and regression trees (CART) by Breiman [14]. We will limit our discussion to this particular algorithm.

A tree-based modeling approach is determined predominantly by three components:

1. the **splitting criterion** which is used to select the next split in the recursive partitioning,
2. the **pruning method** that shrinks the overly large tree to an optimal size after the partitioning has finished in order to reduce variance,
3. the **estimation method** that determines the prediction within a given leaf.

The most common choice for the splitting criterion is the least squares error (LSE). While this criterion is consistent with the objective of finding the conditional expectation, it can also be interpreted as a measure of the improvement of the approximation quality of the conditional distribution estimate. Tree induction searches for local neighborhood definitions that provide good approximations for the true conditional distribution *P(Y|x)*. So an alternative interpretation of the LSE splitting criterion is to understand it as a measure of dependency between $Y$ and an $x_i$ variable by evaluating the decrease of uncertainty (as measured by variance) through conditioning. In addition, the use of LSE leads to implementations with high computational efficiency based on incremental estimates of the errors for all possible splits.

Pruning is the most common strategy to avoid overfitting within tree-based models. The objective is to obtain a smaller sub-tree of the initial overly large tree, excluding those lower level branches that are unreliable. CART uses Error-Complexity pruning approach which finds a optimal sequence of pruned trees by sequentially eliminating the subtree (i.e., node and all its ancestors) that minimizes the increase in error weighted by the number of leaves in the eliminated subtree:

$$g(t,T_t) = \frac{Err(t) - Err(T_t)}{S(T_t) - 1}$$

where $Err(T_t)$ is the error of the subtree $T_t$ containing t and all its ancestors, and $Err(t)$ is the error if it was replaced by a single leaf, and $S(T_t)$ is the number of leaves in the subtree. $Err()$ is measured in terms of the splitting criterion (i.e., for standard CART it is squared error loss). Given an optimal pruning sequence, one still needs to determine the optimal level of pruning and Breiman [14] suggest cross validation on a holdout set.

Finally CART estimates the prediction for a new case that falls into leaf node $l$ similarly to the kNN algorithm as the mean over the set of training responses $D_l$ in the leaf:

$$\hat{y}_l(x) = \frac{1}{n_l} \sum_{y_j \in D_l} y_i$$

where $n_l$ is the cardinality of the set $D_l$ of training cases in the leaf. Given our objective of quantile estimation, the most obvious adjustment to CART is to replace the sample mean estimate in the leaves with the quantile estimate using the empirical local estimate $\hat{G}_{D_l}(c)$ of *P(Y|x)* as in equation (3).

### *5.3.4 Post-processing*

Since the REALISTIC opportunity is defined as a high quantile of the conditional distribution, the predicted opportunity will be smaller than the realized IBM revenue for some companies. In particular, for a quantile of 90% we would expect that about 10% of companies to generate IBM revenue that is larger than our opportunity forecast. While we do not know the exact IBM revenue for the next year, we use the last year revenue as a proxy and report in the MAP tool the maximum of the opportunity model and last years revenue in the brand.

## 5.4 Evaluation of Opportunity Models

In deciding on which method to implement in the MAP tool, we needed to first identify the most appropriate quantile for the opportunity estimation, and then choose an appropriate evaluation criterion for model comparison. The main challenge in these objectives is the fact that we never directly observe the true REALISTIC IBM opportunity, and hence we need a reference solution. So rather than using potentially unreliable survey data, we decided to utilize the expert feedback collected in the initial round of sales workshops conducted in 2005. In particular, we built various opportunity models using D&B firmographics and IBM revenues for the ~30K companies that compose 6000 major MAP accounts discussed in the 2005 interviews.

Before describing these results, we discuss the expert feedback obtained during the initial (2005) MAP workshops. Recall that the experts could either accept the model estimate, or revise this estimate in any way. Figure 8 shows the "validated" (expert-specified) opportunity for a major IBM software brand for about 1200 accounts, as a function of the original opportunity estimates

that were provided in the MAP tool using the initial nearest neighbor model.

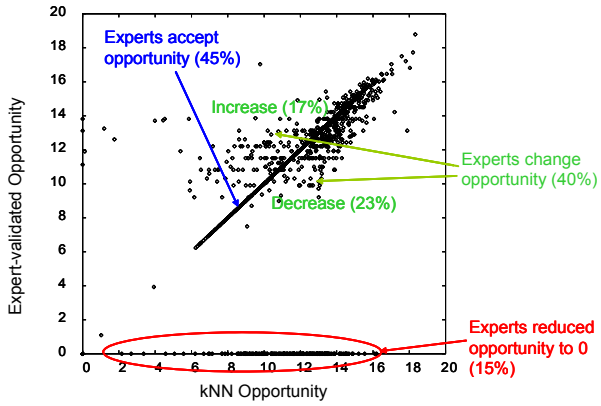



**Figure 8: Comparison of Predicted and Expert-Validated Opportunity**

The plot supports a number of interesting observations:

1. 45% of the opportunity estimates are accepted without alteration. The majority of the accepted opportunities are for smaller accounts. This shows a strong human bias towards accepting the provided numbers (this is broadly known as anchoring).
2. For 15% of the accounts, the experts concluded that there was NO opportunity – mostly for competitive reasons that cannot be known to our revenue-opportunity model. For that reason we decided to exclude those accounts from the model evaluation.
3. Of the remaining 40% of accounts, opportunity estimate were decreased (23%) slightly more often than they were increased (17% )
4. The horizontal lines reflect the human preference towards round numbers.
5. The opportunities and the feedback appear almost jointly normal in a log plot. This suggests that the opportunities have an exponential distribution with potentially large outliers, and that the sales experts corrected the opportunities in terms of percentage.

Given the high skew of the distribution, residual-based evaluation is not robust [17]. To account for this fact, we evaluated model performance on three scales: original, square root, and log. In addition to the sum of squared errors for each scale, we also considered the absolute error. This provides us with a total of 6 different performance criteria. For this analysis, we built nearly 100 different models, including multiple variants of each of the discussed quantile estimation approaches (linear quantile regression, quantile regression trees, k-nearest neighbors, Quanting and graphical models).. We then ranked all models according to each of the 6 performance criteria, and compared how often a given model appears within the top 10 of all models. Based on this analysis for three major product brands, we concluded that the linear quantile regression model showed the most consistently good performance for a quantile of 80%. In other words, this quantile regression model provided the best agreement with the expert feedback collected during the initial 2005 MAP workshops. Hence, this model was selected to provide the revenue opportunity estimates for MAP workshops conducted in 2006.

Selecting one model out of a large set of potential candidates based on the performance on a limited test invites potential of 'overfitting' in the sense of selecting a model that looks particularly good on the particular test set. While we are unable to correctly assess the significance of the model performance due to multiple comparison problem, we are confident with our choice for three reasons: 1) quantile regression is a low-variance model compared to the alternative models, 2) it performed consistently well across a range of parameter setting and 3) it performed consistently well across multiple test sets for the different product brands.

# 6. Solution Deployment and Business Impact

An essential component of initiatives such as OnTARGET and MAP is that we be able to quantify the impact of the delivered solution against the overall business objectives. In general, it is challenging to isolate the impact of a given tool or process, when it is injected into a broad, complex, and dynamic business environment. In this section, we describe several measures of business impact for each of these solutions.

## 6.1 OnTARGET Business Impact

### *6.1.1 Adoption*

The user population for OnTARGET has grown steadily over the last two years, from about 1,000 users at the end of 2005 to approximately 7,000 worldwide users by the close of 2006. These sales professionals are in 21 countries across three major geographies. Interest in the application continues to increase and the growth of the user community is anticipated to continue as new countries and sales personnel are added. As noted earlier, sales reps are often reluctant to use tools that do not enhance their productivity, so the adoption rate of OnTARGET is a significant measure of its impact, especially since use of the tool is not mandated.

While the initial deployment was for the IBM Software Group, the fact that the database includes information from across most business areas has made the application useful as an enterprise-wide application covering multiple lines of business. Therefore, the application quickly moved to other divisions and has become an enterprise application. It is currently being used by sales professionals from the IBM Software, Systems Technology (Server), and Services organizations.

### *6.1.2 Productivity Gains*

The main OnTARGET users are face-to-face and call-center sales personnel looking for the best potential opportunities in their space. During a recent survey of our user base, the average productivity gain identified was 2 hours per week. This productivity gain can be attributed to the fact that the user can quickly create focused targeting lists and does not have to use multiple tools to access additional data and research on prospective clients. OnTARGET users accessed and downloaded over 235,000 company-detail reports in 2006.

### *6.1.3 Impact of Propensity Models*

An obvious question concerns the degree to which the propensity models provide quantifiable business impact. As described in

detail in Section 3.3, we have examined the scores assigned to closed opportunities (i.e. a sale within a product brand) by models built in the quarter *before* the sale was recorded. This analysis, summarized in Table 1, provides strong evidence that the models are indeed identifying new sales opportunities that lead to closed revenue.

### *6.1.4 Pipeline Influence*

OnTARGET was designed to help identify the best potential opportunities for sales people, and hence we have focused on measuring the impact on the sales opportunity pipeline[4]. We define an opportunity to have been influenced by OnTARGET if the detail page for the prospective company has been viewed[5] within a three-month window prior to the opportunity creation date. By this metric, OnTARGET influenced over 17% of won opportunities in 2006 across the included lines of business world-wide. These OnTARGET-influenced opportunities represent over 23% of the pipeline dollar volume in 2006. Based on the most recent data available (through May, 2007), this fraction has increased to 29% in 2007. Given the size of IBM's Software and Systems Technology (Server) businesses, this fraction represents a very significant amount of revenue ultimately influenced by use of OnTARGET.

It is interesting to note that the average revenue associated with won opportunities influenced by OnTARGET has been larger. For example, in our software business, the average revenue associated with won opportunities influenced by OnTARGET in 2006 is over 45% larger than those not influenced by the tool.

### *6.1.5 Return on Investment*

It is difficult to accurately estimate the return on investment (ROI) for OnTARGET because we cannot isolate the gross profit associated with use of the tool, although, as discussed above, it is possible to characterize the impact on the sales pipeline. In terms of the investment, the initial OnTARGET prototype for the US and Canada was delivered in less than 1 year by approximately 10 people (a combination of skills in machine-learning, database development, web-site development, and business development). The subsequent investment to extend and maintain a full Enterprise version in multiple geographies has been approximately 40 people in each of 2006 and 2007. Based on this cumulative investment to date, and estimates of pre-tax income associated with the pipeline influence discussed in Section 6.1.4, the ROI can be argued to be at least a factor of five.

## 6.2 MAP Business Impact

### *6.2.1 Deployment and Adoption*

The MAP initiative has been widely adopted throughout IBM, and continues to play a key role in the deployment of IBM's sales resources. Since its initial deployment in the US in 2005, MAP has been rolled out to 32 countries around the world, which constitute more than 95% of IBM's total revenue. During the 2006 deployment, approximately 420 MAP workshops were conducted with sales teams globally, involving nearly 3,000 sellers across all of IBM's sales units. More than 2,200 individual accounts were discussed, representing approximately 55% of the total modeled revenue opportunity.

Following the interview process, client accounts for each business sector are classified within a two-dimensional segmentation defined by IBM revenue and validated revenue opportunity. With respect to the business objective of improving resource allocation, the most relevant segment is the "*Invest*" accounts where the validated revenue opportunity is significantly greater than current IBM revenue. Using the MAP segmentation, specific measurable decisions were made to optimize coverage of accounts. As a result of the 2005 deployment, a total of 380 sellers were reassigned to Invest accounts, with coverage of approximately 50 lower-opportunity accounts shifted to the ibm.com coverage channel.

The MAP prioritization framework has been adopted globally by all of IBM's business units, resulting in a common, cross-IBM view of clients. The Software Group in particular has embraced the MAP methodology, using it to identify investment accounts supported by 665 sales representatives dedicated exclusively to investment territories.

As resource deployment investments are expected to drive incremental revenue growth, it is very important to measure the MAP impact from a revenue growth perspective. At the same time, we would not expect the impact of those investments to occur quickly, as any shifted resource will need time to ramp up to full productivity. The impact can therefore be assessed by comparing the year-over-year revenue growth across each of the MAP account segments (e.g. Invest). In particular, the growth of the Invest accounts relative to the growth for US accounts as a whole is a key measure of performance. In addition to revenue growth, we can use sales pipeline growth and sales quota attainment for the population of sellers who were shifted to cover these Invest accounts as further measures of impact.

### *6.2.2 Revenue Growth*

If we look at large client accounts in the US, the year-over-year revenue growth during the first half of 2006 was 5% higher in the MAP-identified Invest accounts than for the background of all US accounts. While we cannot state unequivocally that all of this 5% growth is due to the MAP process, internal analysis suggests that there is some causal effect. It is expected that this contribution will increase as shifted resources are given more time to produce results.

### *6.2.3 Sales Pipeline Growth*

The sales pipeline, as derived through the opportunity management system, is an important leading indicator of future revenue. Here again, growth of the investment account segment, as well as the contribution of the investment segment to the total pipeline, is an important indicator of impact. It is also important to recognize that any impact on the sales pipeline resulting from MAP will occur over some period of time beyond the current quarter. We can therefore use a rolling four quarters worth of validated pipeline as an appropriate measure. As of week 12 in 3Q 2006, the validated sales pipeline of Invest accounts (over a rolling four quarter period) grew year over year at a rate of 14% greater than the total US sales pipeline. As sales pipeline is a leading indicator of revenue, the fact that pipeline growth is

[4] The sales pipeline refers to the volume of potential sales opportunities with some probability of closing. A sales opportunity is labeled as "won" when a sales contract is signed.

[5] The OnTARGET application logs the time and user ID associated with all company-detail page accesses.

greater than revenue growth in the Invest accounts is further evidence of the financial impact of MAP.

### *6.2.4 Quota Attainment*

A further measure of impact is the performance of those sales resources that are either shifted or dedicated to Invest sales territories as a result of MAP. For the first two quarters in 2006, the year-to-date quota attainment of the shifted resources was 45%, compared to 36% for resources shifted as a result of other initiatives. This suggests that MAP has identified greater sales opportunities, and that movement of resources to these accounts has yielded increased productivity.

## 7. Conclusions

OnTARGET and MAP are examples of analytics-based solutions that were designed from the outset to address specific business challenges in the broad area of sales force productivity. Although they address different underlying issues, these solutions implement a common approach that is generally applicable to a broad class of operational challenges. Both solutions rely on rigorously defined data models that integrate all relevant data into a common database. Choices of the data to be included in the data model are driven both by end-user requirements as well as the need for relevant inputs to analytical models. Both business problems have a natural mapping to applications of predictive modeling: predicting the probability to purchase in the case of OnTARGET, and estimating the realistic revenue opportunity in the case of MAP. Delivering the underlying data and the analytic insights directly to frontline decision makers (sales representatives for OnTARGET and sales executives for MAP) is crucial to driving business impact, and a significant effort has been invested in developing efficient web-based tools with the necessary supporting infrastructure. Both solutions have been deployed across multiple geographic regions, with a strong focus on capturing and quantifying the business impact of the initiatives. Indeed, we have field evidence that the analytical models developed for OnTARGET are predictive. MAP is a more recent initiative, but preliminary evidence suggests that sales force allocations made within the MAP process are leading to measurable improvements in sales efficiency. Finally, although we have implemented these solutions within IBM, we believe that the underlying methodologies, business processes, and potential impact are relevant to enterprise sales organizations in many other global industries.

## 8. Acknowledgements

We acknowledge the significant contributions of the following people to the development and deployment of the OnTARGET and MAP web-based tools: Jorge Arroyo, Matt Callahan, Matt Collins, Alexey Ershov, Sheri Feinzig, Mark Niemaszyk, Georges Atallah, Kevin Bailie, Madhavi Bhupathiraju, Mike Burdick, Upendra Chitnis, Steve Garfinkle, Elizabeth Hamada, Katherine Hanemann, Joan Kennedy, Keith Little, Kyle Keogh, Shiva Kumar, John LeBlanc, Imad Loutfi, Agatha Liu, John Pisello, Mike Provo, Colleen Siciliano, Ashay Sathe, Diane Statkus, Shan Sundaram, Ruth Thompson, Nancy Thomas, Lisa Yu, and Brian Zou.

## 9. REFERENCES

[1] Ledingham, D., M. Kovac, and H. Simon, *The New Science of Sales Force Productivity.* Harvard Business Review, September, 2006: p. 124-133.

[2] Zikmund, W.G., R. McLeod, and F.W. Gilber, *Customer Relationship Management: Integrating Marketing Strategy and Information Technology*. 2002: John Wiley.

[3] Berry, M.J.A. and G.S. Linoff, *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. 2004: John Wiley.

[4] Morgan, A.J. and S.A. Inks, *Technology and the Sales Force - Increasing Acceptance of Sales Force Automation.* Industrial Marketing Management, 2001. 30(5): p. 463-472.

[5] Speier, C. and V. Venkatesh, *The Hidden Minefields in the Adoption of Sales Force Automation Technologies.* Journal of Marketing, 2002. 68(3): p. 98-111.

[6] *IBM Rational Data Architect,* http://www.ibm.com/software/data/integration/rda/.

[7] *IBM WebSphere DataStage,* http://www.ibm.com/software/data/integration/datastage/.

[8] Rosset, S. and R. Lawrence. *Data Enhanced Predictive Modeling for Sales Targeting*. In *2006 SIAM Conference on Data Mining*. 2006. Bethesda, MD: SIAM.

[9] Hastie, T., R. Tibshirani, and J.H. Friedman, *Elements of Statistical Learning,*. 2003: Springer.

[10] Rosset, S., E. Neumann, and U. Eick. *Evaluation of Prediction Models for Campaign Planning*. In *The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2001. San Francisco, California: ACM.

[11] Koenker, R., *Quantile Regression*. Econometric Society Monograph Series. 2005, Cambridge University Press.

[12] Langford, J., R. Oliveira, and B. Zadrozny. *Predicting Conditional Quanting via Reduction to Classification*. In *22nd Conference on Uncertainty in Artificial Intelligence*. 2006. Cambridge, MA, USA: MIT.

[13] Merugu, S., S. Rosset, and C. Perlich. *A New Multi-View Regression Approach with an Application to Customer Wallet Estimation*. In *12th SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006. Philadelphia, PA: ACM.

[14] Breiman, L., et al., *Classification and Regression Trees*. 1984, Belmont, CA: Wadsworth International Group.

[15] Perlich, C., S. Rosset, R. Lawrence, and B. Zadrozny, *High Quantile Modeling for Customer Wallet Estimation with Other Applications.* In *13th SIGKDD International Conference on Knowledge Discovery and Data Mining* 2007, San Jose, CA: ACM

[16] Mitchell, T.M., *Machine Learning*. 1997: WCB McGraw Hill.

[17] Rosset, S., C. Perlich, and B. Zadrozny, *Ranking-Based Evaluation of Regression Models.* Knowledge and Information Systems, 2006

# How Investigative Data Mining Can Help Intelligence Agencies in Understanding and Splitting Terrorist Networks

Nasrullah Memon
Aalborg Universitet Esbjerg
Niels Bohrs Vej 8
DK-6700, Esbjerg, Denmark

David L. Hicks
Aalborg Universitet Esbjerg
Niels Bohrs Vej 8
DK-6700, Esbjerg, Denmark

Henrik Legind Larsen
Aalborg Universitet Esbjerg
Niels Bohrs Vej 8
DK-6700, Esbjerg, Denmark

Abdul Qadeer Khan Rajput
Mehran University of
Engineering and Technology
Jamshoro, Sindh, Pakistan

## ABSTRACT

This article focuses on the study and development of recently introduced new measures, theories, mathematical models and algorithms to support the analysis, visualization and destabilizing of terrorist networks. Specific models and tools are described, and applied to case studies to demonstrate their applicability to the area. A large knowledge base of terrorist events that have occurred or been planned, is developed by harvesting the Web. An investigative data mining tool is also described that packages together many of the techniques to support the analysis of terrorist networks. We are confident that the tool described can help intelligence agencies in understanding and splitting terrorist networks.

## 10. INTRODUCTION

The threats facing society today require new methods for modeling and analysis. Civil security decision makers, analysts and field operators fighting terrorism and organized crime all need front-line integrated technologies to support their cooperative work. Our opponents are no longer organized in hierarchical structures, but instead consist of individuals and groups that are loosely organized in "dark networks". Instead of large-scale military attacks, they stage attacks or set bombs against unprotected civilians, or seek to influence crowds of legitimate demonstrators so that critical riot situations occur.

In order to construct decision support systems that take account of these new factors, new, more powerful methods and techniques from several technological domains need to be brought together and integrated. Experience shows that the networks can be unwound and analyzed after the events. Although it provides the necessary evidence for bringing criminals to justice, it is then too late to prevent loss of life and material damage.

Mathematical methods used in our research on Investigative Data Mining [1] [2][3] [4] are clearly relevant to law enforcement intelligence work and may provide tools to discover terrorist networks in their planning phase and thereby prevent terrorist acts and other large-scale crimes from being carried out. Relevant patterns to investigate include connections between actors (meetings, messages), activities of the involved actors (specialized training, purchasing of equipment) and information gathering (time tables, visiting sites).

Investigative Data Mining (IDM) offers the ability to firstly map a covert cell, and to secondly measure the specific structural and interactional criteria of such a cell. This framework aims to connect the dots between individuals and "map and measure complex, covert, human groups and organizations". The method focuses on uncovering the patterning of people's interaction, and correctly interpreting these networks assists "in predicting behavior and decision-making within the network".

IDM borrows social network analysis (SNA) and graph theory techniques for connecting the dots; our goal is to propose mathematical methods for destabilizing terrorist networks after linking the dots between them.

In investigative data mining, a number of variations exist in the literature. One is known as link analysis (see for example [5] [6]). Link analysis research uses search and probabilistic approaches to find structural characteristics in the network such as hubs, gatekeepers, pulse-takers [7], or identifying potential relationships for relational data mining. Link analysis alone is insufficient as it looks at one side of the coin and ignores complex nonlinear relationships that may exist between the attributes. Another approach depends purely on visualization, such as NetMap [8]. Unfortunately, these tools that depend on visualization alone - despite being useful to provide some insight - are insufficient and rely on the user to carry out many tedious and time consuming tasks, many of which could be automated.

In addition to the previous discussion, most of the work on link analysis or network visualization ignores the construction of the hidden hierarchy of covert networks. Uncovering a relationship among or within attributes (connecting the dots) is an important step, but in many domains it is more important to understand how this relationship evolved. Hence, understanding network

dynamics and evolution is needed to complete the picture. Once we understand the dynamics and evolution of these relationships and construct the hidden hierarchy, we can search for ways to disconnect the dots when and if needed.

The three innovative points of our research are:

1. The use of a new measure of position role centrality on the pattern of efficiency introduced by Vito Latora and Massimo Marchiori [52]. This measure identifies key players (gatekeepers/ leaders) and followers in the network.
2. The use of another measure known as dependence centrality which discovers who is depending on whom in a network.
3. The estimate of a possible hierarchical structure of a complex network by applying degree centrality and Eigenvector centrality from social network analysis (SNA) literature and combining it with the new measure of dependence centrality.

This paper presents some case studies of the terrorist events that have occurred or been planned, using a software prototype that we have developed. The structure of the rest of this paper is as follows. In Section 2, a brief introduction about investigative data mining is presented, whereas Section 3 describes a brief introduction about graph theory. Section 4 presents an overview of terrorist network analysis. Section 5 discusses a concise review to point the reader to several key papers in the literature; whereas Section 6 and Section 7 discusses about topological analysis of terrorist networks; whereas Section 8 describes the models we have used for destabilizing terrorist networks. Section 9 discusses an overview of the iMiner software prototype system; whereas two case studies are then presented in Section 10. Conclusions are then drawn in Section 11, whereas future recommendations are discussed in Section 12.

## 11. INVESTIGATIVE DATA MINING[1]

Defeating terrorist networks requires a nimble intelligence apparatus that operates actively and makes use of advanced information technology. Data mining for counterterrorism (also known as investigative data mining[6]) is a powerful tool for intelligence and law enforcement officials fighting against terrorism [8]. Investigative data mining is a combination of data mining and subject-based automated data analysis techniques. "Data mining" actually has a relatively narrow meaning: the approach that uses algorithms to discover predictive patterns in datasets. Subject-based automated data analysis applies models to data to predict behaviour, assess risk, determine associations, or do other types of analysis [8]. The models used for automated data analysis can be used on patterns discovered by data mining techniques.

Although these techniques are powerful, it is a mistake to view investigative data mining techniques as a complete solution to security problems. The strength of IDM is to assist analysts and investigators. IDM can automate some functions that analysts would otherwise have to perform manually. It can help to prioritize attention and focus an inquiry, and can even do some early analysis and sorting of masses of data. Nevertheless, in the complex world of counterterrorism, it is not likely to be useful as the only source for a conclusion or decision.

In the counterterrorism domain, much of the data could be classified. If we are to truly get the benefits of the techniques we need to test with actual data. But not all researchers have the clearances to work on classified data. The challenge is to find unclassified data that is, representative of the classified data. It is not straightforward to do this, as one has to make sure that all classified information, even through implications, is removed. Another alternative is to find as good data as possible in an unclassified setting for researchers to work on. However, the researchers have to work not only with counterterrorism experts but also with data mining specialists who have the clearances to work in classified environments. That is, the research carried out in an unclassified setting has to be transferred to a classified setting later to test the applicability of data mining algorithms. Only then do we get the true benefit of investigative data mining.

Investigative data mining is known as a data-hungry project for academia. It can only be used if researchers have good data. For example, in monitoring central banking data, in the detection of an interesting pattern, e.g., a person who is earning less and spending more continuously for the last 12 months, the investigative officer may send a message to an investigative agency to keep the person under observation. Further, if a person, (who is under observation of an investigative agency) has visited Newark airport more than 5 times in a week (a result received from video surveillance cameras), then this activity is termed a suspicious activity and investigative agencies may act. It was not possible for us to get this type of sensitive data. Therefore, in this research, we harvested unclassified data of the terrorist attacks that occurred, or were planned in the past, from a number of authenticated web resources [9].

IDM offers the ability to map a covert cell, and to measure the specific structural and interactional criteria of such a cell. This framework aims to connect the dots between individuals and to map and measure complex, covert, human groups and organizations [10]. The method focuses on uncovering the patterning of people's interaction, and correctly interpreting these networks assists in predicting behaviour and decision-making within the network [10]. The technique is also known as subject based link analysis. This technique uses aggregated public records or other large collections of data to find the links between a subject— a suspect, an address, or a piece of relevant information—and other people, places, or things. This can provide additional clues for analysts and investigators to follow [8].

IDM also endows the analyst with the ability to measure the level of covertness and efficiency of the cell as a whole, and the level of activity, ability to access others, and the level of control over a network each individual possesses. The measurement of these criteria allows specific counter-terrorism applications to be drawn, and assists in the assessment of the most effective methods of disrupting and neutralising a terrorist cell [10]. In short, IDM provides a useful way of structuring knowledge and framing further research. Ideally it can also enhance an analyst's predictive capability [10].

[6] The term is firstly used by Jesus Mena in his book Investigative Data Mining and Criminal Detection, Butterworth (2003) [15].

On the other hand, traditional data mining commonly refers to using techniques rooted in statistics, rule-based logic, or artificial intelligence to comb through large amounts of data to discover previously unknown but statistically significant patterns. However, in the application of IDM in the counterterrorism domain, the problem is much harder, because unlike traditional data mining applications, we must, find an extremely wide variety of activities and hidden relationships among individuals. Table 1 gives a series of reasons for why traditional data mining isn't the same as investigative data mining.

In this research we have chosen to use a very small portion of data mining for counterterrorism; and we have borrowed techniques from social network analysis and graph theory for connecting the dots. Our goal is to propose mathematical methods, techniques, and algorithms to assist law enforcement and intelligence agencies in destabilizing terrorist networks.

| Traditional Data Mining | Investigative Data Mining |
|---|---|
| Discover comprehensive models of databases to develop statistically valid patterns | Detect connected instances of rare patterns |
| No starting points | Known starting points or matches with patterns estimated by analysts |
| Apply models over entire data | Reduce search space; results are starting points for human analysts |
| Independent instances (records) | Relational instances (networked data) |
| No correlation between instances | Significance autocorrelation |
| Minimal consolidation needed | Consolidation is key |
| Dense attributes | Sparse attributes |
| Sampling is used | Sampling destroys connections |
| Homogeneous data | Heterogeneous data |
| Uniform privacy policy | Non-uniform privacy policy |

Traditionally, most of the literature in SNA has focussed on networks of individuals. Although SNA is not conventionally considered as a data mining technique, it is especially suitable for mining a large volume of association data to discover hidden structural patterns in terrorist networks. SNA primarily focuses on applying analytic techniques to the relationships between individuals and groups, and investigating how those relationships can be used to infer additional information about the individuals and groups [11]. There are a number of mathematical and algorithmic approaches that can be used in SNA to infer such information, including connectedness and centrality [12].

Law enforcement personnel have used social networks to analyze terrorist networks [13, 14] and criminal networks [13]. The capture of Saddam Hussein was facilitated by social network analysis: military officials constructed a network containing Hussein's tribal and family links, allowing them to focus on individuals who had close ties to Hussein [16].

After the 9/11 attacks, SNA has increasingly been used to study terrorist networks. Although these covert networks share some features with conventional networks, they are harder to identify, because they mask their transactions. The most significant complicating factor is that terrorist networks are loosely organized networks having no central command structure; they work in small groups, who communicate, coordinate and conduct their campaign in a network like manner. There is a pressing need to automatically collect data of terrorist networks, analyze such networks to find hidden relations and groups, prune datasets to locate regions of interest, detect key players, characterize the structure, trace the point of vulnerability, and find the efficiency of the network [17]. Hence, it is desirable to have tools to detect the hidden hierarchical structure of horizontal terrorist networks [18] in order to assist law enforcement agencies in the capture of the key players so that most of the network can be disrupted after their capture.

## 12. GRAPH THEORY[7]

Throughout this article the topology of a network is represented by a *graph* G = ($V$, $E$) which is an abstract object, formed by a finite set $V$ of vertices (m = | $V$ |) and a finite set $E$ of edges (n = | $E$ |). An edge $e = (u, v) \in E$ connects two vertices $u$ and $v$. The vertices $u$ and $v$ are said to be *incident* with the edge $e$ and *adjacent* to each other. The set of all vertices which are adjacent to $u$ is called the neighbourhood $N(u)$ of $u$: ($N(u) = \{v: (u, v) \in E\}$). A graph is called *loop-free* if no edge connects a vertex to itself. An *adjacency matrix A* of a graph $G = (V, E)$ is a $(m \times m)$ matrix, where $a_{uv} = 1$ if and only if (u, v) $\in E$ and $a_{uv} = 0$ otherwise. Graphs can be *undirected* or *directed*. The adjacency matrix of an undirected graph is symmetric. An undirected edge joining vertices $u, v \in V$ is denoted by $\{u, v\}$.

In directed graphs, each directed edge (arc) has an *origin* (*tail*) and a *destination* (*head*). An edge with origin $u \in V$ is represented by an ordered pair $(u, v)$. As a shorthand notation, an edge $\{u, v\}$ can also be denoted by $uv$. It should be noted that, in a directed graph, $uv$ is short for $(u, v)$, while in an undirected graph, $uv$ and $vu$ are the same and both stand for $\{u, v\}$. Graphs that can have directed as well undirected edges are called *mixed graphs*, but such graphs are encountered rarely. Let $(e_1,\ldots,e_k)$ be a sequence of edges in a graph $G = (V, E)$. This sequence is called a *walk* if there are vertices $v_0,\ldots,v_k$ such that $e_i = (v_{i-1}, v_i)$ for $i = 1, \ldots, k$. If the edges $e_i$ are pairwise distinct and the vertices $v_i$ are pairwise distinct the walk is called a *path*. The *length* of a walk or path is given by its number of edges, $k = |\{e_i,\ldots,e_k\}|$. The *shortest path* between two vertices $u$, $v$ is a path with minimum length, all shortest paths between $u$ and $v$ are called *geodesics*.

The *distance* (dist $(u, v)$) between two vertices $u$, $v$ is the length of the shortest paths between them. The vertices $u$, $v$ are called *connected* if there exists a walk from vertex $u$ to vertex $v$. If all pairs of distinct vertices of graph $G = (V, E)$ are connected, the graph is called *connected*.

[7] Most of the concepts discussed in this section are taken from [19]

In the remainder of this article, we consider only undirected, loop-free connected graphs.

# 13. TERRORISM NETWORK ANALYSIS

The threat from modern terrorism manifests itself worldwide in locally and internationally operating network structures. Before focusing on the development and composition of these networks, we will initially attempt to answer the question: what exactly is a terrorist network?

Persons involved in support, preparation or commission of terrorist attacks almost never operate alone, but as members of - sometimes overlapping - network structures. Within these networks they co-operate with individual members or small groups of members (operational cells). A modern terrorist *network* differs from other terrorist groups and organizations in that it lacks a formal (hierarchical) structure, and has an informal, flexible membership and fluctuating leadership. It is incorrect, however, to conclude that such a network possesses no structure whatsoever. There is always a pattern of connections between individuals who communicate with one another with a view to achieving a common goal. In some cases these communication lines converge in one or more core groups, which thus play a coordinating and controlling role. In other cases there are random communication patterns between all members while the network functions practically without any leadership or central control. It is also possible for several groups to be active within one network.

The flexible and informal character of such a network makes it easy for individual members to establish temporary ad-hoc contacts, in addition to more permanent relations. It also leaves room for personal initiative. The relations within a network are constantly changing in character and duration. In most cases we can distinguish a core group surrounded by a diffuse network of individuals, with central control usually restricted to a minimum. Personal ties between members bind the network together. These relationships are usually based on a shared political-religious ideology, mutual trust, family or friendship ties, shared origin and/or shared experiences in training camps or jihad areas. The notion of a common *enemy* also stimulates bonding among network members.

The above characteristics lead to the following definition:

*A terrorist network is a fluid, dynamic, vaguely delineated structure comprising a number of interrelated persons who are linked both individually and on an aggregate level (cells / groups). They have at least a temporary common interest, i.e. the pursuit of a jihadism-related goal (including terrorism).*

Persons within such a network are referred to as members. A member is a person who contributes actively and consciously to the realization of the aforementioned goal within the bounds of the network.

This definition is in line with the definition of criminal networks used in Criminology, which does not refer to permanent structures, but to temporary, flexible co-operative structures between individuals, based on kinship, friendship, business opportunism, coincidence, necessity, temptation and force, or to the fact that members are colleagues, neighbors or fellow convicts. This co-operation gradually evolves into certain customs and traditions which lead to 'habituation, mutual interdependence and trust, and hierarchical relations [20]. This assessment of fluid and dynamic criminal networks was described in an extensive study into organized crime [21].

## 4.1 Centrality measures for Analyzing Terrorist Networks

Centrality is one of the most important and widely used measures for analyzing social networks. Nearly all empirical studies try to identify the most important actors (also known as vertices / nodes) within the network. Four measures of centrality are commonly used in network analysis: degree, closeness, betweenness, and Eigenvector centrality. The first three were described in modern form by Freeman [22] while the last was proposed by Bonacich [23]. Let us begin with degree centrality.

"The *degree* of a node, *v* is simply the count of the number of encounters with other nodes, that are adjacent to it, and with which it is, therefore in direct contact" [22]; it is known as a measure of activity. The degree centrality $C_D(v)$ of a vertex *v* is simply defined as the degree *d*(*v*) of *v* if the considered graph is undirected. The degree centrality is, e.g., applicable whenever the graph represents something like a voting result. These networks represent a static situation and we are interested in the vertex that has the most direct votes or that can reach most other vertices directly. The degree centrality is a local measure, because the centrality value of a vertex is only determined by the number of its neighbours. The most commonly employed definition of degree centrality is:

$$C_d(u) = \sum r(u,v) \quad \textbf{(1)}$$

Where *r* (*u*, *v*) is a binary variable indicating whether a link exists between nodes *u* and *v*.

A degree based measure of node centrality can be extended beyond direct connections to those at various path distances. In this case, the relevant neighbourhood is widened to include the more distant connections of the nodes. A node may, then, be assessed for its local centrality in terms of both direct (distance 1) and distance 2 connections—or, indeed, whatever cut-off path distance is chosen. The principal problem with extending this measure of node centrality beyond distance 2 connections is that, in graphs with even a very modest density, the majority of the nodes tend to be linked through indirect connections at relatively short path distances.

Thus a comparison of local centrality scores at a distance 4 is unlikely to be informative if most of the nodes are connected to most other nodes at this distance.

The degree, therefore, is a measure of local centrality, and a comparison of the degrees of various nodes in a graph can show how well connected the nodes are with their local environments.

This measure of local centrality has one major limitation. That is comparisons of centrality scores can only meaningfully be made among members of the same graph or between graphs that are the same size. The degree of a node depends on, among other things,

the size of the graph, and so a measure of local centrality cannot be compared when graphs differ significantly in size.

Local centrality is, however, only one conceptualization of node centrality, and Freeman [22] has proposed a measure of global centrality based on what he terms the closeness of the nodes.

The second measure relates to the *closeness* or the *distance* between nodes. According to Freeman [22], this closeness measure can be conceptualized as independence (the extent to which an actor can avoid the control of others) or efficiency (extent to which an actor can reach all other actors in the shortest number of steps). Thus, it measures independent access to others. A central actor can reach other actors through a minimum number of intermediary positions and is therefore dependent on fewer intermediary positions than a peripheral actor.

Suppose a terrorist organization wants to establish a new camp, for example, a human bomb training camp, such that the total distance to all persons interested to kill themselves, for a cause, in the region is minimal. This would make travelling to the camp as convenient as possible for most people who are living in that region and are willing to be used for human bombs in the near future.

We denote the sum of the distances from a vertex $u \in V$ to any other vertex in a graph $G = (V,E)$ as the total distance $\sum_{u \in V} d(u,v)$. The problem of finding an appropriate location can be solved by computing the set of vertices with a minimum total distance.

In SNA literature, a centrality measure based on this concept is called *closeness* centrality. The focus lies here, for example, on measuring the closeness of a person to all other people in the network. People with a small total distance are considered as more important as those with high total distance. The most commonly employed definition of closeness is the reciprocal of the total distance:

$$C_C(u) = \frac{1}{\sum_{v \in V} d(u,v)} \quad \textbf{(2)}$$

$C_C(u)$ grows with decreasing total distance of $u$, therefore it is also known as a structural index. Unlike degree centrality this measure is a global metric.

The third measure is called *betweenness* and is the frequency at which a node occurs on a geodesic that connects a pair of nodes. Thus, any node that falls on a shortest path between other nodes can potentially control the transmission of information or effect exchange by being an intermediary. "It is the potential for control that defines the centrality of these nodes" [24]. Thus, if two persons A and C are connected only through person B, B would fall between A and C and would have control of any resources that flow between A and C. This measure easily discovers gatekeepers.

Let $\delta_{uw}(v)$ denotes the fraction of shortest paths between $u$ and $w$ that contain vertex $v$:

$$\delta_{uw}(v) = \frac{\sigma_{uw}(v)}{\sigma_{uw}} \quad \textbf{(3)}$$

where $\sigma_{uw}$ denotes the number of all shortest-paths between $u$ and $w$. The ratio $\delta_{uw}(v)$ can be interpreted as the probability that vertex $v$ is involved into any communication between $u$ and $w$. Note, that the measure implicitly assumes that all communication is conducted along shortest paths. Then the betweenness centrality $C_B(v)$ of a vertex $v$ is given by:

$$C_B(v) = \sum_{u \neq v \in V} \sum_{w \neq v \in V} \delta_{uw}(v) \quad \textbf{(4)}$$

Any pair of vertices $u$ and $w$ without any shortest path between them will add zero to the betweenness centrality of every other vertex in the network. This measure is also a global metric like closeness centrality.

Actually the definition of betweenness centrality explores an actor's ability to be "irreplaceable" in the communication of two random actors. It is of particular interest in the study of destabilizing terrorists by network attacks, because at any given time the removal of the maximum betweenness actor seems to cause maximum damage in terms of connectivity and average distance in a network.

A more sophisticated version of the same idea is the so-called Eigenvector centrality. The *Eigenvector centrality* ($C_{ev}$) of a node in a network is defined to be proportional to the sum of the centralities of the node's neighbours, so that a node can acquire high centrality either by being connected to a lot of others (as with simple degree centrality) or by being connected to others that themselves are highly central.

The eigenvector centrality can be understood as a refined version of degree centrality in the sense that it recursively takes into account how neighbour nodes are connected.

The idea is that even if a node has a few ties, if those few nodes influence many others (who themselves influence still more others), then the first node in that chain is highly influential. In a terrorist network, if a person has the potential to get bomb making training from a few neighbours, and those neighbours have already trained in terrorist camps and are on high security risk, the potential risk of getting bomb making training for the first person is very high.

It is defined as:

$$x_i = \frac{1}{\lambda} \sum_{j=1}^{n} A_{i,j} x_j \quad \textbf{(5)}$$

where n is the total number of nodes and λ is the largest eigenvalue to assure the centrality is non-negative. Thus, $x_i$ is the $i^{th}$ component of eigenvector associated with the largest eigenvalue λ of the network. While the eigenvector centrality of a network can be calculated via standard methods [25] using the adjacency matrix representation of the network, it can also be computed by an iterative degree calculation method, known as accelerated power method [26]. This method is not only more efficient, but consistent with the spirit of the refined version of degree centrality.

### 4.1.1 Example

Figure 1 shows an example of a terrorist network, which maps the links between terrorists involved in the tragic events of September 11, 2001. This graph was constructed by Valdis Krebs [13] using the public data that were available before, but collected after the event. Even though the information mapped in this network is by no means complete, its analysis may still provide valuable insights into the structure of a terrorist organization. This graph is reconstructed in this paper, using metadata (additional information) of every terrorist involved in the attacks.

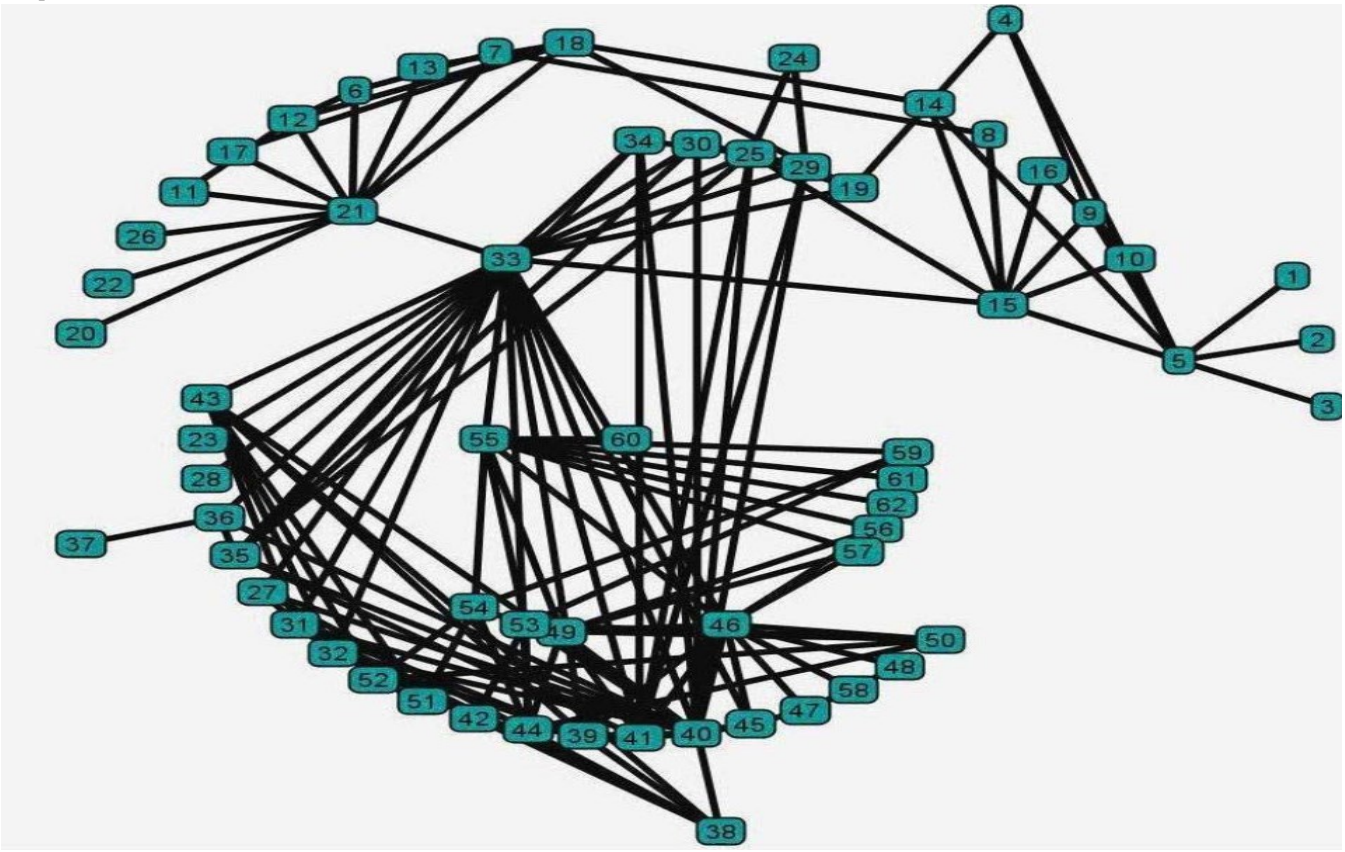



**Figure 1. The dataset of 9/11 hijackers and their affiliates. The dataset was originally constructed by Valdis Krebs [13], but is here re-constructed in our investigative data mining prototype iMiner, using metadata of every terrorist.**

According to Kreb's analysis [13], this network had 62 members in total, of which 19 were kidnapers, and 43 assistants: organizers, couriers, financiers, scouts, representatives, coordinators, counterfeiters, etc. Allen [27] found that successfully functioning large networks typically comprise 25-80 members, with an optimal size between 45 and 50. A close match exists between the results of Allen's analysis of collaborating networked groups and this particular example of a terrorist group.

Inspection of this network by standard measures of network structure reveals firstly its low connectedness. A member of this network holds only 4.9 connections with other members on average (also known as degree centrality), which means that average members were rather isolated from the rest of the network. The density (which is defined as the number of actual links divided by the number of possible links) of this network is only 0.08, meaning that only 8% of all possible connections in the network really exist (see Figure 2).

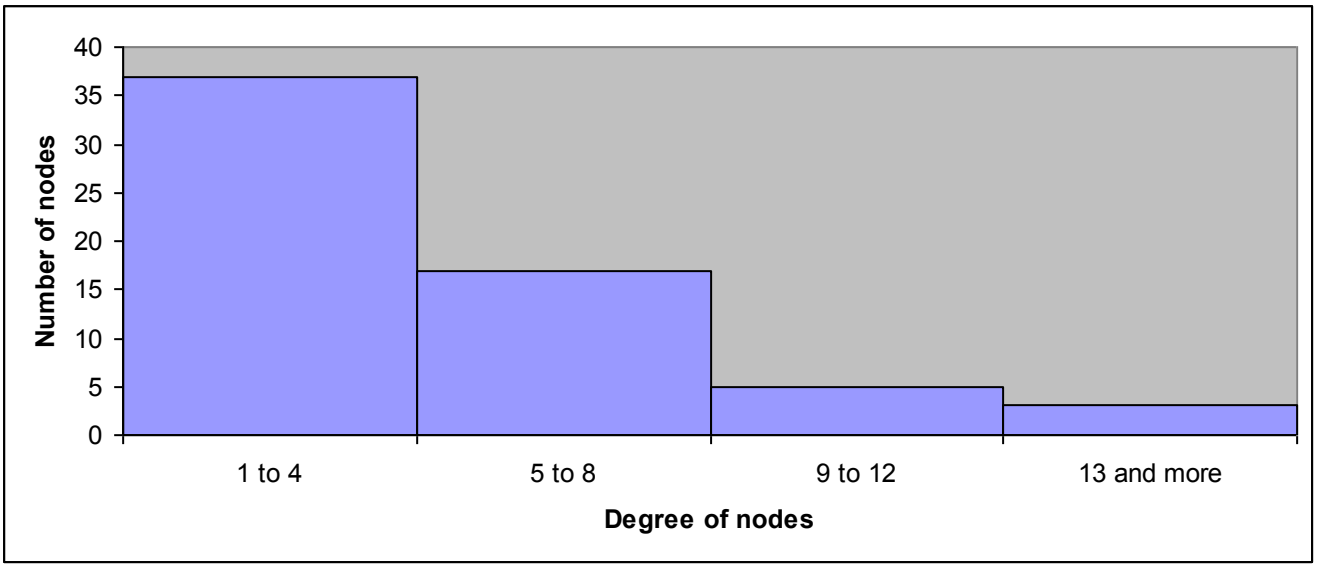



**Figure 2. Distribution of degrees of nodes in the network (see Figure 1) of kidnappers and their supporters.**

In spite of low connectedness, however, the nodes of this network are relatively close. The average closeness of nodes is 0.35. Betweenness as stated above is another important measure in SNA and it indicates a node's importance for communication among other nodes. The average betweenness of this network is 0.032, indicating relatively high average redundancy. However the betweenness of 40 nodes is in fact less than 1% and only 6 nodes have betweenness higher than 10%. These 6 nodes are critical for information flow, especially one with betweenness of almost 60%, meaning that almost 60% of communication paths among other nodes pass through this central node. The node (33) represents *Mohamed Atta*, the leading organizer of the attack whose central position in the network is confirmed by other centrality indicators as well.

The distribution of the degrees of nodes is particularly interesting. The degrees of the nodes are exponentially distributed: the degree of most of the nodes is small, while few nodes have high degree (see Figure 2). This property characterizes the so called scale free networks [28] [29]. Scale free networks form spontaneously, without needing a particular plan or the interventions of a central authority. Nodes that are members of the network for a longer time, that are better connected with other nodes, and that are more significant for the network's functioning, are also more visible to new members, so that the new members spontaneously connect more readily to such nodes than other, relatively marginal ones.

On the pattern of scale free networks, the Al Qaeda Training Manual [30] states: "The cell or cluster methods should be organized in a way that a group is composed of many cells whose members do not know each other, so that if a cell member is caught, other cells would not be affected, and work would proceed normally".

## 4.2 Link Analysis

Link analysis is an analytic technique for making relationships explicit. Link analysis is the process of building up networks of interconnected objects through relationships in order to expose patterns and trends. Link analysis uses item-to-item associations to generate networks of interactions and connections from defined datasets. Link analysis methods add dimensions to an analysis that other forms of visualization do not support. Link analysis can be used to construct inferential structures of organizations or interactions which can be tested later. It is very well suited for hypothesis construction and can be applied to a variety of problems in Armed Forces Intelligence (for example, orders of battle), in Political Intelligence and in Sociological Intelligence research and analysis [31].

Despite the seeming novelty of link analysis, the federal government in the USA has used link analysis, for nearly fifty years. Karl Van Meter describes the two main types of link analysis: the village survey method and traffic analysis [32]. The village survey method was created and used by Ralph McGehee of the CIA in Thailand in the 1960s to understand family and community relationships. He conducted a series of open-ended interviews and in a short time was able to map out the clandestine structure of local and regional Communist organizations and associated sympathetic groups. Traffic analysis (also known as communication link analysis) began during World War II and its importance continues to this day. This technique consists of the study of the external characteristics of communication in order to get information about the organization of the communication system. It is not concerned with the content of phone calls, but is interested in who calls whom and the network members, messengers, and brokers. Traffic analysis was also used by the British MI5 internal security service to combat the IRA in the 1980s and 1990s and continues to be used across the world by

law-enforcement agencies including the US Defense Intelligence Agency (DIA) Office of National Drug Control Policy [32].

The *Analyst Notebook* [33] is the primary software used for link analysis. Currently on its sixth version, this software is recognized as one of the world's leading analytical tools and is employed in more than 1,500 organizations. SNA improves upon link analysis by moving from single variable analysis to multivariate analysis, allowing the individual to control many factors at once. The change from single variable to multivariate analysis is quite significant when researching terrorism: a number of factors affect terrorism, not one single factor. For example, the prediction for one to participate in a terrorist activity might not be strongly affected by the single variable of being related to a terrorist member. However, the combination of multiple variables, such as poverty and type of government, combined with the link to a terrorist member, may cause a person to participate in a terrorist activity. Multivariate analysis allows us to take into account these multiple variables and their effects when controlling another variable.

From the outside, it is difficult to understand precisely how link analysis is being used in federal governments. Confidentiality prevents government analysts from discussing their work with researchers and private companies without security clearances. Despite this lack of information, it is clear that governments are interested in using network techniques in dealing with counterterrorism strategies. Many government agencies, such as Defense Advanced Research Projects Agency (DARPA), U.S. Army Research Labs, the U.S. Office of Naval Research (ONR), the National Security Agency (NSA), the National Science Foundation (NSF), and the Department of Homeland Security (DHS) have funded research related to link analysis.

The link analysis systems are being used in the investigative world, but from unclassified work, it is known that they are only used for identifying central players and some interesting patterns from the available datasets. Apparently little work is carried out for the destabilizing of terrorist networks [1] [2] [4] [34] [35]. The motivation behind this study is to connect the dots in order to assist law enforcement agencies to disconnect / destabilize most of the network by capturing/eradicating some key players.

# 5. EXISTING APPROACHES

Existing terrorist network research is still at its incipient stage. Although previous research, including a few empirical studies, have motivated the call for new approaches to terrorist network analysis [34] [36] [37], studies have remained mostly small-scale and used manual analysis of a specific terrorist organization. For instance, Krebs [13] manually collected data from public news releases after the 9/11 attacks and studied the network surrounding the 19 hijackers. Sageman [38] analyzed the Global Salafi Jihad network consisting of 171 members using a manual approach and provided an anecdotal explanation of the formation and evolution of this network. None of these studies used advanced data mining technologies that have been applied widely in other domains such as finance, marketing, and business to discover previously unknown patterns from terrorist networks.

The papers [34] [39] provide examples of social network analysis in counterterrorism applications and indicate both usefulness and some limitations of social network analysis as a basis for quantitative methods for situation awareness and decision-making in law enforcement applications. Raab & Milward [39] discuss the organizational structure of certain drug trafficking, terrorism, and arms trafficking networks, showing how some of them have adapted to increased pressure from the States and international organizations by decentralizing into smaller units linked only by function, information, and immediate need. They also describe ways and structures of cooperation between different kinds of criminal networks, for example in financing terrorists by illegal diamond and drug trafficking.

Recently, computer scientists have become interested in network analysis. This has led to an increased emphasis on studying the statistical properties of large networks, such as the Internet, criminal networks, and, even, infrastructure networks [40]. This influx of people to the field has also led to several new approximate algorithms to compute important properties [1] [2] [41] [42]. Most of the above mentioned literature has used graph theory and SNA techniques for terrorist network analysis.

Jonathan D. Farley presented a new mathematical approach [35] to destabilize terrorist networks using order theory. This paper pointed out that modeling terrorist networks as graphs does not give enough information to deal with the threat, "modeling terrorist cells as graphs ignores an important aspect of their structure, namely their hierarchy, and the fact that they are composed of leaders and followers ". Jonathan D. Farley proposed an alternative approach that better reflects an organization's hierarchy. In this case, the relationship of one individual to another in a cell becomes important. Leaders are represented by the topmost nodes in a diagram of the ordered set representing a cell and foot soldiers are nodes at the bottom. Disrupting the organization would be equivalent to disrupting the chain of command, which allows orders to pass from leaders to foot soldiers [35].

What is needed is a set of integrated methods, technologies, models, and tools to automatically mine data and discover valuable knowledge from terrorist networks based on large volumes of data of high complexity.

In the system to be proposed, we represent a terrorist network by an undirected graph; then we convert it into a directed graph with the help of centrality measures. We propose three strategies for destabilizing terrorist networks as discussed in Section 8. In addition we also found that it is imperative for intelligence agencies to understand the characteristics of terrorist networks. For this purpose we present a topological analysis of terrorist networks to identify with the characteristics of terrorist networks.

# 6. STATISTICAL MEASURES OF NETWORK TOPOLOGY

In graph theory a number of measures have been proposed to characterize networks. However, three concepts are particularly important in contemporary studies of the topology of complex networks: degree distribution, clustering coefficient, and average path length [43] [44] [45].

## 6.1 Degree Distribution

The degree $k_i$ is the number of edges connecting to the $i^{th}$ vertex. The vertex degree is characterized by a distribution function $P(k)$, which gives the probability that a randomly selected vertex has $k$ edges. Recent studies show that several complex networks have a heterogeneous topology, i.e., some vertices have a very large number of edges, but the majority of the vertices only have a few edges. That is, the degree distribution follows a power law $P(k) \sqcup k^{-\gamma}$ for large $k$ (i.e., $P(k)/ k^{-\gamma} \rightarrow 1$ when $k \rightarrow \infty$). The average degree $\langle k \rangle$ of a graph with $N$ vertices and $M$ edges is $\langle k \rangle = 2M / N$.

## 6.2 Clustering Coefficient

Many complex networks exhibit an inherent tendency to cluster. In social networks this represents a circle of friends in which every member knows each other. The clustering coefficient is a local property capturing "the density" of triangles in a graph, *i.e.*, two vertices that both are connected to a third vertex are also directly connected to each other. An $i^{th}$ vertex in a network has $k_i$ edges that connect it to $k_i$ other vertices. The maximum possible number of edges between the $k_i$ neighbours is $\binom{k_i}{2} = k_i (k_i - 1) / 2$. The clustering coefficient of the $i^{th}$ vertex is defined as the ratio between the number $M_i$ of edges that actually exist between these $k_i$ vertices and the maximum possible number of edges, i.e., $C_i = 2M_i / k_i(k_i - 1)$. The clustering coefficient of the whole network is given by:

$$C = (1/N) \sum_{i=1}^{n} C_i \sum_{i=1}^{n} C_i$$

## 6.3 Average Path Length

The distance $l_{uv}$ between two vertices $u$ and $v$ is defined as the number of edges along the shortest path connecting them. The average path length $l = \langle l_{uv} \rangle = [1 / N (N - 1)] \sum_{u \neq v \in V} l_{uv}$ is a measure of how a network is scattered. Sometimes, the diameter $d$ of a graph is defined as the maximum path length between any two connected vertices in the graph. However, in other situations the concept of diameter relates to the average path length, i.e., $d = l$. The so-called small-world property appears to characterize many complex networks. Despite their often-large size, there is a relatively short path between any two vertices in a network: the average shortest paths between a pair of vertices scales as the logarithm of the number of vertices.

# 7. GRAPHS AS MODELS OF REAL-WORLD NETWORKS

The study of networks, and in particular the interest in the statistical measures of the topology of networks, has given birth to three main classes of network models. The *random graph* was introduced by Erdos and Renyi in the late 1950s and is one of the earliest theoretical models of a network [46]. This is the easiest model to analyze mathematically and it can serve as a reference for randomness. Watts and Strogatz introduced the so called *small world model* in 1998 [47]. This model combines high clustering and a short average path length.

In 1999, Barabasi and Albert (BA) addressed the origin of the power-law degree distribution, evident in many real networks, with a simple model (also known as the *scale-free network* model) that put the emphasis on how real networks evolve [48].

Three models have been employed to characterize complex networks: the random graph model, the small-world model, and the scale-free model [49]. Most complex systems are not random but are governed by certain organizing principles encoded in the topology of the networks.

A small-world network has a significantly larger clustering coefficient while maintaining a relatively small average path length. The large clustering coefficient indicates that there is a high tendency for nodes to form communities and groups. Scale-free networks, on the other hand, are characterized by the power-law degree distribution, meaning that while a large number of nodes in the network have just a few links, a small fraction of the nodes have a large number of links. It is believed that scale-free networks evolve following the self-organizing principle, where growth and preferential attachment play a key role for the emergence of the power law degree distribution [50]

Although the topological properties of these networks have been discovered, the structures of terrorist networks are largely unknown due to the difficulty of collecting and accessing reliable data (Krebs, 2001). Do terrorist networks share the same topological properties with other types of networks? Do they follow the same organizing principle? How do they achieve efficiency under constant surveillance and threat from authorities? [50].

We apply the small-world network metrics of Watts & Strogatz [47] to Figure 1. One of the key metrics in the small-world model is the average path length, for individuals and for the network overall [51]. A good score for an individual means that he/she is close to all of the others in a network –they can reach others quickly without going through too many intermediaries. A good score for the whole network indicates that everyone can reach everyone else easily and quickly. The shorter the information paths for everyone, the quicker the information arrives and the less distorted it is when it arrives. Another benefit of multiple short paths is that most members of the network have good visibility into what is happening in other parts of the network –a greater awareness. They have a wide *network horizon* which is useful for combining key pieces of distributed intelligence. In an environment where it is difficult to distinguish signal from noise, it is important to have many perspectives involved in the sense-making process [51]

We found that members in the 9/11 terrorist network (Figure 1) are extremely close to their leaders. The terrorists in the network are on average only 1.79 steps away from Mohamed Atta, meaning that Mohamed Atta's command can reach an arbitrary member through only two mediators (approximately). Despite its small size (62), the average path length is 3.01, Information flows quicker, with less distortion, and Mohamed Atta is more involved.

The other small-world topology characteristics a high clustering coefficient, is also present in this network. The clustering coefficient of this network is 0.49, significantly high. Previous studies have also shown the evidence of groups and teams in this network. In these groups and teams, members tend to have denser and stronger relations with one another. The communication between group members becomes more efficient, making a crime or an attack easier to plan, organize, and execute [50].

In addition, this network can be viewed as a scale-free system (as discussed above). The degree distribution decays much more slowly for small degrees than for that of other types of networks, indicating a higher frequency for small degrees.

## 8. DESTABILIZING TERRORIST NETWORKS

It is a good strategy if intelligence agencies and law enforcement know whom to capture from a terrorist network, so that by killing or eradicating one terrorist, a maximum network is disrupted. In this research we propose three strategies, which may assist law enforcement and intelligence agencies in the disruption of terrorist networks:

1. Determination of how much the efficiency of a network is affected by removing one or more nodes in the network.
2. Determination of how many nodes in a network are dependent on one node. If many nodes are dependent on one single node, then it might be a good strategy to spend energy in capturing that independent.
3. Finally hidden hierarchies, if there is a possibility to know the hierarchical structure of a terrorist network, then it may be easy for intelligence agencies to view the command structure of a network and decide accordingly whom is to be captured, so that the maximum network is disrupted.

In this section we present a theory behind analyzing and destabilizing of terrorist networks. We have implemented all the models discussed in this section in our investigative data mining toolkit known as iMiner.

### 8.1 The Efficiency of a network

Complex networks can be understood from a point of view of efficiency, i.e., networks demonstrating small-world properties are supposed to be very efficient in terms of information propagation across a network [52]. The network efficiency $E(G)$ is a measure to quantify how efficiently the nodes of a network exchange information. To define efficiency of a network $G$, first we calculate the shortest path lengths $d_{ij}$ between the $i^{th}$ and the $j^{th}$ nodes. Let us now suppose that every node sends information along the network, through its links. The efficiency in the communication between the $i^{th}$ node and the $j^{th}$ node is inversely proportional to the shortest distance: when there is no path in the graph between the $i^{th}$ and the $j^{th}$ nodes, we get $d_{ij} = +\infty$ and efficiency becomes zero. Let $N$ be known as the size of the network or the numbers of nodes in the graph, the average efficiency of the graph (network) of $G$ can be defined as:

$$C_{eff} = E(G) = \frac{1}{N(N-1)} \sum_{i \neq j \in G} \frac{1}{d_{ij}} \quad (6)$$

The above formula gives a value of $C_{eff}$ $E$ in the interval of [0, 1].

Note that this measure is different from closeness centrality which is an arithmetic mean of the shortest distances rather than a harmonic mean.

### 8.2 Position Role Centrality

This centrality is a newly introduced measure which highlights a clear distinction between followers and gatekeepers. It depends on the basic definition of efficiency as discussed in equation (6). The efficiency of a network in presence of followers is low, in comparison to their absence in the network. This is because followers are usually less connected nodes and their presence increases the number of low connected nodes in a network, thus decreasing its efficiency.

$$C_{pr} = E(G) - E(G - v_i), i = 1...N \quad (7)$$

where $G - v_i$ indicates the network obtained by deactivating node $v_i$ in the graph $G$. If we plot the values on the graph, the nodes which are plotted below the x-axis are followers, whereas the nodes higher than the remaining nodes with higher values on positive y axis are the gatekeepers.

### 8.3 Dependence Centrality

Dependence centrality [53] represents how much a node is dependent on other nodes in a network. Consider a network representing a symmetrical relation, "communicates with" for a set of nodes. When a pair of nodes (say, $u$ and $v$) is linked by an edge so that they can communicate directly without intermediaries, they are said to be adjacent. Suppose a set of edges links two or more modes ($u$, $v$, $w$) such that node $u$ would like to communicate with $w$, using node $v$. The dependence centrality can discover how many times node $u$ uses node $v$ to reach node $w$ and how many shortest paths node $u$ uses to reach node $w$. There can, of course, be more than one geodesic, linking any pair of nodes.

Let $\zeta_{(u,v)}(w)$ = dependence factor of the node $u$ on node $v$ to reach any other node (*i.e.*, node $w$) in the graph of communication as shown in (8):

$$\zeta_{(u,v)}(w) = \frac{occurence(u,v)}{d(u,v) \times path(u,w)} \qquad (8)$$

where $occurence(u,v)$ = the number of times (shortest paths), the node *u* uses node *v* in the communication with one another, $path(u,w)$ = the number shortest paths between node *u* and node *w*, and $d(u,v)$ = the geodesic distance between node *u* and node *v*.

As mentioned earlier, the dependence centrality of a node represents how much a node is dependent on other nodes. Usually the nodes which are adjacent to a node are always important for that node, as all activities of that node depend on the nodes which are adjacent to it (or directly connected to that node).

Now we define dependence centrality as the degree to which a node, *u*, must depend upon another, *v*, to relay its messages along geodesics to and from all other reachable nodes in the network. Thus, for a network containing *n* nodes, the dependence centrality of *u* on *v* can be found by using equation (9):

$$C_{dep(u,v)} = 1 + \sum_{w=1}^{n} \zeta_{(u,v)}(w), \qquad u \neq v \neq w \qquad (9)$$

We have used 1 in the above formula, because we expect every graph/ network we use is connected.

We can calculate the dependence centrality of each vertex on every other vertex in the network and arrange the results in a matrix $D = [C_{dep(u,v)}]$. The value of the dependence matrix can be normalized by dividing each value with (*n*—1) where *n* represents the total number of nodes in the network.

Each entry in D is an index of the degree to which the node designated by the row of the matrix must depend on the vertex designated by the column to relay messages to and from others. Thus *D* captures the importance of each node as a gatekeeper with respect to each other node—facilating or perhaps inhibiting its communication.

The dependence matrix benefits an analyst by providing an even clearer picture than betweenness or closeness centrality, not only identifying how much a particular node is dependent on others, but also how much others depend on that particular node. Consider, for example, a fictitious small terrorist network as shown in Figure 3. The nodes *a*, *b*, and *c* show a relatively low score based on overall centrality. The (column) sum of each node (*a*, *b*, *c*) is approximately 1, in comparison to the sums of other nodes in the network. The following is the summary of the dependency matrix shown in Table 1.

The *lowest sum of values in a row* points out that the nodes that are most difficult to be deactivated, as its communications are least damaged with the capture of other nodes. These nodes are least dependent on others. Its communications are uniformly distributed. Whereas the *highest sum of values* in a row points out that the nodes that can be easily deactivated. These nodes are mostly dependent on others.

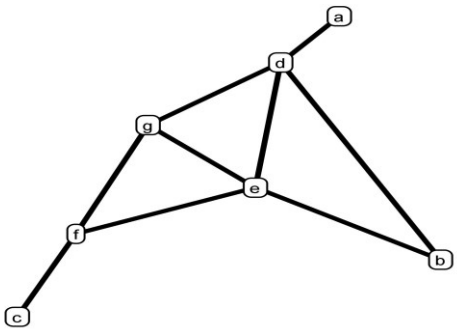



**Figure 3. A fictitious Terrorist Network**

**Table 1. Summary of Dependence Matrix**

| Node | g | e | F | D | b | c | a | Sum |
|---|---|---|---|---|---|---|---|---|
| **g** | | 0.25 | 0.33 | 0.42 | 0.17 | 0.17 | 0.17 | 1.51 |
| **e** | 0.17 | | 0.33 | 0.33 | 0.17 | 0.17 | 0.17 | 1.34 |
| **f** | 0.33 | 0.5 | | 0.25 | 0.17 | 0.17 | 0.17 | 1.59 |
| **d** | 0.33 | 0.33 | 0.25 | | 0.17 | 0.17 | 0.17 | 1.42 |
| **b** | 0.17 | 0.58 | 0.25 | 0.42 | | 0.17 | 0.17 | 1.76 |
| **c** | 0.25 | 0.33 | 1.00 | 0.22 | 0.17 | | 0.17 | 2.19 |
| **a** | 0.25 | 0.25 | 0.22 | 1.00 | 0.17 | 0.17 | | 2.06 |
| **Sum** | 1.5 | 2.24 | 2.38 | 1.64 | 1.02 | 1.02 | 1.02 | |

*The lowest sum of values in a column* tells us that minimum communication takes place through these nodes. The capture of these nodes will be of least damage for a network. Whereas the *largest sum of values in a column* points out the nodes whose capture would be most disruptive to the network.

The dependence matrix benefits the analyst by giving an even clearer picture than betweenness centrality. The matrix not only identifies how much a particular node is dependent on others, but also it discovers how much others depend on that particular node. Thus an analyst may be able to tell the intelligence agencies about the over all picture of the nodes in a network: keeping in view the strength and weaknesses of the node being considered for capture.

Looking at the values of betweenness, we note that the betweenness centrality for the nodes *a* and *b*, i.e., $C_B(a)$ and $C_B(b)$, is zero, as these nodes participate least in the paths of communication. One may only deduce that other nodes in the network are not much dependent on these nodes for communication.

If we view the dependence matrix which is tabulated in Table 1, the row of node *a* tells us how much node *a* depends on others. It is crystal clear that node *a* mostly depends on the node *d* (by removing node *d*, node *a* will automatically be isolated). If we analyze the row of node *b*, the row is more uniformly distributed as compared to the row of node *a*. The node *b* mostly depends on node *e* (which has the highest value of 0.58).

**Table 2 . Betweenness Centrality of the network shown in Figure 3**

| Node | g | e | f | d | b | C | a |
|---|---|---|---|---|---|---|---|
| **Betweenness** | 0.13 | 0.30 | 0.33 | 0.37 | 0 | 0 | 0 |

The node *b* also has some other alternatives, for example, it also depends on node *d* (it has value 0.42). As these two nodes are not frequently in communication as resulting from betweenness, there is significant difference as indicated by the dependence matrix. It should be noted that if intelligence agencies desire to know which node should be removed or eradicated (either node *a* or node *b*), the analyst by looking at the dependence centrality matrix can assist in advising to capture node *b* in comparison to node *a*, because node *b*'s capture will cause more harm than node *a*. These points have significant effects, which are not indicated by betweenness centrality.

We applied the above mentioned measures (described in subsection 8.1-8.2) in the network of terrorists involved in the tragic events of September 11, 2001 (as shown in Figure 1). The results are depicted in Figure 4. The results show that node 33 (Mohamed Atta) as a key player in the plot. The position role centrality of this node is higher than all nodes which proves that this node played an important role in the plot and worked as a gatekeeper and that removing with this node the efficiency of the graph is decreased from 0.395 to 0.32. This clearly identifies the importance of this node in the network.

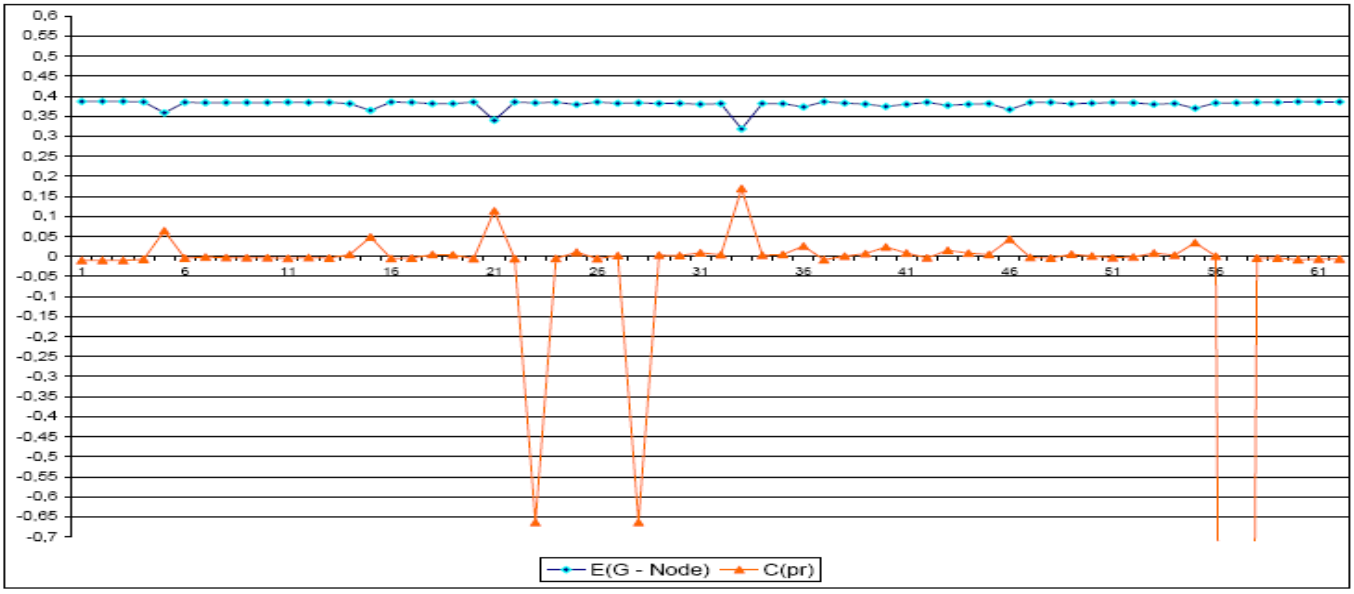



**Figure 4. The efficiency of the original network *E* (*G*) = 0.395. The removed node is shown on x-axis. The efficiency of the graph once the node is removed is shown as *E* (*G* – $v_i$); while importance of node. The newly introduced measure position role centrality is shown as C(pr).**

## 8.4 Detecting Hidden Hierarchy

Hierarchy, as one common feature of many real world networks, attracts special attention in recent years [55] [56] [57] [58]. Hierarchy is one of the key aspects of a theoretical model to capture statistical characteristics of terrorist networks.

In the literature, several concepts are proposed to measure the hierarchy in a network, such as the hierarchical path [57], the scaling law for the clustering coefficients of nodes in a network [55], etc. These measures can tell us the existence and extent of hierarchy in a network. We address herein another problem of how to the construct hidden hierarchy of terrorist networks (which are known as horizontal networks) [54].

Discovering hierarchy from a terrorist network is *a process of comparing the different centrality values of different nodes to identify which node is more powerful, influential or worthy to neutralize than others*.

From the above definition, it is clear that we may use different centrality measures for finding the corresponding hierarchical view of a graph/ network. Every centrality measure is associated with a particular meaning, for example, degree tells us about powerful nodes, and betweenness gives us an idea about gatekeepers. Similarly these measures can be used to build hierarchies to detect the organizational view of a corresponding terrorist network/ organization.

Currently, experts have agreed that real destabilization is about isolating leaders from enough followers, thus disabling them from executing any terrorism plans. This idea has certainly been a central motive for investigative tools. The algorithms for detecting hidden hierarchy from terrorist networks can be found in the article [2].

## 9. THE SOFTWARE PROTOTYPE

iMiner is an experimental system, which provides answers to the following questions.

- Who is likely to be an important person (node) in the network?
- Why s/he is important?
- Which terrorist is highly/ less connected?
- What are the various position roles in the network?
- Is there any command structure in the network?
- Is it possible to construct a hidden hierarchy of the non-hierarchical networks?
- Which nodes represent key players?
- What is the efficiency of the network?
- How much is the efficiency of the network affected by eradicating one or more terrorists?
- How can the law enforcement agencies use (often incomplete and faulty) network data to disrupt and destabilize terrorist networks?

The system architecture of the investigative data mining toolkit (iMiner prototype) is depicted in Figure 5. The first stage of network analysis development is intended to automatically identify the strongest association paths, or geodesics, between two or more network members. In practice, such a task often entails intelligence officials to manually explore links and try to find association paths that might be useful for generating investigative leads.

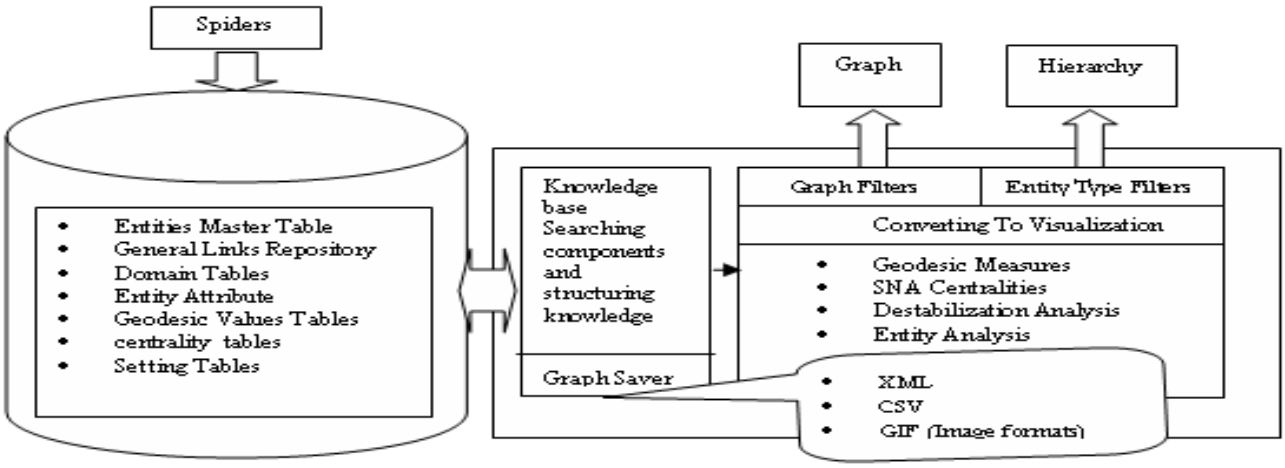



**Figure 5. The System Architecture of the Investigative Data Mining Toolkit**

The iMiner supports domain-based study of a network system. The domain can be similar to a particular case study (for example the Bali bombing or 9/11 case studies) with resizable boundaries. An investigating officer may expand the domain by including additional entities or reduce it by excluding less important or partially involved entities. The main advantage of this type of approach is to isolate a piece of data in which an investigating officer is interested, and to concentrate analysis activities only on those entities.

## 10 CASE STUDIES

In this section, we present two case studies of terrorist attacks that have occurred or been planned in the past to test the algorithms and models recently published [54].

### 10.1 Case 1: Bali Night Club Bombing Network

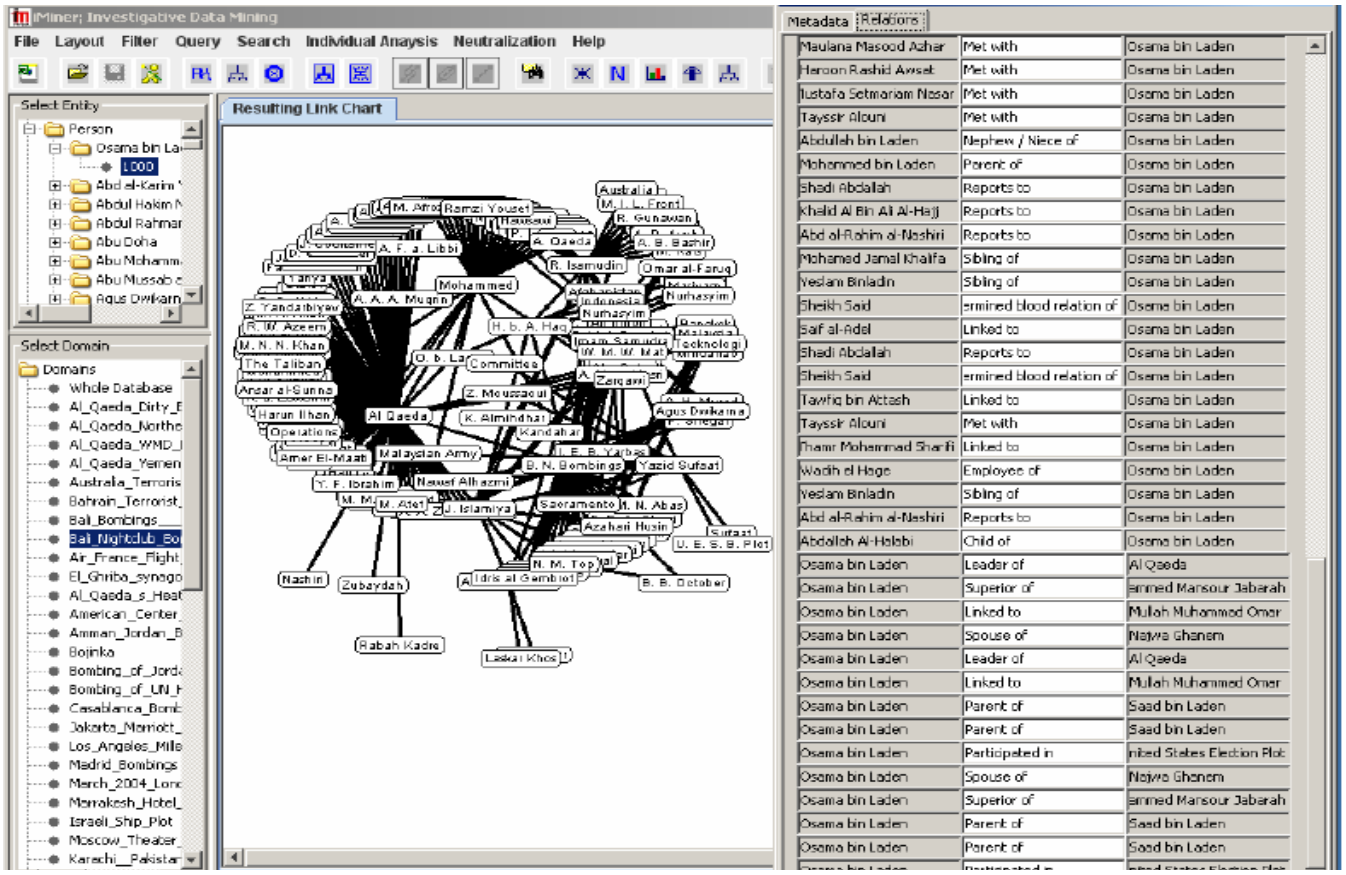



**Figure 6. Bali Night Club Bombing Terrorists Network**

The 2002 Bali bombing[8] occurred on October 12, 2002 in the tourist district of Kuta on the Indonesian island of Bali. The attack was the deadliest act of terrorism in the history of Indonesia, killing 202 people, 164 of whom were foreign nationals (including 88 Australians), and 38 Indonesian citizens. A further 209 were injured.

The attack involved the detonation of three bombs: a backpack-mounted device carried by a suicide bomber; a large car bomb, both of which were detonated in or near popular nightclubs in Kuta; and a third much smaller device detonated outside the United States consulate in Denpasar, causing only minor damage.

Various members of Jemaah Islamiyah (JI), a violent group, were convicted in relation to the bombings. Riduan Isamuddin, generally known as *Hambali* and the suspected former operational leader of Jemaah Islamiyah, is in U.S. custody in an undisclosed location, and has not been charged in relation to the bombing or any other crime. Khalid Sheikh Mehmood[9], in March 2007, claimed that he was the master mind of the 2002 Bali bombing attacks. Huda-bin-Abdul Haq[10] alias Mukhlas, a senior and influential JI leader and head of regional network that includes Sumatra, Singapore, Malaysia and Southern Thailand. He was the co-founder of the religious school in Malaysia, which functioned as training ground for JI operatives. Directly involved in Bali bombings and helped in money transaction and explosives procurements.

We have collected a dataset of the Bali bombing terrorists' network which is shown in Figure 6.

Inspection of this network by standard centrality measures of network structure reveals first its low connectedness. A member of this network holds only *2.87 connections* with others, which means that average members were rather isolated from the rest of the network. The density of this network is only 0.03, meaning that only 3% of all possible connections in the network really exist.

In spite of low connectedness, however, the nodes of this network are relatively close. We found that members in this network were very close to their leaders as shown in Table 3. The terrorists in the this network are on average, only 2 steps away from *Khalid Sheikh Mohammed*, 2.3 steps away from *Riduan Isamudin*, 2.4 steps away from *Osama bin Laden*, and 2.5 steps away from *Huda bin Abdul Haq*, meaning that these leaders' commands can reach an arbitrary member through only two mediators The average betweenness of this network is also very low, indicating high average redundancy.

[8] http://en.wikipedia.org/wiki/2002_Bali_bombings

[9] http://en.wikipedia.org/wiki/Khalid_Shaikh_Mohammed

[10] http://www.ipcs.org/agdb05-JI.pdf

**Table 3. The Characteristics of Key Players of 2002 Bali Bombing Terrorist Network**

| Key Players | Drop of efficiency | PR Centrality | Degree | Closeness | Betweenness | EVC | Av. Path Length |
|---|---|---|---|---|---|---|---|
| **Huda bin Abdul Haq** | 0.374 | 0.004 | 0.097 | 0.405 | 4.45E-4 | 121 | 2.5 |
| **Osama bin Laden** | 0.375 | 0.003 | 0.024 | 0.415 | 1.141E-4 | 91.0 | 2.4 |
| **Riduan Isamudin** | 0.333 | 0.114 | 0.185 | 0.43 | 0.004 | 119 | 2.3 |
| **Khalid Shaikh Mohammed** | 0.306 | 0.185 | 0.218 | 0.5 | 0.006 | 121 | 2.0 |

The efficiency of the network to communicate with its members is 0.38. The drop of efficiency in the case of capturing member by intelligence agencies or law enforcement is shown in Table 3.

The position role centrality of the key conspirator of the plot, *Khalid Sheikh Mohammad* is higher than other leaders, and shows his conspiracy position in the network.

The degree distribution of the nodes in the network (as shown in Table 4) is particularly appealing. The degree of nodes are exponentially distributed: the degree of most of the nodes is small, while very few nodes have high degree. This property characterises the so called scale- free networks.

Using algorithms for detecting hidden hierarchy from horizontal networks [54], we succeeded in constructing the hidden hierarchy of the Bali bombing terrorist attack using algorithms as discussed above, as shown in Figure 7. The H. B. A. Haq node and its descendants form a group (this cluster acted as executive cluster), while the cluster of Khalid Shaikh Mohammed and his affiliates was well known as a strategic cluster, whereas the cluster of R.Isamudin (known as Hambali) and his associates is known as a tactical / logistical cluster . The accuracy of the software can be determined by the fact that all of H. B. A. Haq, Khalid Shaikh Mohammed and R. Isamudin were key players. H. B. A. Haq was termed as a potential leader while Khalid Sheikh Mohammed was the key conspirator.

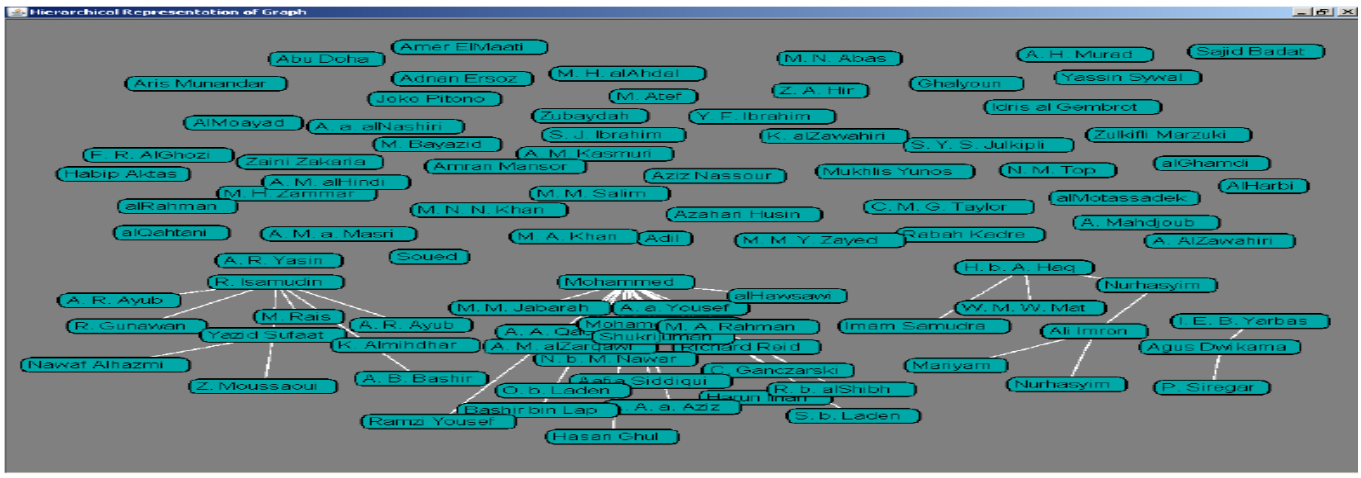



**Figure 7. Hidden Hierarchy in the Bali Bombing Terrorists Network**

**Table 4. Degree Distribution for the 2002 Bali Bombing Terrorist Network**

| Degree | Nodes |
|---|---|
| 1 | 80 |
| 2 | 18 |
| 3-6 | 13 |
| 7-9 | 7 |
| 11-12 | 3 |
| 23 | 1 |

| | |
|---|---|
| 27 | 1 |
| 29 | 1 |
| 52 | 1 |

## 10.2 Case 2: September 11, 2001 Terrorist Plot

The September 11, 2001 attacks (often referred to as 9/11—pronounced "nine eleven") consisted of a series of coordinated terrorist suicide attacks upon the United States, predominantly targeting civilians, carried out on Tuesday, September 11, 2001.

That morning, 19 terrorists affiliated with al-Qaeda hijacked four commercial passenger jet airliners. Each team of hijackers included a trained pilot. Two aircraft (United Airlines Flight 175 and American Airlines Flight 11) crashed into the World Trade Center in New York City, one plane into each tower (WTC 1 and WTC 2). Both towers collapsed within two hours, followed by WTC 7 later that day. The pilot of the third team crashed American Airlines Flight 77 into the Pentagon in Arlington County, Virginia. Passengers and members of the flight crew on the fourth aircraft (United Airlines Flight 93) attempted to retake control of their plane from the hijackers; that plane crashed into a field near the town of Shanksville in rural Somerset County, Pennsylvania. As well as the 19 hijackers, a confirmed 2,973 people died and another 24 are missing but presumed dead as a result of these attacks..

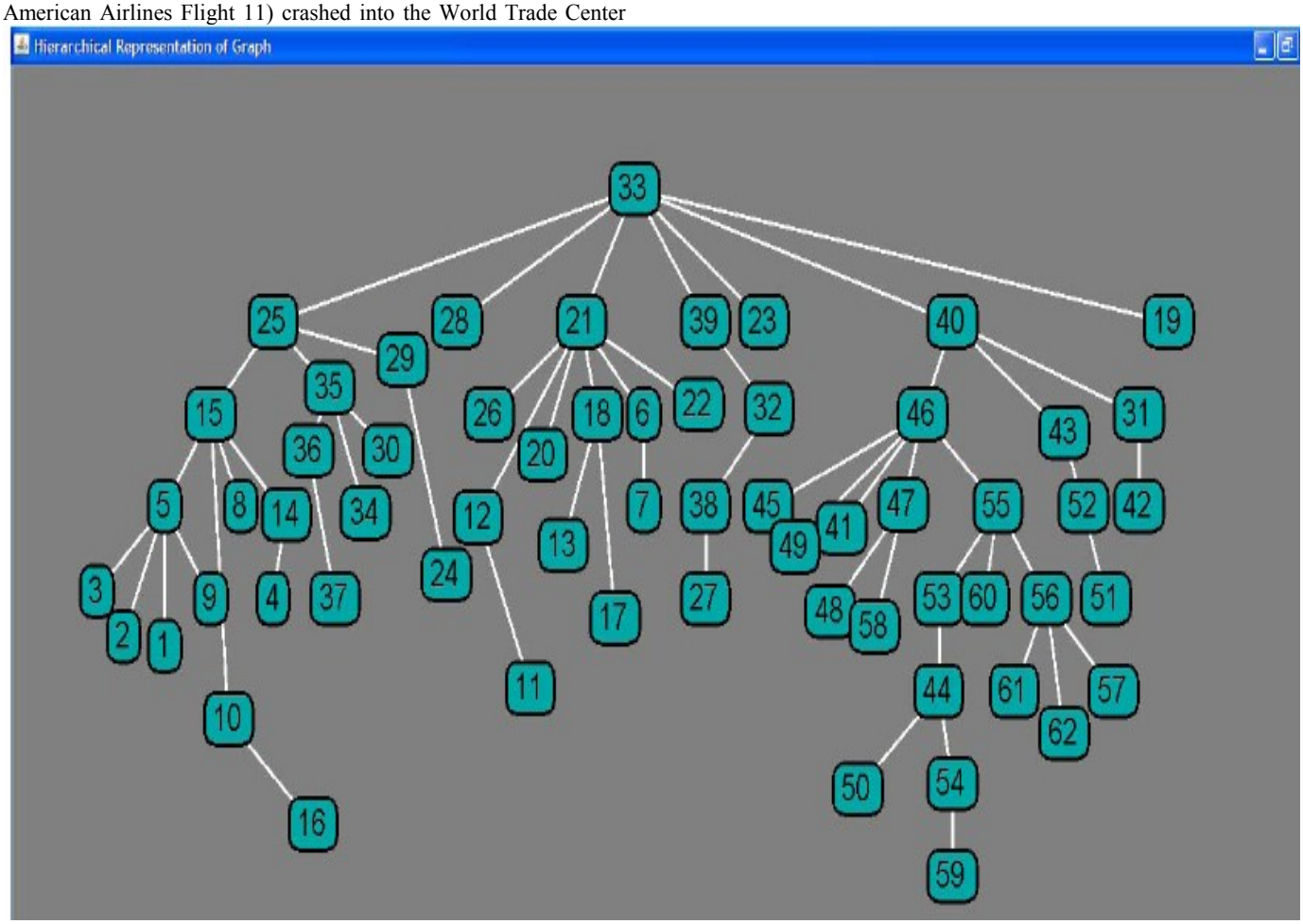



**Figure 8. The hierarchy clearly suggests that Muhammad Atta (node # 33) was the most powerful person (leader) of the plot.**

The renowned Social Network Analyst Valdis Krebs [13] mapped the network of 9/11 (hijackers and their affiliates) as shown in Figure 1. Using the algorithms for detecting hidden hierarchy of non-hierarchical terrorist networks [2], we tested the network of

terrorists involved in 9/11 tragic events and results are depicted in Figure 8.

# 11 CONCLUSIONS

In this paper, we presented an overview of an investigative data mining toolkit (iMiner software prototype) which we have developed for undertaking analysis of terrorist networks. In general investigative data mining has been shown to be a promising and potentially powerful area of research. The paper presented interesting patterns gleaned from the data. We discussed three innovative ideas of our research which were already published in [54] and featured in Government Computer News [59]. The mathematical models and algorithms discussed in the paper are implemented in the prototype. The *iMiner* demonstrates key capabilities and concepts of a terrorist network analysis toolkit. Using the toolkit investigating officials can predict overall functionality of the network along with key players. Thus the counterterrorism strategy can be designed keeping in the mind that destabilization not only means disconnecting the dots (nodes) but disconnecting those key players from the peripheries by which a maximum network disruption can be achieved. Investigative data mining can be used to understand terrorist networks, and we are of the view that an investigative data mining tool like iMiner, improves on the traditional analysis of networks with large volumes of data and investigative data mining could reduce the consequent overload on analysts. The results presented in this paper are our findings based on a limited exercise in exploring the utility of investigative data mining in analyzing terrorist networks. This tool may also be used for law enforcement agencies for destabilizing of terrorist networks for capturing the key nodes. The intelligence agencies may also evaluate the efficiency of the networks in the case of the capture of a particular node. Further real-time or near real-time information from a multiplicity of databases could have the potential to generate early warning signals of utility in detecting and deterring terrorist attacks. It is necessary, of course, to have 'experts' in the loop. This analysis has provided a substantive and in-depth analysis of terrorist networks. Furthermore this analysis has provided a richer and deeper understanding and insight into terrorist networks and has provided approaches to destabilize the networks.

In this paper we presented the system architecture of our software prototype and the process by which we harvested data from the Web and stored it in the knowledge base. The focus of the knowledge base we have developed is the agglomeration of publicly available data and integration of the datasets with the software prototype in order to investigate interesting patterns.

# 12 FUTURE RECOMMENDATIONS

This research sought to develop new theory and measures / mathematical models and practical algorithms for analyzing, visualizing and destabilizing terrorist networks. These measures could be useful for law enforcement and intelligence agencies to disrupt the effective operation and growth of these networks or destroy some terrorist cells entirely. Although these adversaries can be affected in a number of ways, this research focuses upon capturing / eradicating a terrorist organization's most influential persons or finding susceptible points of entry and conveying information or influence that contribute to winning the war against terrorism. There remain a number of research opportunities in this new area of investigative data mining. Refinement in the measures we presented in this paper to detect different roles in terrorist networks is one of them. It is possible to find models to detect the following roles in terrorist networks as proposed by Williams [60]:

1) **Organizers** are the core ensuring a network's direction. It is they who determine the scale and scope of activities, as well as the guidance and impetus necessary for performing those activities.

2) **Insulators** are individuals or groups charged with insulating a core from dangers posed by infiltration and compromise situations to which it is exposed. These actors transmit directives or guidance from a core to a periphery. They also ensure that the flow of communication from a periphery in no way compromises a core.

3) **Communicators** are individuals who ensure that communication flows effectively from one actor to another throughout the network. Unlike insulators, communicators must gather feedback regarding directives that they transmit to other actors in a network. Williams claims that there can be conflicts between those who act as insulators and those who act as communicators, or that the same individuals may assume both roles simultaneously to avoid these conflicts.

4) **Guardians** ensure network security and take necessary measures to minimize its vulnerability to infiltrations or external attack. Their role also consists in watching over recruitment to a network and ensuring the loyalty of recruits through a variety of ritual oaths and latent coercion directed against new members and their families. Guardians seek to prevent defections from the network actors and to minimize damages when defections occur.

5) **Extenders** extend the network by recruiting new members and also by negotiating collaboration with other networks and encouraging collaboration with the business sector, government and justice. Various tactics are used to this end. They range from voluntary recruitment through bribery and corruption to involuntary recruitment through coercion, occasionally supported by incentives and rewards.

6) **Monitors** are dedicated to the network's effectiveness. Their responsibilities consist in providing information to organizers regarding weaknesses and problems within the network so that the organizers can resolve them. Monitors ensure that the network is able to adjust to new circumstances and maintain the high degree of flexibility that is necessary to circumvent law enforcement.

7) **Crossovers** are part of a terrorist network, but continue to work in legal institutions, whether governmental, financial or commercial. As such, these individuals provide invaluable information and contribute to the protection of a network.

Investigative data mining can also be used to understand the psychological effects of terrorism. One of the main effects of terrorism is fear, which is spread through network structures such as the media, the Internet, and personal relationships. For example, the number of ties an individual has to victims of

terrorism may impact the individual's perception of the risk of terrorism.

We would like to see further research on network structure evolution. It would be interesting to compare structures of multiple terrorist networks to see how they evolve over time. The network structure may impact the ability of an organization to endure over the years and to complete attacks. It is important for intelligence agencies to understand how to split up a network; they could potentially exploit the destabilizing techniques discussed in this research work including small world topology by eliminating weak ties in order to isolate the network and diminish its reach and power. The removal of individuals in key network regions may be even more important than attacking the traditional leaders of a group. It will also be helpful for intelligence agencies to look at how tightly knitted individuals are in a network as discussed in this Section of this work. New algorithms may need to be developed considering the social cohesion between terrorist groups.

We mentioned in this article, that we harvested data from the Web and stored it in our knowledge base. A fuzzy knowledge base should be developed in this very important field. Semantic Web languages, such as RDF, RDF Schema and OWL can be considered for this purpose. The metadata used in homeland security projects are fuzzy by nature, and the semantic web is capable of representing fuzzy data.

Considering the nature of this war, understanding the structure of these networks is paramount. No longer fitting the traditional paradigm of combat between great armies, this war involves not only defeating the individuals actively threatening our National Security, but also alleviating the environments that nurture the development and continuity of such groups. In order to accomplish this, analysts must at a minimum (1) improve the understanding of why people would undertake such activities and mentally prepare to be used as human bombs; (2) identify vulnerabilities existing within these networks and how to exploit them; and, (3) determine what consequences may follow an operation to minimize the likelihood that actions executed unintentionally contribute to the environments that promote extremism.

## REFERENCES


[1] Memon Nasrullah and Henrik Legind Larsen. (2006) *Practical Approaches for Analysis, Visualization and Destabilizing Terrorist Networks*. In the proceedings of ARES 2006: The First International Conference on Availability, Reliability and Security, Vienna, Austria, IEEE Computer Society, pp. 906-913.

[2] Memon Nasrullah and Larsen Henrik Legind. (2006) *Practical Algorithms of Destabilizing Terrorist Networks*. In the proceedings of IEEE Intelligence Security Conference, San Diego, Lecture Notes in Computer Science, *Springer-Verlag*, Vol. 3976: pp. 398-411 (2006)

[3] Memon Nasrullah and Larsen Henrik Legind. (2006) Detecting Terrorist Activity Patterns using Investigative Data Mining Tool. *International Journal of Knowledge and System Sciences*, Vol. **3**, No. 01, pp. 43-52.

[4] Memon Nasrullah, Qureshi A. R.. (2005). Destabilizing Terrorist Networks. *In WSEAS Transactions on Computers*, Issue 11, Vol. **4,** pp.1649-1656

[5] Taskar Ben, Pieter Abbeel, Ming-FaiWong, and Daphne Koller. (2003) *Label and Link Prediction in Relational Data*, in IJCAI Workshop on Learning Statistical Models from Relational Data.

http://kdl.cs.umass.edu/srl2003_upload/files/taskar-paper.pdf

[6] M. Barlow, J. Galloway, and H. Abbass. (2002). *Mining Evolution through Visualization*. In Proceedings of Workshop on Beyond Fitness: Visualization Evolution at the 8th International Conference on the Simulation and Synthesis of Living Systems,

http://www.alife.org/alife8/workshops/15.pdf

[7] Q&A with Professor Karen Stephenson, April 18, 2006

[8] DeRosa M., (2004), *Data Mining and Data Analysis for Counterterrorism*, CSIS Report.

[9] Memon Nasrullah (2007). A First Look on iMiner's Knowledge Base and Detecting Hidden Hierarchy of Riyadh Bombing Terrorist Network. In: *Proceedings of International Conference on Data Mining Applications,* Hong Kong, 21-23 March, 2007.

[10] Memon N., Larsen H. L.. (2006). Structural Analysis and Mathematical Methods for Destabilizing Terrorist Networks. In: *Proceedings of the Second International Conference on Advanced Data Mining Applications, (ADMA 2006)*. Springer Verlag Lecture Notes in Artificial Intelligence (LNAI 4093), 2006. p. 1037-1048

[11] Degenne, A., Forse, M. (1999). *Introducing Social Networks*. London: Sage Publications.

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

[13] Krebs, V. E. (2202) Mapping Network of Terrorist Cells. *Connections* 24(3): 43-52

[14] Stewart, T. (2001). *Six Degrees of Mohamed Atta.* Accessed on 24/1/2006

*http://money.cnn.com/magazines/business2*

[15] Mena, J. (2003). *Investigative Data Mining for Security and Criminal Detection*. Butterworth Heinemann.

[16] Hougham, V. (2005). *Sociological Skills Used in the Capture of Saddam-Hussein.*
*http://www.asanet.org/footnotes/julyaugust05/fn3.html*
Accessed on 22/2/2005

[17] Memon, N.; Kristoffersen, K. C.; Hicks, D. L.; Larsen, H. L. (2007). Detecting Critical Regions in Covert Networks: A Case Study of 9/11 Terrorists Network. In: *Proceedings of International Conference on Availability, Reliability, and Security 2007, Vienna, Austria, March 10-13, 2007*.

[18] Memon N., Larsen H. L. (2007). *Detecting Hidden Hierarchy from Terrorist Networks. Mathematical Models for Counterterrorism*, Springer (In Press).

[19] West B. D. (2001). Introduction to Graph Theory (Second Edition) *Prentice Hall*.

[20] Klerks, P. (2002) The Network Paradigm Applied to Criminal Organisations: Theoretical nitpicking or a relevant doctrine for investigators? Recent developments in the Netherlands. *Connections* 24(3): 53-65.

[21] Fijnaut Cyrille J.C.F., Frank Bovenkerk, Gerben Bruinsma. (1998) Organized Crime in Netherlands. The *Hauge Kulwar Law International*.

[22] Freeman, L.C. (1978) Centrality in Social Networks: I. Conceptual clarification. *Social Networks*, **1**:215-39

[23] Bonacich, P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*. **2**:113-120 (1972)

[24] Frantz T. and K. M. Carley. (2005). *Relating network topology to the robustness of centrality measures*. Technical Report CMU-ISRI-05-117, School of Computer Science, Canregie Mellon University.

[25] Press William, S. Teukolsky, W. Vetterling, and B. Flannery, (2002) Numerical Recipes: The Art of Scientific Computing, *Cambridge University Press*.

[26] Hotelling Harold (1936). Simplified calculation of principal components, Psychometrika, Vol 1, pp. 27-35.

[27] Allen, C. *The Dunbar Number as a Limit to Group Sizes* (2004). Retrieved May 31, 2006, from http://www.lifewithalacrity.com/2004/03/the_dunbar_numb.html

[28] Watts, D.J.: *Six Degrees – The Science of a Connected Age*. W.W. Norton & Company, New York, 2003.

[29] Kreisler, H.: International Relations in Information Age: Conversation with John Arquilla, Professor at University of California, Berkeley 2003. Retrieved on June 01, 2006 from

http://globetrotter.berkeley.edu/people3/Arquilla/arquilla-con0.html

[30] Al Qaeda Training Manual. Retrieved on June 01, 2006 from

http://www.fas.org/irp/world/para/manualpart1.html

[31] Walter R. Harper and Douglas H. Harris. (1975). The Application of Link Analysis to Police Intelligence," *Human Factors* 17: 157-164.

[32] Karl M. Van Meter, (2001). "Terrorists/Liberators: Researching and Dealing with adversary social networks," *Connections* **24 (**3): 66-78.

[33] i2: Analyst's Notebook (Accessed on November 14, 2006. http://www.i2.co.uk/Products/Analysts_Notebook/default.asp

[34] Carley, Kathleen Ju-Sung Lee, David Krackhardt. Destabilizing Networks, *Connections* **24**(3):31-34. (2001)

[35] Farley, J.D. Breaking Al Qaeda Cells: A Mathematical Analysis of Counterterrorism Operations (A Guide for Risk Assessment and Decision Making). *Studies in Conflict and Terrorism*, 26: 399-411 (2003)

[36] McAndrew, D.: *Structural Analysis of Criminal Networks. Social Psychology of Crime: Groups, Teams, and Networks, Offender Profiling Series*, III. L. Allison. Dartmouth, Aldershot (1999)

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

[38] Sageman, M.: Understanding Terror Networks. Pennsylvania, University of Pennsylvania Press (2004)

[39] J. Raab and H. B. Milward. Dark Networks as Problems. *Public Administration Research and Theory*, 13(4): 413-439 (2003)

[40] P. Svenson, C. Mårtenson, C. Carling. *Complex Networks: Models and Dynamics*, FOI-R-1766-SE (2005)

[41] A. Clauset, M. Newman, C. Moore. Finding Community Structure in Very Large Networks, *Physical Review* E70, 066111 (2004)

[42] M. Newman. A Measure of Betweenness Centrality based on Random Walks. *Social Networks* 27: 39-54 (2005)

[43] Albert R. and A. L. Barabasi. (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(47).

[44] Dorogovtsev, S. N., & Mendes, J. F. F. (2002). Evolution of networks. *Advances in Physics*, *51*, 1079–1187.

[45] Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review*, *45*, 167–256.

[46] Bollobas, B. (1985). *Random Graphs*. London: Academic Press.

[47] Watts, D. J. and S. H. Strogatz.(1998). Collective dynamics of "small-world" networks. *Nature*, 393:440–442.

[48] Barabasi A. L. and R. Albert. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

[49] Albert, R., Jeong, H., & Barabasi, A.L. (2000). Error and Attack Tolerance of Complex Networks. *Nature*, *406*, 378–382.

[50] Chen, H. (2006). Intelligence Security Informatics for Information Sharing and Data Mining, *Integration Series in Information Systems, Vol. 10*., Springer US, 85-104.

[51] Krebs, V., (2005), *Organizational Hierarchy: Adapting Old Structures to New Challenges*. Available Online, Accessed on 25/3/2007. http://www.orgnet.com/orgchart.html

[52] Latora, V., Massimo Marchiori, How Science of Complex Networks can help in developing Strategy against Terrorism, *Chaos, Solitons and Fractals* 20, 69-75 (2004)

53] Memon Nasrullah, Hicks David Lane, Larsen Henrik Legind. (2007). How Investigative Data Mining Can Help Intelligence Agencies to Detect Dependence of Nodes in Terrorist Networks. In: R. Alhajj et al. (Eds.): ADMA 2007, LNAI 4632, pp. 430–441, 2007.

[54] Memon Nasrullah. (2007). Investigative Data Mining: Mathematical Models for Analyzing, Visualizing and Destabilizing Terrorist Networks. PhD Thesis. Aalborg University Denmark. ISBN: 978-87-7606-020-6.

[55] Ravasz, E. A. L., Barabasi, (2003). Hierarchical Organization in Complex Networks. *Physical Review*, E **67**, 026112.

[56] Costa, L. D. F. (2004). Hierarchical Backbone of Complex Networks. *Phys. Rev. Lett*. **93**, 098702.

[57] Trusina A.,Sergei Maslov, Petter Minnhagen, and Kim Sneppen. (2004). Hierarchy Measures in Complex Networks. *Phys. Rev. Lett*. **92**, 178702.

[58] Variano, E. A. *et al* (2004). Networks Dynamics and Modularity. *Phys. Rev. Lett*. **92**, 188701.

[59] Jackson Joab, NSA and Social Networking: TRENDS & TECHNOLOGIES that affect the way government does IT, Government Computer News, 24th July 2006. Available online

http://www.gcn.com/print/25_21/41403-1.html

[60] Williams, P. (2001). "*Transnational Criminal Networks*" in J. Arquilla and D. Ronfeldt (eds), Networks and Netwars: The Future of Terror, Crime and Militancy. Santa Monica: Rand Corporation, pp. 61–97.

# Smartmatch

Brendan Kitts and Gang Wu
Microsoft Corporation
One Microsoft Way, Redmond, WA. USA
{bkitts,simonwu}@microsoft.com

Paid Search auctions work by defining ads which are returned based on a user query. For instance if a user types in “pink carnations”, an advertiser who is bidding on “pink carnations” will have their ad returned.

Unfortunately there is tremendous variation in search phrases typed by users and a variety of phrasings may mean the same thing. This has resulted in hundreds of thousands of keywords being selected by advertisers. For instance “pink carnations”, “pink bouquets”, “pink roses”, “flower pink”, “pink flwr”, “pink flower for decoration”, “pink boutonnière”, “pink corsage”, “pastel carnations” may all be useful to an advertiser who is bidding on “pink flowers” – yet would not match unless the advertiser has selected all of those variations.

The objective of the Smartmatch project was to augment the standard exact, phrase and broad string matching that is employed by standard search engines, with “smart” matches based on analysis of user intent. One could think of this as “spell correction” but allowing for other common phrasings, contractions, acronyms, mutations, based on observed human re-write behavior.

Achieving good keyword expansions is difficult – only the most relevant expansions could be used and poor rewrites resulted in user dissatisfaction. As a result Smartmatch was allowed multiple research teams to develop algorithms and have them tested experimentally This was possible because keyword expansions were represented as (keyword,expansion,algorithm) in a large lookup table. At run-time a different algorithm was randomly selected, presented to the user, and then the resulting impression and click recorded along with the algorithm responsible for the expansion. The representation of rewrites meant it was extremely easy to develop a new algorithm. Teams could even combine the best performing rewrites into an optimized/combination algorithm.

Smartmatch was conceived in January 2006, piloted in September 2006 on 5% of US traffic, and rolled out around December 2006 on 95% of traffic. Control groups are maintained to provide information on whether algorithms need to be retrained. Since deployment over 55 algorithms from 9 research teams have been tested, and the best algorithm used for the bulk of expansions. Smartmatch has improved all business measures monitored by adCenter including human relevance of ads, advertiser conversion rate, search engine clickthrough rate, and is now a model program for similar data mining initiatives underway at Microsoft.

# Finding Duplicates in the 2010 Census

Edward H. Porter and Michael Ikeda
U.S. Census Bureau
4600 Silver Hill Road
Washington, DC 20233
Edward.h.porter@census.gov

The United States Census Bureau counts people as living in two separate living arrangements called "housing units" and "group quarters." Housing Units consist of homes, apartments and any room intended as a separate living quarters, but not transient quarters. All other types of living arrangements are classified as Group Quarters, such as dormitories, military bases, nursing homes and prisons. Often times, people may have been counted in both types units. Other times an individual may be counted in housing units in different locations.

This talk presents techniques used at the Census Bureau in finding duplicates. Historically, Record Linkage programs such as the Matcher and BigMatch help us find duplicates. Data Elements are first compared exactly by a procedure known as blocking—for example, in the first blocking pass only those records whose phone numbers match exactly are compared. Then Data elements, such as first name, surname and date of birth, are assigned scores for agreement and disagreement. If the cumulative score exceeds a certain threshold the records are considered to be duplicates. If the score does not exceed the threshold the records may be set aside for review-which can be expensive-or declared a non-duplicates.

This talk will also discuss how cultural and procedural changes will aid or hinder the finding of duplicates and how those changes will affect is in the modeling. With expanded computing power that allows us to compare millions of pairs of data per second plus subsequent mining and modeling on elements of the data will aid the Census Bureau in identifying false positive matches and will aid the Bureau in finding duplicates in the 2010 Census. This of course will produce a more accurate Census of the Unites States in 2010.

# Any-time clustering of high frequency news streams

Fabian Moerchen
Siemens Corporate Research
755 College Road East
08540 Princeton, NJ, USA
001-609-734-3529
fabian.moerchen@siemens.com

Klaus Brinker
Siemens Corporate Research
755 College Road East
08540 Princeton, NJ, USA
001-609-734-3312
klaus.brinker@gmail.com

Claus Neubauer
Siemens Corporate Research
755 College Road East
08540 Princeton, NJ, USA
001-609-734-6567
claus.neubauer@siemens.com

## ABSTRACT

We describe a large scale system for clustering a stream of news articles that was developed as part of the Geospace & Media Tool (GMT). The GMT integrates the news feed with geospatial, census, and human network information to provide a research tool for members of Congress and their staffs. News articles covering the same event are summarized for the user through the clustering component. The clustering result is available to the user at any time without additional on-demand clustering steps. The documents are grouped into clusters on-the-fly without any assumptions on the number of clusters and without retrieving previous documents. High efficiency is achieved by utilizing locality sensitive hashing (LSH) as a means to determine a small set of candidate clusters for each document. This way a large number of clusters can be considered while keeping the number of expensive document to cluster comparisons low. Our experiments with the system reveal interesting aspects of large-scale text processing in general and news clustering in particular. We demonstrate how the LSH based approximation achieves a large speedup at the cost of only few and small errors. On a high-frequency benchmark data set a clustering quality comparable to one of the best non-streaming document clustering algorithms is obtained.

## Categories and Subject Descriptors

*H.4* **[Information Systems Applications]***: Miscellaneous*

## General Terms

Algorithms, Performance.

## Keywords

text mining, clustering, data streams.

## 14. INTRODUCTION

The automated processing of large amounts of text is an important tool in knowledge management [9, 18]. Classification and clustering of text documents can help to structure a document collection and make it more accessible. At a coarse level an assignment to topics or categories can help to navigate a corpus. If some of the documents are already assigned to topics, a supervised classification approach can help to label the remaining documents and any new documents [34, 43, 29].

Otherwise unsupervised clustering can help to discover the underlying topical structure of existing text collections [15, 35, 52]. In many applications the document collection is dynamic [5, 48, 30] in the sense that new documents are continuously being added to the database and need to be processed. In this scenario both supervised and unsupervised methods need to be able to cope with new topics in the text that do not fit the previous classification or cluster model [46, 20].

Consider news articles from newspapers and news agencies that form a stream of text documents. For a given event different news sources publish similar articles and even the same news source covers a developing story over several days. Clustering can be used to aggregate the news articles that cover the same story and offer the user a better overview of the current events [5, 48, 30]. For each cluster a summary can be generated from using headlines and content of representative articles and the most relevant keywords. Additional articles from the cluster can be displayed on demand to provide more details. A well known implementation of such a news aggregation system is Google News (http://news.google.com).

We describe a large scale system for clustering a stream of news articles that is a core component of Geospace & Media Tool (GMT) developed in cooperation with the Parsons Institute for Information Mapping (PIIM), The New School, NY. The GMT integrates the news feed with geospatial, census, and human network information to provide a research tool for the members of Congress and their staffs. We describe the requirements of the text clustering component within this application and describe an efficient solution. We report the results of extensive experiments regarding the trade-off between speed and quality, the handling of dynamic content, and the merits of using meta-data to improve the clustering quality. The same system could be applied to cluster Blog entries, emails, customer service requests, medical reports, and similar potentially high rate text streams. Section 2 briefly describes the different components and use cases of the GMT and translates this into requirements for the clustering of new articles. In Section 3 we review related work on text clustering. The data set used in evaluating the system is described in Section 4. Sections 5-6 describe our online text clustering system. The evaluation with news articles


*DMCS Workshop, KDD'07*, August 12–15, 2007, San Jose, CA, USA.

in Section 7 shows the high clustering quality and discusses the influence of some key parameters on the performance. The results and lessons learned are discussed in Section 8 and the achievements are summarized in Section 9.

# 15. GEOSPACE & MEDIA TOOL

The GMT is a government funded tool developed under the lead of the Parsons Institute for Information Mapping (PIIM), The New School, NY. It integrates a stream of news articles provided by Factiva (http://www.factiva.com) with geospatial, census, and human network information to provide a research tool for the United States' senators and their staff. The news articles are processed by a customized entity extraction module that combines off-the-shelve software for named entity and geo location detection with algorithms for disambiguation. The articles and the detected entities are then processed by the clustering component described in detail in this study. The news articles and clusters are stored in a relational database system for access by the web-based GMT client. The client lists the currently active top stories and supports user-defined keyword and location searches. The ranking of top stories and search results utilizes statistics pre-computed during the clustering. For each news cluster several representative headlines and keywords are displayed. The extracted locations of the news articles and clusters are used to display them on a zoomable map. Census data provided by ESRI (http://www.esri.com) can be displayed on top of the map to provide context to news stories. The connections between extracted people and organizations can be explored with network displays that are complemented with biographical and contact information.

The clustering component of the GMT serves to summarize news articles about one particular event or topic for the users. The grouping of similar articles eases the browsing of the vast amount of content. The ranking helps in determining the currently most important topics. The users of the GMT are further interested in small stories with local scope. The geo-location of news articles helps in searching such articles via the interactive map or by specifying regions along with tops keywords. For the clustering it means that all news articles need to be clustered, not only articles about the major topics. Existing systems like Google News only display top stories and it is unclear what happens to the rest of the news. Preliminary user tests resulted in positive feedback about the usefulness of the system to congressional staff to solve the daily task of researching various current topics.

Each resulting news cluster should contain as many articles about an event as possible (high recall) and only contain only few articles with low relevance (high precision). The system needs to find a tradeoff between these two quality measures for both large clusters and small clusters simultaneously.

The news stream used in the GMT contains about 50k-100k articles per day. The news articles need to be clustered as fast as possible to provide timely information to the users. With such a high rate of incoming articles clustering of the complete data set will quickly become infeasible. Even clustering only the articles from the last few days will be demanding and more importantly useless, because the result will be outdated when it is available. Data stream clustering techniques are needed [24, 2, 22, 3, 1] to continuously process the feed. The data stream model is commonly characterized such that "the data elements in the stream arrive online", "the system has no control over the order [...]", "data streams are potentially unbounded in size" [6] and there is only one chance of fast access for each data element. For data mining in general [17] and clustering in particular [8] this has been translated into certain requirements.

In summary, we identified the following conditions as being crucial for the clustering the news stream in the GMT:

- Process the stream in a single pass using a small constant time per record and only a fixed amount of main memory [17].
- Process all documents [8], i.e., do not use load-shedding or outlier removal.
- Create a clustering similar in quality to non-streaming algorithms [17].
- Make the clustering available at any point in time [17].
- Do not make assumptions about the number of clusters.
- Dynamically adjust to changing content.

We developed a new solution for this problem because no previously proposed method met the above requirements sufficiently (see Section 3). Even though our method is quite simple it achieves excellent clustering quality in an application to news processing.

# 16. RELATED WORK

Introductions to text clustering can be found in [35, 9, 52, 18]. An analysis of the efficiency and quality of various building blocks from the popular k-Means and Scatter- Gather [15] algorithms for large datasets is done in [31]. In [35] the bi-secting k-Means algorithm performed better than k-means and hierarchical algorithms. The extensive evaluation of hierarchical clustering algorithms concludes that "partitional methods are suitable for producing flat and hierarchical clustering solutions for document datasets effectively and efficiently" [52].

Two well known incremental hierarchical clustering algorithms are BIRCH [49] for numerical data and COBWEB [19] for categorical data. In [39] a variant of COBWEB for text documents is described. The algorithms can update the current clustering model upon arrival of new data points, but they are not well suited for data stream processing [8]. The time and memory requirements of COBWEB can degrade because the internal tree structure is not balanced. BIRCH is designed to use slow secondary memory for the cluster model. Even if the cluster model can be kept in memory completely by using pruning techniques it does not necessarily correspond to a natural cluster structure, a final clustering of the leaf nodes is required [49].

The incremental competitive learning algorithm of [7] is targeted toward text documents and produces a flat clustering. Exactly k clusters of about the same size clusters are created. For a detailed clustering of a collection very different cluster sizes are more desirable as noted for news articles in [45]. Here, a cluster model of a news archive is updated daily in a batch process, thus not making the result available in real time as new articles arrive. In [50, 51] self splitting competitive learning is

advocated to free the user from choosing the number of clusters. In order to assign the vectors of a splitted cluster to the new sub-clusters the vectors of previous data points are needed, violating the one pass requirement.

Clustering of text streams has also been used for retrospective theme detection [36] and topic detection and tracking in the TDT workshops [48]. Topic detection aims to find the first incoming document of a new topic. Here, comparing new documents to all previous documents has worked better than comparing them to clusters [10]. Topic tracking aims at discovering all document that belong to the topic given a certain number of example documents. Topic tracking can be done in retrospective or online [5, 48] with supervised and unsupervised methods [47, 21]. Unsupervised online clustering corresponds to combined topic detection and topic tracking given the first detected document. For the retrospective analysis several clustering algorithms are compared in [48]. Hierarchical clustering performed best, but online single-pass clustering was almost as good. This was explained with the temporal proximity of topics in the data set. In contrast to our requirements some topic tracking systems do not assign all documents to clusters [10].

In [16] the very fast k-Means (VFKM) algorithm is proposed for huge datasets, but it requires several passes over the data using a sample with increasing size in each pass. A one-pass approximation to k-Medians is described in [37, 24]. The data points are clustered in batches to obtain a large number of weighted medians which are successively re-clustered until only k centers remain. The number of clusters k needs to be specified at least roughly and the clustering result is not available at any time. Whenever a clustering of the data observed so far is desired, the clustering up to the final level needs to be initiated.

The approach proposed in [2, 3, 1] does not make any assumptions on the number of clusters, but also divides the process into an online and offline component. The online part keeps a collection of so-called micro-clusters with sufficient statistics and temporal information about the data points assigned to it up to date. At certain time points a snapshot of these clusters is saved. The offline part constructs the final clustering from the stored micro-clusters for a specified time horizon on demand. This step is potentially very expensive for large time horizons. A density based micro-clustering is described in [12]. In [28] the data stream is first segmented by a change point detection algorithm. Each time a segment boundary is found the preceding segment is clustered. This can potentially cause a large delay between the time a data point is observed and when it is clustered.

# 17. DATA SET

Our text clustering system was developed to process the stream of news from many (online) news sources collected by the news provider Factiva. Depending on the number of original sources about 50k-200k articles per day need to be processed. The goal was to provide an efficient way to cluster the news articles at a story level to support the browsing of the articles. This is in contrast with text categorization [34] or topic tracking (e.g. [5]) where broad categories or only major news stories are of interest. The ground truth data available with commonly used benchmark datasets like Reuters-21578 [33] or RCV1 [32] corresponds to (coarse) topic categories and not individual news stories. The TDT datasets contain labeled news stories but the data is collected over several months resulting in relatively low daily and hourly rates of articles that do not represent our target stream well. Moreover, certain datasets like TDT2 [14] include structurally diverse news items, such as radio and TV broadcasts, or have been compiled to benchmark different tasks, such as supervised adaptive topic tracking and multilevel hierarchical topic detection (TDT56). The labeling procedure of the TDT datasets is further biased towards larger stories [48].

We therefore collected the daily news articles from 02/16 to 02/28 in 2006 and labeled 80 news stories of various sizes as our ground truth for evaluation. Several large stories were identified monitoring the websites of major newspapers during the same time period. Smaller local stories were found by filtering the complete data with keyword queries like “Phoenix” and performing an initial very fine grained clustering on this subset. For each story a thorough semiautomatic labeling process was performed involving the following steps executed repeatedly as necessary:

- Search the database for more articles containing important keywords present in the current story.
- Rank the selected articles by their similarity (see Section 5) to the center of the current selection or the closest article in the current selection to help distinguishing relevant and irrelevant articles.
- Search the database for more articles with high similarity to any selected article.
- Rank unselected candidate articles by their similarity to the closest article in the current selection to add very similar articles.

For borderline articles the decision about what was included in a story was checked by at least two persons. The 80 true stories were split into a training part used in optimizing the crucial system parameters and an independent test part to evaluate them. For each data set we added about 1000 unlabeled articles per day randomly selected from the stream. Some statistics of the final data sets are shown in Table 1.

**Table 1: Characteristics of benchmark data sets derived from high density text stream**

| Data set | Articles | | Articles per story | | |
|---|---|---|---|---|---|
| | labeled | unlabeled | min | median | max |
| Train | 4168 | 12416 | 7 | 34 | 682 |
| Test | 3674 | 12371 | 4 | 40 | 643 |

Large stories included a discussion about the hunting accident of Cheney, the resignation of the president of Harvard University and calls for the closing of Guantanamo. Typical medium sized stories were the delays at Delphi and the hostages abducted from an oil platform in Nigeria. Among the smallest stories were the decision about Measure 37 in Oregon and a gunman in Phoenix. Several stories were specifically selected to be very similar, e.g., patent issues of Blackberry and of Adidas vs. Nike.

## 18. PREPROCESSING

The incoming documents are preprocessed with the standard text mining chain of methods [40]. First all words from a list of stop words are removed. For our news application the list included the names of several big news agencies. We further removed all Internet links and Email addresses. From the remaining words a list of lower case word stems is generated with Porter's stemming algorithm [38]. The frequencies of the word stems are saved for each document.

For the news application the names of locations, persons, and organizations were extracted from the original (un-stemmed) text and saved separately with their occurrence frequencies. Each article was further assigned a list of subject categories similar to the RCV1 data. Each category code was treated as a word stem. The importance of this meta-information and of words from the headline and the abstract can optionally be emphasized by artificially increasing their frequencies.

Each word stem frequency is then mapped to a numerical value with the incremental TFIDF (Term Frequency Inverse Document Frequency) scheme of [11] as used in many topic detection approaches [48, 4, 10]:

$$\hat{\mathrm{TF}} = \frac{\mathrm{TF}}{\mathrm{TF} + 0.5 + \frac{1.5 \cdot \mathrm{DL}_N}{\frac{1}{N} \cdot \sum_{i=1}^{N} \mathrm{DL}_i}} \quad (1)$$

$$\mathrm{IDF} = \frac{\log\left(\frac{N+0.5}{\mathrm{DF}}\right)}{\log(N+1)} \quad (2)$$

$$\mathrm{TFIDF} = \hat{\mathrm{TF}} \cdot \mathrm{IDF} \quad (3)$$

where TF is the term frequency (how many times did the term appear in the document), *N* the number of documents processed, $DL_k$ is length of the k-th document in words, and DF is the document frequency (in how many documents did the word appear so far). The more frequently a word appears in a document, the higher the corresponding feature value is. The document frequencies indicate how common a word stem is in the document collection. The more frequently a word appears in the corpus, the lower the corresponding feature value is.

The value of *N*, the sum of the $DL_k$ values, and the DF for each word stem are updated incrementally as the stream is processed. In order to ensure bounded memory consumption we limit the list of word stems and their corresponding DF to a fixed number. When this limit is exceeded we discard the words stems that have not appeared in any document for the longest amount of time. This way the system can adapt to changing topics in the document stream. The most common word stems in the corpus correspond to very common terms of the English language that are not quite general enough to be used as stop words. Their frequencies stabilize quickly [11]. Many word stems with medium document frequencies correspond to currently important topics because they appeared in a significant number of documents. The word stems with very low document frequencies mostly correspond to noise (e.g. misspellings) or very small stories that have a short lifespan in the stream.

## 19. CLUSTERING

In order to meet our requirements we need a single pass algorithm that processes all documents with limited memory and processing time. The documents need to be assigned to clusters immediately to minimize the delay between the point in time when a document is ingested into the database and when it is available to the user as part of a cluster. We cannot make any assumptions on the number of clusters a priori. In Algorithm 6.1 we list the basic single-pass clustering algorithm [48, 44].

Each incoming document is compared to a set of candidate clusters determined by ρ. Different variants of ρ are discussed below. If the distance to the closest cluster is below the threshold T, the document is assigned to this cluster. Otherwise a new cluster containing the current document vector is created. For $\delta(\cdot, \cdot)$ we use the cosine distance of a document vector to the cluster centroid where each vector is normalized to length one. For efficiency the centroid can be pruned to contain only the k largest entries [48].

**Algorithm 6.1** Single pass anytime clustering algorithm.

**Input:**

Document vector stream $V = \{v_i | i = 1, ..., \infty\}$.
Cluster candidate selection function $\rho(\cdot, \cdot)$.
Distance function $\delta(\cdot, \cdot)$ between vectors and clusters.
Distance threshold $T$.

**Output:**

Set of clusters $C = \{c_j \subset V | j = 1, ..., \infty\}$.

$C := \emptyset$
**for all** $i = 1, ..., \infty$ **do**
  $\hat{C} := \rho(C, v_i) \subseteq C$
  $\hat{d} := min\{\delta(v_i, c) | c \in \hat{C}\}$
  **if** $\hat{d} < T$ **then**
    $\hat{c} := c \in C \; \delta(v_i, c) = \hat{d}$
    $\hat{c} := \hat{c} \cup \{v_i\}$
  **else**
    $C := C \cup \{\{v_i\}\}$
  **end if**
**end for**

Multiple cluster memberships can be supported with the following variation: The document is added to all clusters that are sufficiently similar and a new cluster is only created if no cluster with a distance below the threshold is found. The time needed to process a document depends on the size of the candidate cluster set C hat returned by ρ. If efficiency is not an issue we can simple choose

$$\rho_{all}(C, v) := C \quad (4)$$

i.e., compare each incoming document to all existing clusters. Clearly this will not be scalable because as the document stream progresses more and more existing clusters will need to be considered. Using only the most recent clusters from a sliding time window ensures limited memory consumption and processing time independent of the amount of previously clustered documents. Given a maximum age *A* and an age function $a(\cdot, \cdot)$ for clusters we can define:

$$\rho_{\mathrm{window}}(C, v) := \{c \in C | a(c, v) \leq A\} \quad (5)$$

For the age of a cluster we use the difference between the time stamp of the current document and the most recent document in the cluster, other formulations are possible.

For high density streams and long time windows this will still return too many candidates to achieve real-time processing. Let's assume we have 10k documents per day that are on average clustered into clusters of size 10. If we want to keep 1k clusters from each of the last 7 days this leaves us 1.2ms for a single comparison of a document to a cluster including the time needed for IO and preprocessing. With 50k documents per day only 0.05ms are available. We need to reduce the number of document cluster comparison to handle such high density streams together with large time windows. We propose to use locality-sensitive hashing (LSH) [27, 26, 41] to overcome this obstacle. LSH provides an index structure that can be used to determine approximate nearest neighbors, can be updated incrementally, and can deal with high dimensional data [41]. For a given vector the hash returns a (small) set of candidate clusters. These candidate clusters include the most similar cluster with high probability. Let *LSH(·, ·)* be this hash function, then we can define

$$\rho_{\text{hash}}(C, v) := \{c \in \text{LSH}(v, \rho_{\text{window}}(C, v)) \qquad (6)$$

The exhaustive search is only carried out over the much smaller set of clusters as determined by the hash function. This way large time windows can be supported without a prohibitively large increase in processing time. The hash supports online updates. New clusters are added to the hash and changed clusters are updated by removing the old cluster vector and adding the current vector. The hash returns a variable number of clusters for each request. To ensure a limited run time we keep track of the sizes of the clusters stored in the hash and use at most the M largest candidate clusters. We used an efficient variant of the hash functions described in [26]. Instead of implicitly mapping the range of each TFIDF features to several binary columns we used only one binary feature indicating whether the TFIDF value is greater than zero, i.e., whether the word stem is present in the document or not.

Some notes on implementation. The major cost of such a clustering system is the retrieval of the text from the database or files and the storage of vectors and cluster information in the database. If C hat is large the corresponding vectors cannot be kept in memory. The usage of LSH thus saves not only distance calculations but many vector retrieval operations either from a disk cache or the database. The set of relevant clusters C hat can be maintained incrementally or reinitialized in large intervals, e.g., daily. While the algorithm is formulated such that each document is processed individually, it is more efficient to process (small) batches of documents together. This will only cause a small delay in the clustering of the documents early in the batch.

# 20. EVALUATION

We performed a thorough evaluation of our clustering system to optimize parameters for deployment and to analyze the tradeoff between speed and quality. First we analyzed the approximation quality of our LSH based solution to find out how much scale-up one can expect for making certain errors. Next the most important parameters of the clustering algorithms were optimized on the training set and evaluated independently on the test set. We demonstrate the resulting high quality of the online system by comparing our system to a non-streaming method. Finally we performed experiments varying the size of the feature space and the size of the cluster representation to see if we can save more time and space without sacrificing the achieved quality.

## 20.1 Methods

During the following evaluations we fixed several parameters of the system based on prior experience to avoid a combinatorial explosion. We used 250 random hash functions based on 2 random permutations each. During clustering the hash was filled with at most 10k clusters at the beginning of each day selected from the previous 7 days based on size and age. Unless otherwise noted we used a clustering threshold of 0.76 and emphasized all meta data (see Section 7.3.1). The number of active features was at most 50k and the size of cluster vectors was not limited. The documents were processed in batches of size 100.

The clustering quality was evaluated with precision, recall, and F1 [40]. For each ground truth story we evaluated all clusters that contained at least one article of the story. Precision measures how dominant this story is in the cluster, whereas recall measures how much of the story is contained in the cluster. F1 measures a compromise between precision and recall as each can be optimized individually with a trivial solution (one cluster with all documents or one cluster per document). The cluster with the highest F1 score was selected for each story and the unweighted average of the F1 values was used to evaluate a clustering result. We chose not to use weighting by size because we want small clusters to be well represented. To de-emphasize the influence of the random number generation in the hash structure we used several repetitions for each parameter setting and report mean and standard deviation. The test data is only explicitly used in parameter optimization to avoid over fitting and the comparison to the non-streaming algorithm to ensure reproducibility. The other experiments were performed to analyze the system's behavior and give recommendations for the parameter selection.

## 20.2 Nearest cluster approximation

The deployed LSH speeds up the search of the cluster closest to a document vector, but it provides only an approximation. There can be cases where the nearest cluster is not part of the candidate set. We performed some experiments to evaluate the approximation quality similar to [23]. In order to investigate the trade-off between speed and error we varied the number of candidate clusters M that we use for exhaustive nearest cluster search. The more candidates we consider, the more likely the true closest cluster should be found or the smaller the possible error should be but at the same time more distance calculations and vector retrieval operations are needed.

**Table 2: Quality of LSH-based nearest cluster selection.**

| Candidates | Error rate | | Absolute Error | | Speedup |
|---|---|---|---|---|---|
| | LSH | Random | LSH | Random | |
| 250 | 0.39 | 0.93 | 0.0614 ± 0.1057 | 0.1718 ± 0.1742 | 20.4 |
| 500 | 0.31 | 0.89 | 0.0498 ± 0.0881 | 0.1606 ± 0.1696 | 11.3 |
| 750 | 0.26 | 0.87 | 0.0419 ± 0.0700 | 0.1547 ± 0.1669 | 8.5 |
| 1000 | 0.24 | 0.85 | 0.0382 ± 0.0581 | 0.1543 ± 0.1684 | 7.2 |
| 1250 | 0.23 | 0.84 | 0.0354 ± 0.0478 | 0.1514 ± 0.1669 | 6.6 |
| 1500 | 0.22 | 0.84 | 0.0345 ± 0.0450 | 0.1515 ± 0.1674 | 6.3 |

We compared the LSH based cluster candidate selection to an exhaustive search over all clusters in the hash as in [23] and to a random selection that picks the same number of clusters as returned by the LSH method. The results for 5 repetitions of each setting on the training data are listed in Table 2. The test data was not used, as the results are independent of the labeling. We recorded the fraction of erroneous decisions in finding the closest cluster (error rate). The standard deviation is not listed because it was 2 orders of magnitude smaller than the mean.

For all wrong decisions we calculated the mean and standard deviation of the absolute difference between the distance to the best cluster and distance to the selected cluster (absolute error) within each clustering run and list the mean values over the repetitions. To evaluate the speedup in comparison with the exhaustive optimal search we report the percentage of necessary distance calculations that also correspond to the number of vector retrieval operations. The overhead needed to maintain and query the hash is very low as the hash fits into main memory and mainly integer operations are used. We did not use wall clock timing because they are heavily influenced by the configuration of the caches and the memory management of the database and the operating system.

The LSH-based selection proved to be very effective. It leads to much fewer wrong decisions and much smaller absolute errors. With 1000 candidates less than a fourth of the decisions are wrong with a mean absolute error of only 0.04. Compared to the optimal exhaustive search a high speed can be achieved. Above 1k candidates a saturation effect is observed. This is probably due to the necessarily limited capability of the hash structure with fixed parameters. We further investigated the influence of the number of candidates on the cluster quality. The F1 values for the LSH-base approximation are shown in Figure 1. The quality increases clearly up to 500 candidates, above 1k candidates little improvement is observed. Similar results were obtained on the test set. We chose this candidate set size for further experiments.

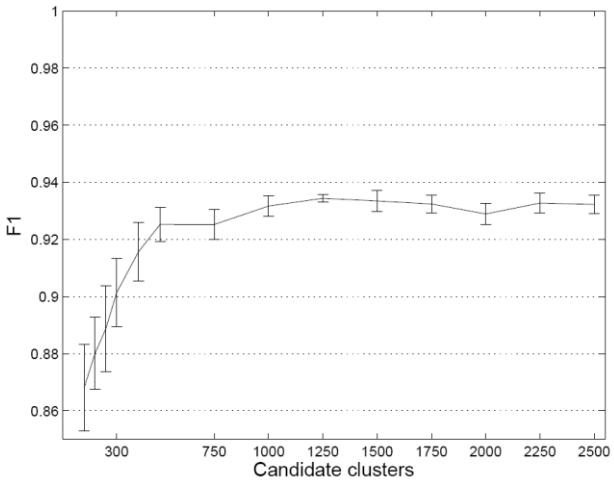



**Figure 1: Cluster quality for LSH approximation with different numbers of cluster candidates.**

## 20.3 Cluster quality

### *20.3.1 Clustering threshold*

Optimization of the threshold parameter is crucial for achieving a good quality with single-pass clustering [48, 25]. We performed a parameter study involving the threshold and different weighting schemes for emphasis of the following meta-data: locations, persons, organizations, categories, headline, and abstract. The baseline setting does not use any meta-information. The best emphasized variant utilizes the meta-information from locations, organizations, categories, and the abstract. In order to further investigate the importance of meta-data we implemented a third weighting scheme simulating the absence of locations, persons, and organizations by removing the corresponding word stems.

**Table 3: P-values for comparison of the cluster quality with a threshold of 0.76.**

| vs. baseline | Training | Test |
|---|---|---|
| Emphasized meta-data | $< 10^{-13}$ | $< 0.0165$ |
| Removed meta-data | $< 10^{-4}$ | $< 10^{-15}$ |

All tests were repeated 20 times to enable an evaluation of significance with the t-test. The results for the training data are shown in Figure 2 with mean and standard deviation for several thresholds.

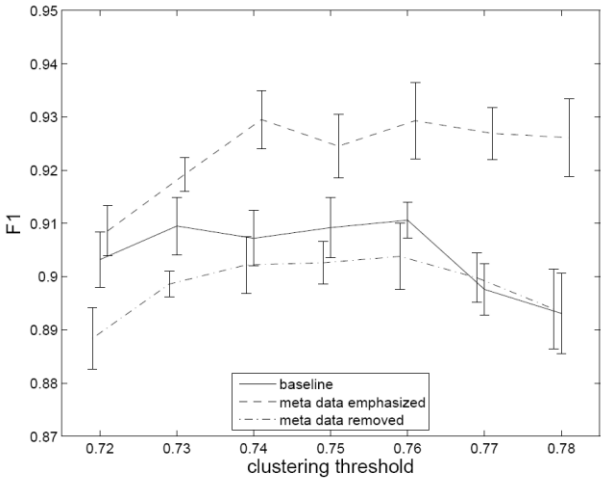



**Figure 2: Cluster quality on the training data for different thresholds and meta-data weighting schemes.**

The emphasis of meta data clearly improves the clustering quality on the training data. The p-values from the comparison using the best clustering threshold of each setting are shown in Table 3. This result can also be reproduced on the test data using the same threshold values as on the training data. The absolute difference in quality is smaller on the test data and for the larger thresholds (corresponding to higher recall but lower precision) it is even better than the emphasized variant. The removal of meta-data results in a significant decrease in cluster quality on both datasets as can be seen from Figures 2-3, and Table 3. It seems that the off-the-shelve text preprocessing already does a decent job in detecting which word stems are most important for the news articles. Nevertheless, emphasizing meta-data explicitly can further improve the quality significantly.

### *20.3.2 Comparison with hierarchical clustering*

Our online clustering system performs quite well as demonstrated in the previous sections. In addition one needs to consider that even a thorough manual labeling process will never be perfect so F1 values of 100% are not to be expected. In order to estimate how much quality is lost due to the online constraints we compared our algorithm to groupwise average hierarchical clustering, one of the best offline text clustering algorithms [48, 47, 31, 25, 52]. Starting with one cluster per document the two closest clusters are merged based on the average similarity of all pairs of documents from the two clusters.

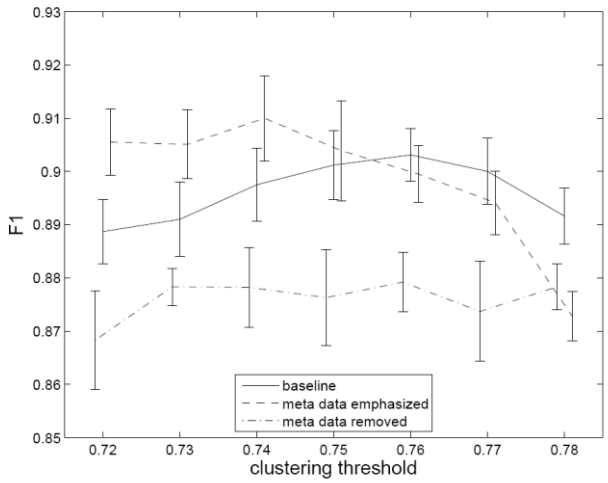



**Figure 3: Cluster quality on the test data for different thresholds and meta-data weighting schemes.**

We stopped merging as soon as the number of clusters was equal to that created by the single-pass algorithm. It turned out that under these conditions the single-pass algorithm performed better than groupwise average clustering on the training data. We varied the threshold of the single-pass and thus the number of clusters for groupwise averaging until an optimum was found. The best results for each method are shown in Table 4.

**Table 4: Cluster quality in comparison with offline group-wise average clustering.**

| Data set | Single-pass | Groupwise |
|---|---|---|
| Training | 0.9378 ± 0.0018 | 0.9453 ± 0.0001 |
| Testing | 0.9091 ± 0.0073 | 0.9258 ± 0.0005 |

For both datasets the single-pass clustering achieves F1 values that are comparable to offline hierarchical clustering. The absolute differences in the F1 values are smaller than 0.01. Of course hierarchical clustering does not scale up to high frequency text streams because it requires the calculation of all pair wise distances which is quadratic in the size of the document collection.

### *20.3.3 Size of the feature space*

An important feature of our system is the dynamic feature space. Only a limited number of word stems can be used at any time to avoid an unbounded increase in the run time of the system. A larger feature space will generally lead to a slower system and extremely large feature spaces will contain a lot of irrelevant word stems. On the other hand the feature space should not be chosen too small because then important words might be discarded if they don't occur in the currently processed document(s). If they occur again at a later point in time they will be added as a new feature with a different vector position. This can lead to errors in the assignment of a document to an existing cluster with similar documents, because such word stems will incorrectly increase the distance. We varied the maximum number of active feature from 15k to 175k and measured the number of re-appearing word stems including duplicates and the cluster quality on the training data as shown in Figure 4 and

Figure 5, respectively. Similar results were obtained on the test set.

The number of re-appearing features is very high for small numbers of active features. If at most 25k features are active at any time a feature is assigned a different vector position than for a previous occurrence more than 100k times. For 100k active features this happens only around 7k times. This is also reflected in the cluster quality that rises steeply up to 50k-75k features and does not improve past 100k features.

Both curves are certainly somewhat data set dependent, the number of unique features in this data set is about 166k. For real life high density streams we recommend to use at least 100k.

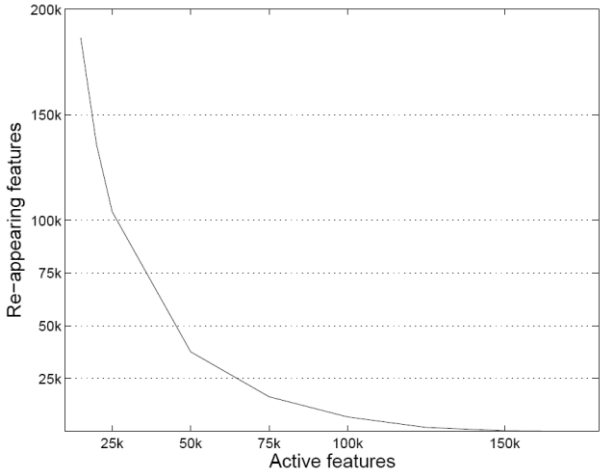



**Figure 4: Number of re-appearing features for different maximum numbers of active features.**

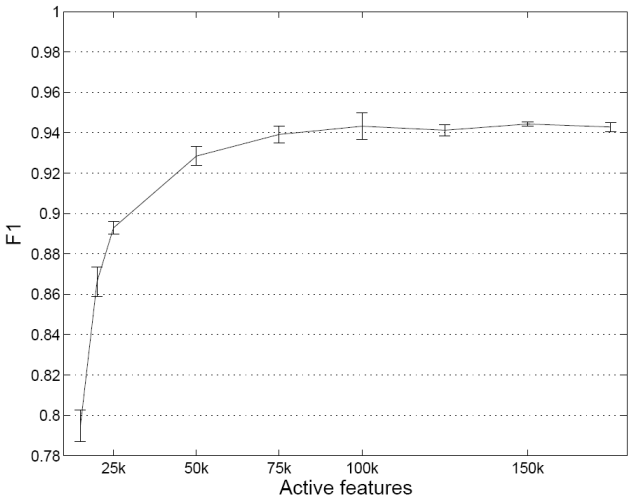



**Figure 5: Cluster quality for different maximum numbers of active features.**

### *20.3.4 Size of the cluster representation*

Apart from the global limitations on the number of features, the vector based representation of each cluster can be limited. For clusters that contain many documents, the number of non-zero entries in the vector calculated as the sum of the individual documents vectors can become very large. The vector sum can be pruned by keeping only the largest d entries and setting additional values to zero [42]. This will reduce noise and increase the speed of distance and hash calculations. We varied the maximum number of features per cluster from 100 to 10k and measured the cluster quality on the training set as shown in Figure 6. Similar results were obtained on the test set.

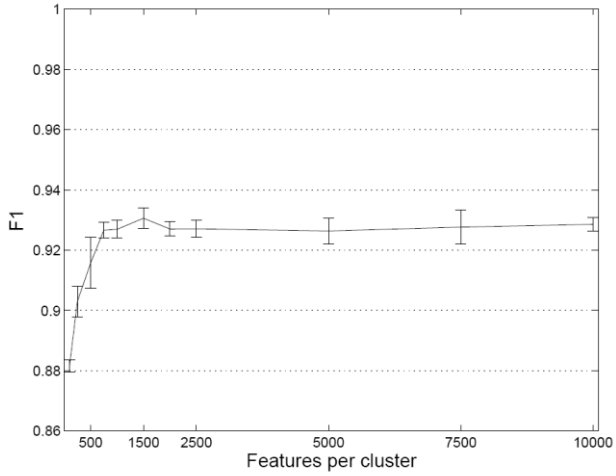



**Figure 6: Cluster quality for different numbers of feature per cluster.**

The clustering quality rises up to a maximum of 1.5k features per clusters. Beyond this no significant degradation could be observed. The influence of noise seems to be negligible but pruning is still worthwhile because it saves memory and processing time. We assume that the entries pruned beyond the largest 1.5k have small TFIDF values and thus do not significantly influence the clustering quality.

## 21. DISCUSSION

We designed a high performance text clustering system and applied it to the real world problem of news aggregation. Our main contribution is the usage of LSH to make the single-pass clustering algorithm [48, 44] scale up to high frequency text streams using a very simple hash function. Previous research used datasets with much lower density of news articles per day and or a posterior analysis of news archives. Under these conditions all clusters from a long time range can be considered for an incoming document and even iterative algorithms might be feasible. For high-frequency text streams our solution creates a good solution efficiently. We showed empirically that a very high level of cluster quality is maintained even though only a small fraction of the distance calculations and the associated retrieval of cluster vectors are needed. Our reported experiences give insight into the problems that are encountered when dealing with large-scale problems.

We store the centroid vectors of clusters from a sliding time window in the hash structure to find good candidate clusters for an incoming document vector. If memory permits, document vectors could be stored in the hash with the corresponding

cluster id to assign a document to the cluster with the closest document. For topic detection this single-link approach is of advantage [48] whereas for clustering the former group-average paradigm is reported to work better [25].

The clustering quality of the single-pass clustering algorithm has been reported to be almost as good as or even better than iterative algorithms [48, 25] if the clustering threshold parameter is set appropriately. The optimal threshold on our training data was between 0.76 and 0.78 depending on the particular experimental setting. This compares with previously mentioned values and ranges: 0.77 [48] 0.7- 0.9 [25]. We achieved a cluster quality that is comparable with one of the best offline clustering algorithms for text data, namely group-wise average hierarchical clustering.

The single-pass clustering is similar to micro-clustering [3] without the on-demand step. New documents are added to the most similar micro-cluster if the similarity is high enough. If the maximum amount of micro-clusters k is reached inactive clusters are removed. This is similar to our time window. The LSH technique could also be used to speed up micro-clustering.

If we were to execute the on-demand clustering step on a regular basis, several problems would arise within our applications. First of all it would need to done frequently to minimize the delay between the time when the document is ingested into the system and the time when it is available to the user as part of a cluster. Even hourly clustering would mean a significant delay for a news system. Also, it has been reported in [10] that a delay does not necessarily help in new event detection. If each on-demand clustering is done independently, the amount of clusters that are saved to the database is much larger compared to single-pass clustering where one cluster can stretch over a long period of time. Clusters from close-by snapshot times would be very similar creating near duplicates in the database that are hard to detect and filter. A possible solution to this would be the recently proposed evolutionary clustering framework [13] where consecutive clusterings are required to be similar. In this case a previous cluster could be associated with a similar current cluster and saved as a single object in the database. This approach does not, however, support clusters with gaps longer than the clustering interval. Our single-pass clustering supports long cluster lifetimes including gaps up to the duration of the sliding time window. The historical information of the time stamps of documents within a cluster can be easily reconstructed from the database on demand. Finally in [3] there is no mentioning of limiting the number of features which is a potential memory problem.

Our experimental study indicates that utilizing locations, persons, and organizations is of advantage. This is in accordance with previous studies [30, 25]. In [25] one vector for the text and separate vectors for entities and noun phrases are generated. The similarities from comparing the corresponding vectors from different documents are mixed with weights found by regression on a training data set. We integrated the text information and the meta-data into a single vector with emphasized frequencies for meta-data terms. This enables the use of a single hash structure to find a single set of cluster candidates. When using several vectors and similarities the determination of good candidates will be more involved.

When varying the maximum length of cluster centroid vectors we found that the clustering quality on our training increases up to around 1.5k entries and does degrade for larger values. This is in contrast to previous studies that truncated the vectors for efficiency down to 20 [42] or 25 [31] entries. One reason for our observation might be the larger vocabulary and size of our corpus. We recommend to use much higher values.

We did not use any dimensionality reduction techniques like latent semantic indexing (e.g. [9, 18]) because the projection would require additional online computation and we would loose the direct correspondence of feature with words stems that is utilized to generate keywords for each cluster.

## 22. SUMMARY

We presented a system for high performance online text clustering of a stream of news articles that meets all the identified requirements of the Geospace & Media Tool. The system has been tested on the complete news stream over several weeks and successfully discovered top stories as reported by other news sites. The system will be deployed in the near future to aid members of Congress their staffs in analyzing the daily news in connection with geospatial, census, and human network information.

The textual content of the articles is analyzed and similar articles are grouped into clusters on-the-fly without any assumptions on the number of clusters and without retrieving previous documents. The result is available to the user at any time without additional on-demand clustering steps. The system dynamically adjusts to changing topics by gradually adapting the feature space. Efficiency is ensured by limiting the amount of currently active features and by considering only clusters from a finite time horizon for the assignment of incoming documents. Very large time windows can be supported by using locality-sensitive hashing to summarize the clusters and find the most similar cluster for each document with high probability. We demonstrated the effectiveness and efficiency of the system on the very demanding application of news clustering. The clustering quality is comparable to one of the best non-streaming document clustering algorithms and the architecture can easily support several 10k documents per day on off-the-shelve hardware.

## 23. ACKNOWLEDGMENTS

We acknowledge the team at Parsons Institute for Information Mapping, The New School, New York, NY, for their collaboration in integrating our text processing methods into the GMT and providing the extracted meta-data for our experiments. We further thank Marc Muntziger for his help in the experiments regarding the meta-data emphasis and Sheik Abdul-Saboor and Tino Hertlein for helping to label the ground truth stories.

## 24. REFERENCES

[1] C. Aggarwal. Data Streams: Models and Algorithms. Springer, 2007.

[2] C. Aggarwal, J. Han, J. Wang, and P. Yu. A framework for clustering evolving data streams. In Proceedings of the

International Conference on Very Large Databases (VLDB), 2003.

[3] C. Aggarwal and P. Yu. A framework for clustering massive text and categorial data. In Proceedings of the SIAM Conference on Data Mining (SDM), 2006.

[4] J. Allan, V. Lavrenko, D. Malin, and R. Swan. Detections, bounds, and timelines: UMass and TDT-3. In Proceedings of the Topic Detection and Tracking Workshop (TDT-3), 2000.

[5] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 37–45, 1998.

[6] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom. Models and issues in data stream systems. In Proceedings of ACM Symposium on Principles of Database Systems (PODS), pages 1–16, 2002.

[7] J. Banerjee and A. Ghosh. Competitive learning mechanisms for scalable, incremental and balanced clustering of streaming texts. In Proceedings of the International Joint Conference on, Neural Networks (IJCNN), volume 4, pages 2697–2702, 2003.

[8] D. Barbara. Requirements for clustering data streams. SIGKDD Explorations, 3(2):23–27, 2002.

[9] M. W. Berry. Survey of Text Mining: Clustering, Classification, and Retrieval. Springer, 2003.

[10] T. Brants and F. Chen. A system for new event detection. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 330–337. ACM Press, 2003.

[11] J. Callan. Document filtering with inference networks. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 262–269. ACM Press, 1996.

[12] F. Cao, M. Ester, W. Qian, and A. Zhou. Density-based clustering over an evolving data stream with noise. In Proceedings of the SIAM Conference on Data Mining (SDM), 2006.

[13] D. Chakrabarti, R. Kumar, and A. Tomkins. Evolutionary clustering. In Proceedings of International ACM Conference on Knowledge Discovery and Data Mining (KDD), pages 554–560, 2006.

[14] C. Cieri, D. Graff, M. Liberman, N. Martey, and S. Strassel. The TDT-2 text and speech corpus. In DARPA Broadcast News Workshop, 1999.

[15] D. Cutting, D. Karger, J. Pedersen, and J. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 318–329, 1993.

[16] P. Domingos and G. Hulten. A general method for scaling up machine learning algorithms and its application to clustering. In Proceedings of the International Conference on Machine Learning (ICML), pages 106–113. Morgan Kaufmann, 2001.

[17] P. Domingos and G. Hulten. A general framework for mining massive data streams. Journal of Computational and Graphical Statistics, 12:945–949, 2003.

[18] R. Feldman and J. Sanger. The Text Mining Handbook. Cambridge, 2007.

[19] D. Fisher. Knowledge acquisition via incremental conceptual clustering. Machine Learning, 2(2):139–172, 1987.

[20] G. Forman. Tackling concept drift by temporal inductive transfer. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 252–259. ACM Press, 2006.

[21] M. Franz, J. McCarley, T. Ward, and W.-J. Zhu. Unsupervised and supervised clustering for topic tracking. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 310–317, 2001.

[22] M. Gaber, A. Zaslavsky, and S. Krishnaswamy. Mining data streams: a review. ACM SIGMOD Record, 34(2):18–26, 2005.

[23] A. Gionis, P. Indyk, and R. Motwani. Similarity search in high dimensions via hashing. In Proceedings of the International Conference on Very Large Databases (VLDB), 1999.

[24] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. IEEE Transactions on Knowledge and Data Engineering, 15(3):515–528, 2003.

[25] V. Hatzivassiloglou, L. Gravano, and A. Maganti. An investigation of linguistic features and clustering algorithms for topical document clustering. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 224–231. ACM Press, 2000.

[26] P. Indyk and R. Motwani. Approximate nearest neighbor: towards removing the curse of dimensionality. In Proceedings of the ACM Symposium on Theory of Computing, 1998.

[27] P. Indyk, R. Motwani, P. Raghavan, and S. Vempala. Locality-preserving hashing in multidimensional spaces. In Proceedings of the ACM Symposium on Theory of Computing, pages 618–625, 1997.

[28] A. Jain, Z. Zhang, and E. Chang. Adaptive non-linear clustering in data streams. In Proceedings of the ACM International Conference on Information and knowledge management (CIKM), pages 122–131. ACM Press, 2006.

[29] T. Joachims. A statistical learning model of text classification with support vector machines. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 128–136. ACM Press, 2001.

[30] W. Lam, H. Meng, K. Wong, and J. Yen. Using contextual analysis for news event detection. International Journal Of Intelligent Systems, 16(4):525–546, 2001.

[31] B. Larsen and C. Aone. Fast and effective text mining using linear-time document clustering. In Proceedings of

the Text Mining Workshop at the International ACM Conference on Knowledge Discovery and Data Mining (KDD), pages 16–22. ACM Press, 1999.

[32] D. Lewis, Y. Yang, T. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research,, 5:361–397, 2004.

[33] D. D. Lewis. An evaluation of phrasal and clustered representations on a text categorization task. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 37–50, 1992.

[34] D. D. Lewis. Machine learning for text categorization: background and characteristics. In Proceedings of the 21st National Online Meeting, pages 221–226. Information Today, 2000.

[35] S. M., K. G., and K. V. A comparison of document clustering techniques. In Proceedings of the Text Mining Workshop at the International ACM Conference on Knowledge Discovery and Data Mining (KDD), 2000.

[36] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In Proceedings of the International ACM Conference on Knowledge Discovery and Data Mining (KDD), pages 198–207. ACM Press, 2005.

[37] L. O'Callaghan, N. Mishra, A. Meyerson, S. Guha, and R. Motwani. Streaming-data algorithms for high-quality clustering. In Proceedings of IEEE International Conference on Data Engineering, 2002.

[38] M. F. Porter. An algorithm for suffix stripping. pages 313–316, 1997.

[39] N. Sahoo, J. Callan, R. Krishnan, G. Duncan, and R. Padman. Incremental hierarchical clustering of text documents. In Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM), pages 357 – 366, 2006.

[40] G. Salton. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison Wesley, 1989.

[41] H. Samet. Foundations of Multidimensional and Metric Data Structures. Morgan Kaufman, 2006.

[42] H. Schuetze and C. Silverstein. Projections for efficient document clustering. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 74–81. ACM Press, 1997.

[43] F. Sebastiani. Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47, 2002.

[44] D. Shen, Q. Yang, J.-T. Sun, and Z. Chen. Thread detection in dynamic text message streams. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 35–42. ACM Press, 2006.

[45] A. Smeaton, M. Burnett, F. Crimmins, and G. Quinn. An architecture for efficient document clustering and retrieval on a dynamic collection of newspaper texts. In BCS-IRSG Annual Colloquium on IR Research, Workshops in Computing, 1998.

[46] G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. Machine Learning, 23(1):69–101, 1996.

[47] Y. Yang, J. Carbonell, R. Brown, T. Pierce, B. Archibald, and X. Liu. Learning approaches for detecting and tracking news events. IEEE Intelligent Systems, 14(4):32–43, 1999.

[48] Y. Yang, T. Pierce, and J. Carbonell. A study on retrospective and on-line event detection. In Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR), pages 28–36, 1998.

[49] T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In Proceedings of the International ACM Conference on Managemet of Data (SIGMOD), pages 103–114, 1996.

[50] Y.-J. Zhang and Z.-Q. Liu. Self-splitting competitive learning: a new on-line clustering paradigm. IEEE Transactions on Neural Networks, 13(2):369–380, 2002.

[51] Y.-J. Zhang and Z.-Q. Liu. Refining web search engine results using incremental clustering. International Journal of Intelligent Systems, 19(1-2):191–199, 2004.

[52] Y. Zhao and G. Karypis. Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery, 10(2):141–168, 2005.

# Data mining for quality improvement

Françoise Fogelman Soulié
KXEN http://www. kxen.com
25 Quai Gallieni
92 158 Suresnes cedex – France
francoise@kxen.com

Doug Bryan
KXEN http://www. kxen.com
201 Mission Street
CA 94130 San Francisco – USA
Doug.Bryan@kxen.com

## ABSTRACT
In this paper, we describe how data mining can be used in large projects for quality improvement. We first introduce the context of quality performance in SixSigma initiatives, we describe the conventional methods implemented in SixSigma for monitoring quality. We then show how data mining can be used in such context and present three examples of ways a large telco operator is presently using data mining in quality improvement applications. All three applications described demonstrate the same result : by producing models on the large volumes of data available in telco, companies can get a huge return on the investment they put into gathering them, turning data into a strong asset to improve their business processes quality. Yet, deploying data mining on a large scale poses specific constraints which we discuss.

## Categories and Subject Descriptors
G.3 [Probability and Statistics] : *Nonparametric statistics, Robust regression, Statistical computing*

I.2.6 [Learning] : *Knowledge acquisition*

I.5 [Pattern Recognition] : I.5.1 Models – *Statistical.*

J. [Computer Applications]

## General Terms
Algorithms, Experimentation, Quality Performance, SixSigma.

## Keywords
Data Mining, Industrial applications.

## 25. INTRODUCTION
Historically, data mining has been mostly used for applications in CRM : models are built to define targets for marketing campaigns, customers' life time value or customers segmentation. In banking and insurance; to evaluate risk (credit scoring), fraud (credit card); to identify buying behavior in retail, produce on-line recommendations and ratings in e-commerce ... All those sectors have heavily invested in building huge data warehouses (from a few terabytes to peta-bytes) that contain millions of records and thousands of variables. Increasingly, these data warehouses are becoming enterprise-wide encompassing business processes from design to production to maintenance of products and services. It then becomes possible to mine through very rich data sets providing innovative ways for improving process quality. This approach allows, in particular, to improve existing Six Sigma methods deployed in quality & performance management.

I will present three examples of this in telecommunications : for cellular network optimization, for network maintenance and for customer satisfaction when using the internal Information System.

## 26. PERFORMANCE IMPROVEMENT
Today, competition is global and all companies struggle to improve their productivity, which is the key to increase profitability : « As markets liberalize & globalize, the only sustainable source of higher profitability for a firm will be to continually raise productivity higher than its competitors » [1]. On this quest for productivity, companies have started to systematically and continuously analyze their business processes trying to improve their performances. Because technology provides ways today to both collect and exploit huge volumes of data, companies have embarked into enterprise-wide efforts to industrialize their performance improvement initiatives, putting in place quality improvement tools (Lean, Six Sigma), data warehouses and analytics tools. Those who are successful in these efforts (which Thomas Davenport [2] calls the "Analytics Compe-titors") start to differentiate from their competitors and their profitability increases faster than in their business sector : « At a time when firms in many industries offer similar products and use comparable technologies, business processes are among the last remaining points of differentiation. And analytics competitors wring every last drop of value from those processes » [3]

Performance improvement thus focuses on the Enterprise's processes : production (for producing products & services), support (HR, IT), operational monitoring and management … For each process, performance indicators are defined (KPI : Key performance Indicators are those which are critical for



*KDD'07*, August 12, 2007, San Jose, California, USA.

performance) and targets assigned, depending upon the present level achieved by the company and benchmarks against competition. The performance improvement process (Figure 26-1) then aims at reducing the gap between current and target KPI. It first tries to identify *root-causes* of low performances : this is very often done "by hand" by field technicians for production processes, or through Business Intelligence reports for monitoring and management processes. Processes usually produce data, sometimes in large volumes (as we will see in section 29) : these data are most often temporarily collected and looked at to identify the root causes; rarely, do companies go to the effort of collecting and aggregating these data into a datawarehouse (unless they're analytics competitors !)

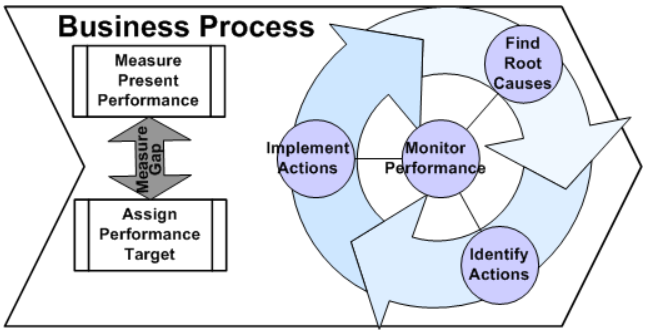



**Figure 26-1 – Performance improvement process**

When root causes have been found, corrective actions must be identified and implemented. In this iterative process, KPIs are measured and monitored; performance continuously improves.

We now introduce a few concepts of one of the most well known performance improvement approaches : Six Sigma. However, data mining can be used in any performance improvement problem in a fashion very similar to what we descrive below.

## 27. SIX SIGMA

Six Sigma was invented in 1986 at Motorola by B. Smith to handle a strong increase in returns under warranty. In his attempt to devise a method to standardize the way defects are counted, Smith developed initial concepts, tools and methodology and started Six Sigma, which rapidly became central to Motorola quality strategy. From there, it spread to many companies, in manufacturing and logistics at first, and is nowadays used in many different domains, in small and medium organizations as well. The reason for this is that Six Sigma brings significant performance improvement : Motorola reports various examples of this [4].

The key feature of the Six Sigma process is that it requires that people gather data on their critical business processes, structure it and analyze it to take their decisions : measures for processes performance must be defined, together with the data needed to evaluate them and goals to be achieved, very much in the fashion shown in Figure 26-1.

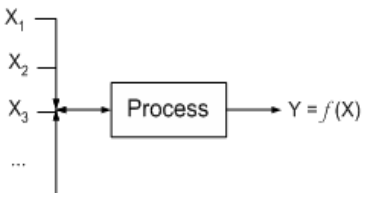



**Figure 27-1 – A business process**

In Six Sigma, a performance-driven project is considered as a process, with input variables (called Xs) and an output variable (called Y) : Xs will be used by the process to produce Y. Y is typically associated with some goal we want to achieve : usually Ys are linked to customer requirement, or some intermediate goals. Of course, all these Xs and Y have to be defined, measured and improved along the project. A transfer function describes the relationship between Xs and Y, and the goal of a Six Sigma project is to understand that relationship and in particular which of the Xs are the most important for Y (these variables are called « root causes » or « vital few » in Six Sigma, « key drivers » in business or « significant variables » in statistics). Even though many Xs can be used in the original definition of a transfer function, once we have identified the key drivers, we can both restrict the number of variables in the function and also act to control those variables only, thus limiting the number of actions to be undertaken.

Six Sigma comes with methods, concepts and tools, but does not use modern data mining technologies. We will show that data mining indeed has the potential to bring value to Six Sigma and quality performance improvement projects.

## 28. DATA MINING FOR SIX SIGMA

The most significant use of data mining for Six Sigma is in building a process model $\hat{Y} = f(X)$, which can then be used in various ways in Six Sigma conventional analysis :

– Identify *root causes* : the model parameters are used to assess the various variables significance. If the model is linear, the relative values of the coefficients – possibly normalized – will provide that information : for example, Figure 28-1 shows the relative importance of variables in a maintenance model [11] : the root causes for a repair intervention are the Company's sector, the maintenance contract level, whether the product has already been repaired and the product installation date. Once the most important variables are identified – *root causes* – then actions can be devised to work on these variables so as to improve performance (Figure 26-1); one can also build a new, simpler model with only these variables.

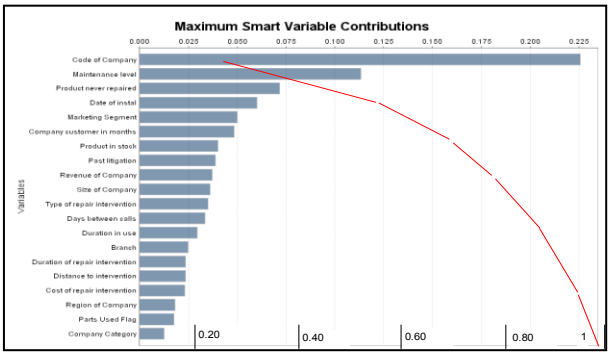



**Figure 28-1 – Importance of variables**

In addition to this, it is possible to also investigate the impact of one precise category of a variable on output Y : some categories of that variable might be positively correlated with

[11] All models and figures in this paper are produced using KXEN Analytic Framework v4.0

Y while others are negatively correlated : Figure 28-2, for example, shows that variable *Company customer in months* (ie how many months has the company be a customer) has categories with various impact on the target Y (Y=1 if product has failed and thus needs repair); namely, new customers (less than 20 months) tend to have high risk of asking repair, while "older" customers tend to not be at risk. Identifying such effects will allow to better understand root causes and better identify corrective actions.

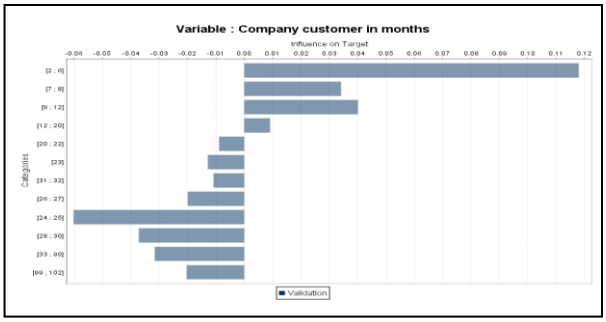



**Figure 28-2 – Impact of variable *Company customer in months* on target Y**

– *Pareto* Analysis : in Six Sigma, it is common to perform a Pareto analysis, identifying those most significant variables which account for most of the effect on Y. In a typical Pareto analysis, one expects some "80-20" Pareto rule, where 20% of the causes account for 80% of the effect (Figure 28-1 actually shows that 40% of the variables account for 80% of the effect.)

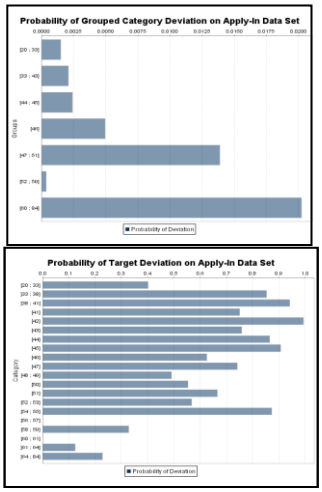



**Figure 28-3 – Deviation in distributions**

– *Failures* detection: when a process runs continuously, data distributions can drift, usually resulting in degradation of performance. This drifting can come either from a change in some of the variables distributions (Figure 28-3 – left shows a period where the distribution of variable *Duration in Use* – how long has the product been in use – has deviated from the previous period where the model was produced, especially in one category); deviation can also come as a change in the cross-distribution of one variable with respect to Y (Figure 28-3 – right shows that the cross-statistics of variable *Duration in Use* with respect to Y has significantly changed in some categories.) In quality problems, such deviations will very often happen because of the failure of some component in the process : analysis of deviations of a model in successive periods can thus be used to detect failures during a period.

The ability to produce models and use them in the above fashions in situations where processes can produce millions of records with thousands of variables per day is actually quite a revolution for Six Sigma and quality improvement projects. Tools typically used for such projects, such as Minitab (http://www.minitab.com/), completely lack this ability, making the full analysis of complex processes very hard. As a result, failures tend to be detected only long after they happened and after lots of hard field work has been devoted to figure out the problem. Data mining thus brings a very innovative and promising path for such projects. We now present some typical examples in telecom.

# 29. APPLICATIONS

The telecommunication industry is a sector which produces very large volumes of data and is one of the earliest and most advanced user of data mining techniques. Today, these techniques are mostly used in CRM (Customer Relationship Management) : very large numbers of models are used for targetting marketing campaign (for example, Vodafone D2 [5] or Cox [6] produce hundreds of models per year with KXEN.) However, many other potential applications exist in other areas of that industry.

We have worked with a very large telco operator and developed applications to handle quality issues in mobile network optimization, network maintenance and customer satisfaction : in each of these applications, which we describe in the next sections, we have used KXEN Analytic Framework v4.0, mostly the regression / classification K2R module briefly described below.

## 29.1 Data mining approach used

KXEN software is based upon the Structural Risk Minimization of Vladimir Vapnik [7] and implements the basic functions of data mining as listed in [8] : classification / regression, segmentation, time series, association rules and attribute importance. In Structural Risk Minimization, an embedded collection of models families is chosen with increasing VC dimension (eg a family of linear models with increasing norm of parameters). The data set is cut into 3 subsets : the estimation set is used to find the best model which fits its data; the various optimal models produced in each family are evaluated on the evaluation set and the best one is selected as that which provides the optimum error on validation set. The final model can then be tested on test set. SRM ensures that the best fit – robustness compromise is achieved, as demonstrated in Vapnik [7].

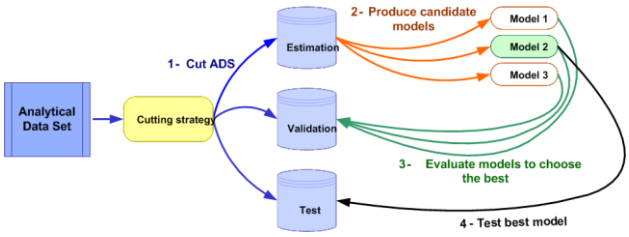



**Figure 29-1 – Modeling process using SRM**

For classification / regression, module K2R uses polynomials (the degree is a free parameter) : the SRM method allows to find the best polynomial. Usually degree 1 – ie linear regression – is sufficient because variables are first automatically encoded : module K2C non-linearly encodes variables of all types (continuous, ordinal, categorical), finding the best encoding through SRM. After that, the classification / regression module K2R works as shown in figure Figure 5-2 : variables are first encoded through K2C, then an embedded family of models with increasing VC dimension is chosen. A learning algorithm – ridge regression – is used to fit the data in the estimation set and the best model in the family is chosen by monitoring the criterion on the evaluation set. The criterion we use – KI – is specific to KXEN : it is related to AUC by $KI = 2 * AUC - 1$.

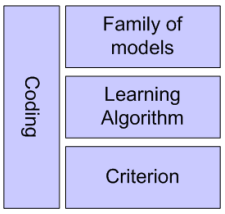



**Figure 29-2 – Structure of modeling in KXEN**

KXEN has been designed to automate the data mining process, making it possible to produce easily thousands of models on very large volumes of data [9] (time to produce a model is typically in the few minutes / hours range, because the lenghty process of selecting / encoding the variables is automatized) : this makes it perfectly suited to large problems in telco such as those we describe in the next section.

One point which was critical in the applications described below is the nature of the teams involved : they were business users (telco engineers, call centers operators or IT engineers). There was no statistician / data miner available at that telco operator and the goal was to allow the business users to handle the model production themselves. So, it was not about comparing the proposed approach to a previously existing solution, it was rather to demonstrate that business users themselves could actually make use of their data to improve the performances of their operations.

## 29.2 Mobile network deployment

Cellular radio networks are complex, heterogeneous, adaptive systems that cover hundreds of square miles and support millions of users. When performance drops the finger-pointing between stakeholders – subscribers, service providers, network operators, handset OEMs, and network OEMs – begins. Often the first step is to identify the vital few key drivers of the performance indicator that dropped, so that systems engineers can develop hypothesis about root causes. This is usually done by the engineers on the field, where they have to go and fine tune the parameters of hundreds of thousands of network components. For that reason, deploying, optimizing and maintening a cellular network is a lengthy process, very costly in terms of skilled telco engineers and technicians. The goal of this project was to see how data mining could be used into this process to help field engineers and optimize performance.

**Table 29-1 – Variables for network optimization**

| Network component | Number of variables |
|---|---|
| cell | 580 |
| channel | 390 |
| carrier | 70 |
| site | 25 |
| neighbor | 40 |
| band | 20 |

Cellular networks are comprised of a grid of cells operating on a hierarchy of components :

- *Sites* contain a group of antennas and often are at the intersection of three cells;
- *Cells* are a geographic area covered by an antenna. A network may contain different kinds of cells depending on the technology and protocols used;
- *Neighbors* are adjacent cells;
- *Channels* are a section of the radio frequency spectrum;
- *Bands* are a range of channels;
- *Carriers* are radio waves having at least one characteristic, such as frequency, amplitude or phase.

We aggregated data from 24 hours of operation of a large urban network. The data were gathered from the cellular network equipment and included 40 million rows and over 1000 variables, just for that one day. Table 5-1 shows the variables we used.

There are thousands of performance indicators used for network deployment : some are very technical (linked to the signal, the cells, the channels …), others are more directly related to Quality of Service (coverage, interrupted calls …) Fine tuning the network elements typically has impact on many indicators in a very intricate way. The idea behind this project is to monitor thousands of indicators to identify, for the most critical ones – the Key Performance Indicators – the critical elements in the network. In what follows, we describe the typical approach for handling one such KPI : it has to be repeated for thousands KPIs.

Various models were then elaborated to explain some of the most critical Key Performance Indicators for the optimization of the cellular network. Most of the work was to aggregate the data and put it into a database where it could be accessed to build the model. This required a few weeks of work for this initial project.

After that, for each KPI Y, a model was built using KXEN K2R module (classification / regression) to explain Y in terms of the variables shown in Table 5-1. Building such a model and refining it typically required 2 to 3 days.

A first result of this model was to provide an encoding of variables (this is done automaticaly by KXEN module K2C) : this non-linear encoding proved to provide very useful information. For example, in the case of a continuous variable Y, the non-linear encoding built by KXEN is much more easy to understand (Figure 29-3 – right shows one variable as a function of Y, the “elbow” which appears is very significant for systems engineers) than the usual scatter plot used in SixSigma (Figure 29-3 – left shows that same variable X as a function of Y)

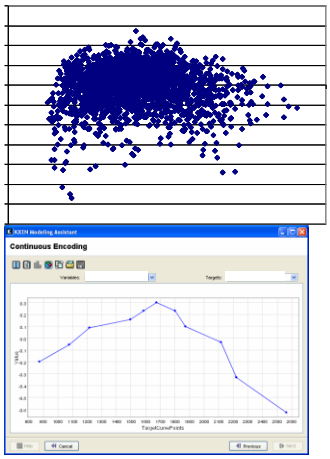



**Figure 29-3 – Variable X as a function of Y**

The initial 1125 variables were automatically reduced to just a few dozens, among which the key drivers were extracted : this process found the same top three drivers than the field engineers had. But the data mining process took 2-3 days while the systems engineering team had previously spent man-years.

The hard benefit of using data mining to analyze network performance is straight-forward : reduce staff hours. However the soft benefits may even be greater :

- Meet service-level agreements in time;
- Reduce false warnings of poor performance by taking into account hundreds of variables, thus allowing management and top engineering resources to focus on what really matters;
- Reduce *time-to-market* for performance enhancements : because data mining model allows to identify root causes of problems in 2-3 days work instead of man-years of field tests, the telco operator can optimize its network faster and bring it to market earlier;

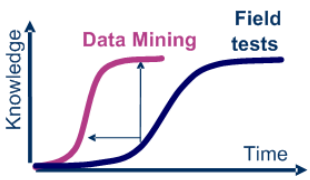



**Figure 29-4 – Data Mining reduces Time-To-Market**

- Increase *customer satisfaction*.

Data warehousing and predictive analytics technologies now make it cost-effective to collect, store, aggregate, and analyze performance data on hundreds of thousands of network components every day. By using the KXEN Analytic Framework to automatically monitor multiple key performance indicators for every site, cell, and carrier in a market, using hundreds of input variables, engineers can spot components with low performance and report those in executive dashboards and balanced scorecards, while taking into account traffic, season, day-of-week, and any other significant variables. Reports on top -10 key drivers for each performance indicator can be produced into an engineering dashboard. Both performance and key drivers indicators can be updated daily.

This telco operator is now deploying a much larger project on his cellular network where it expects to gain critical time-to-market, while saving on human resources.

## 29.3 Network maintenance

This telco operator is running a call center, where customers can call when they have problems with the services they bought : fixed line, cellular, ADSL connection, TV-on-ADSL … Customers can call for many reasons : problems because they do not know how to use their service (especially in the first few weeks after installation), problems with the equipment they have at home or problems because some component somewhere in the network servicing their home failed. The process in place involves analyzing the calls and sending staff on the field to identify the causes of the problem and fix it so that customers can use the service or product to which they subscribed.

This operator implemented a first test where data were collected from four main sources : customer data, subscription data, network data and call data. We collected data from 4 weeks of operation of one regional call center. Each week had about 200 000 records with some 200 variables and we aggregated them with customer and network data into one repository.

We implemented various models :

- *Customer level* : we first eliminated the recent customers (those who have registered recently and do not know yet how to use their equipment). We then built a model for each week to predict whether a customer had called during that week. We identified the key drivers of that model. Among the top 5 key drivers were 2 variables describing network component. By looking into the impact of these variables onto the target (Figure 29-5), engineers could identify the component categories (ie localizations) which had failed, thus causing the customers to call.

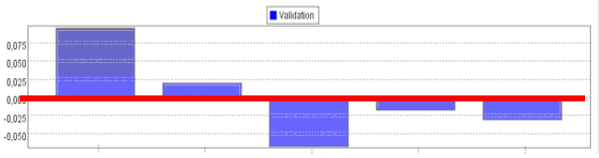



**Figure 29-5 – Groups of network components**

We then looked for deviations of each week $W_t$ model on the next week $W_{t+1}$ data : KXEN deviation detection function gave us the list of the components which had failed that week $W_{t+1}$ (reports are like in Figure 28-3). On-field tests validated all the findings.

- *Call level* : we first eliminated customers who had not called during 2 successive weeks $W_t$ and $W_{t+1}$ and then built a model to predict which week the call was. The model gave us the key drivers : these are those equipments which "acted" different those 2 weeks (ie generated calls one week and not the other).

This operator is presently industrializing the data collection and will put in place the analysis process afterwards.

## 29.4 Customer satisfaction

This telco operator has a few hundred thousands of employees using hundreds of internal applications into its Information System. Internal quality department sends a survey every week to about 4 000 employees : they ask whether employees had problems using the IS applications and what their satisfaction index is; employees can also comment into a free text field. Initially, returned emails were "looked at" and only striking

events were identified. But management wanted to get more out of these emails.

A project was thus set up to produce a fully industrialized process by which all electronic surveys would be incorporated into a data base and analyzed. Each week, models are executed to produce 4 satisfaction indices, and analysis along various axes (application, work position, organisation, business units).

The models were built using KXEN modules (classification / regression K2R and text coder K2C to take advantage of the free-text field); They allowed to identify the major problems in applications (which applications came in the top-5 key drivers), in business domains and even in network equipment.

The text-field was particularly interesting. KXEN KTC works by extracting from a text zone the most frequent "roots" : first we eliminate words in stop-lists (such as "a", "for" …), then apply stemming rules (such as "problems" is-replaced-by "problem"). The roots are then added as additional variables and further used just as other variables (the automatic encoder module K2C handles the text-extracted variables just as it does other variables.)

Usually, running KTC added a few hundreds variables to the initial ones. But, in almost all models, the text-extracted variables were among the top-ten key drivers (Figure 29-6), showing that employees usually told very important things indeed in the free-text zone.

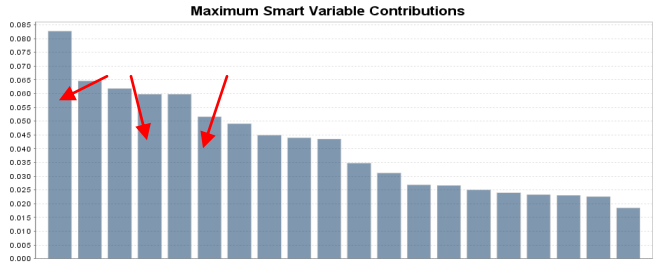



**Figure 29-6 –Most important variables in satisfaction surveys often are text-derived variables (in red)**

This telco operator is now routinely using this application and produces reports to follow-up users satisfaction and identify technical problems (from outside the network system.)

## 30. Conclusion

We have shown that data mining can be used in a variety of performance improvement projects : by building models of a process, the user can identify root causes of a problem, and target those for corrective actions; he can find out when failures occur and which precise component in the process is guilty.

Telecommunication processes are complex; they usually generate large volumes of data and require huge teams and work-load for deployment and maintenance. Data mining can address such issues, provided it is able to handle these large volumes of data and to produce hundreds of models, very often in a limited time frame. In the projects we have presented, we have used KXEN because it has the ability to do just that !

## References


1. McKinsey (2001) "US Productivity Growth 1995-2000; understanding the contribution of Information technology relative to other factors". Report, McKinsey Global Institute.
2. Davenport, Thomas H., Harris, Jeanne G. (2007) Competing on Analytics: The New Science of Winning. Harvard Business School Press
3. Davenport, Thomas (2006) "Competing on analytics". Harvard Business Review, January.
4. Motorola University "Free Six Sigma Lessons, Lesson 1" http://www.motorola.com/content.jsp?globalObjectId=3069-5787#
5. West, Andreas & Bayer, Judy (2005) "Creating a Modeling Factory at Vodafone D2: Using Teradata and KXEN for Rapid Modeling". Teradata Conference, Orlando. *http://www.teradata.com/teradata-partners/conf2005/*
6. Douglas, Seymour (feb 2003) "Cox Communications Makes Profitable Prophecies with KXEN Analytic Framework" Product Review – KXEN Analytic framework. *DMReview Magazine*.
7. Vapnik, Vladimir (1995) "The Nature of Statistical Learning Theory". Springer.
8. Hornick, Mark F., Marcade, Erik, Venkayala, Sunil (2007) "Java Data Mining. Strategy, Standard, and Practice. A practical guide for architecture, design, and implementation". Morgan Kaufmann series in data management systems. Elsevier;
9. Fogelman Soulié, Françoise (2006) Data Mining in the real world. What do we need and what do we have ? KDD'06, Philadelphia, August 20, 2006. Workshop on Data Mining for Business Applications. 49-53, 2006. *http://labs.accenture.com/kdd2006_workshop/dmba_proceedings.pdf*

# A Process to Define Sequential Treatment Episodes for Patient Care

Patricia B. Cerrito
University of Louisville
Department of Mathematics
Louisville, KY 40292
502-852-6010
pcerrito@louisville.edu

## ABSTRACT

In this project, we develop a process to define episodes of patient treatment from claims data. There are two types of patient conditions. The first has well defined treatments and outcomes; the second condition requires a continuum of care. We examine the first type of treatment using an example of treatments for a heart condition requiring either angioplasty or bypass surgery. We want to examine the occurrence of multiple procedures, and the duration between procedures. We use the statistical software, SAS (SAS Institute; Cary, NC), to preprocess the data into treatment episodes, relying on time and cost markers to define episodes. In particular, we want to investigate the use of a new type of stent that was introduced to reduce the occurrence of multiple episodes of angioplasty. Once examined, it appears that the probability of repeat episodes is 12% with this new treatment.

## Categories and Subject Descriptors

J.3. [Life and Medical Sciences]: Health

## General Terms

Algorithms, Management, Measurement

## Keywords

Episode Groupers, Healthcare Applications

## 1. INTRODUCTION

Physicians make many different decisions to treat patients, especially those with multiple chronic illnesses. For example, there are many different medications for the treatment of Type II diabetes, and the physician chooses one or more of them for their patients. There is also the decision to start a patient on insulin; moreover, there is now a choice between insulin injections and inhaled insulin. A patient with blocked arteries can receive angioplasty, or bypass surgery. It is the accumulated consequences of these decisions that result in differing patient outcomes. In a continuum of treatment for chronic conditions, it is difficult to determine where one decision starts and ends. It is the purpose of this project to examine claims data to investigate sequential patterns of physician decision making by defining episodes of patient care.



We first need to preprocess the data to create treatment episodes to construct a sequence of care. We assume that episodes of treatment can be defined. Some treatments, for example, chemotherapy for cancer, can have a start date and an ending period with follow up so that recurrence begins a new treatment episode. However, for chronic conditions such as congestive heart failure and diabetic foot ulcers with chronic osteomyelitis, it is not clear when one treatment ends and another begins. Instead, the continuum of care should be considered.

We then define events in the sequence of treatment that suggest disease changes. We assume that the illness will change over time, for the better or for the worse, and that these markers can be used to examine treatment differences related to outcome. Consider, for example, Type II diabetes. One marker is the initial disease diagnosis followed by drug treatment. A second marker is a change in the type of medication, or the dose. A third marker would be a transition to insulin. We will use survival data mining to investigate the relationship between treatments and time to events for chronic diseases.

The next step is to construct a decision tree based on the analysis of the treatment sequence. We make the assumption that decision trees can be so constructed to examine the competing risks of different treatment sequences.

## 2. BACKGROUND

### 2.1. Drug-Eluting Stents and Physician Decision Making

The monitoring and examination of physician decision making remains in its infancy. Generally, physicians are surveyed concerning their use of medical and surgical procedures, and a consensus is reached to develop guidelines. Then, compliance with these consensus guidelines is examined with a "yes"-"no" result.

Another way to monitor physician decision making is to use observational data, comparing treatment methods. This is commonly done with observational studies that rely on patient charts. However, patient charts still exist mostly on paper, and are difficult to convert to datasets. Therefore, such studies tend to be small. Moreover, they tend to be for short periods of time.

Now that we have electronic billing records, claims data can be used to investigate the success of treatment by looking to future treatments that can result from the initial procedure. We show here an example of one such study that was performed to determine the effectiveness of the use of a new device to treat patients with blocked arteries. Clinical trials prior to approval focused on short-term benefit only; we investigate long-term benefit. As such, this project demonstrates a new approach to the use of claims and billing information, using linkage to connect multiple episodes to the same patient and using the information to examine recurrent time events. It demonstrates how claims data can be used to monitor treatment decisions continuously in relationship to patient outcomes.

Patients with blocked arteries in the heart can be treated with bypass surgery or with angioplasty.[1] Bypass surgery is very traumatic with very high cost. Veins and/or an artery are taken from other parts of the body and connected to the blocked heart artery as a means of diverting blood flow from the blockage through the new connection. The patient's sternum (breast bone) must be broken, and the patient's heart is stopped while the surgery is performed. The recovery period is considerable, as are the risks of mental or physical impairment. While there is a recently developed procedure that allows the heart to continue beating during surgery, it remains an uncommon choice because it has a high rate of error and few surgeons can perform it. The recovery period is considerable, as are the risks of impairment, stroke, and death during surgery when compared to the alternative procedure of angioplasty.

Angioplasty is a simpler procedure that uses a catheter inserted through the artery from the groin to the heart. The catheter has a balloon at the tip that is inflated at the blockage to expand the opening. Often a metal stent is inserted at the reopened artery to keep it open. However, there is a risk that clotting will occur around the stent, closing the artery. In most cases, this procedure is done on an outpatient basis and the patient can return to normal activities after a couple of days. However, there is also higher risk that the artery will re-block, requiring a repeat procedure when compared to the long-term effects of bypass surgery. The cost of angioplasty is considerably less when compared to the cost of bypass surgery.

In April, 2003, a "Teflon-coated" stent, called a drug-eluting stent, was approved for use by the Food and Drug Administration. The stent is coated with a drug that is meant to reduce the likelihood of re-blockage of the artery.[2] The drug is time-released into the blood stream for a period of time before it completely dissipates. Unfortunately, clinical trials tend to be of short term; the longevity of the new stent in preventing a blockage was not known at the time of FDA approval. Although this stent can cost three times as much as the regular stent, it was quickly approved for use by various insurance providers. It was anticipated that it would reduce the need for repeat procedures, and would reduce the use of bypass surgery.[3, 4]

FDA approval of the eluting stent provided no guidelines as to the optimal procedure. Therefore, each patient and physician is faced with the choice of procedure in the general absence of information concerning the long-term effects. While the eluting stent has much promise, only recently have papers appeared in the medical literature concerning the stent's severe side effects.[5] Other papers are appearing that still focus on time periods of less than one year, although one or two go just beyond one year. Because such retrospective analyses are difficult and expensive to perform, it is rare to look at long-term effects of 5 years or more duration. They are difficult, in part, because much of the patient record still exists in paper form, and the paper records must be compiled manually. Since the eluting stent just now has four years of approved use, we are now at the beginning of examining longer-term results.

Because of the expense of the procedures, a record of all procedures performed on a patient will exist in the claims data as long as the patient remains with the same insurer, whether the claims are from a private insurer, Medicare, Medicaid, or some other government program. Some commercial developers of data follow patients when they change insurers, so that all recurrences can be identified. These claims will be linked by a patient identifier. The existence of claims data makes it easier to examine multiple patient events. If a patient has a claim for a second procedure, we can identify the survival period of the first procedure. However, the complexity of claims data makes it difficult to extract the specific events because follow up will be labeled by the same event code. Therefore, we must examine the development of episode groupers that are used to target such events. Such episode groupers are commonly used to identify the total cost of a patient event; they are not yet in use to develop models of decision making.

## 2.2, Episode Groupers

Solutions under the general category of episode grouper have been developed specifically to fuse claims data. The methodology is difficult to find since it is mostly proprietary and little exists in the research literature.[6-8] A brief summary is given in Forthman, Dove and Wooster.[8] The main purpose of these groupers is to identify homogeneous groups of patients so that cost comparisons and summaries can be made. These "episode groupers" are used in analysis with little understanding as to how episodes are defined or how patients are grouped.[9-13] However, it is known that the groupers do not take into consideration the severity of an individual patient's condition.[11]

One method of grouping is to examine medications of a similar nature, and to define the end of an episode if there is at least one day between claims.[14] The Medicare Claims Processing Manual defines an episode of care as having a maximum time period of 60 days or until discharge, although episodes of care can be overlapping.[15] Another study defined episodes as 30-day periods while a third compared 4-month to 9-month absence of treatment as the end of an episode.[16, 17] There are still other definitions of episodes, including one per year.[18]

However, the main method used to define an episode of care is a variable timeframe, or "washout" period, with a continuous time period with an absence of treatment; that time period changes with the definition of the patient's condition.[19-21]

Unfortunately, it is not always clear just what that time period should be. For example, when a bone gets infected with a superbug known as MRSA, recurrence can occur up to a year after treatment is completed. Should this year be the definition of an episode, or should a period of say, six months be used to end the episode? One study that attempted to define an episode concluded that the duration was approximately 5 weeks for treatment of diabetic foot ulcers, excluding all patients who had a bone infection or amputation.[22] Yet most clinical studies of the same problem consider 8-12 weeks as a minimum for healing of the wounds, almost twice the length of the defined episode.[23-25] As another example, chronic diseases that are physician managed will have ongoing treatment if periodic testing and monitoring occur. In that case, an episode has to be defined differently for different treatments for the same patient condition.

Once a patient episode is defined, it is usually examined independently of other episodes for the same patient.[26] The main measure of an episode is its total cost.[27] However, that means that the likelihood that a treatment choice in episode one does lead to episode two is not examined.[28] In particular, we want to determine whether treatment choices lead to additional episodes of care. For example, suppose a treatment standard decreases inpatient stays from 5 days to 4 days, but at the cost of doubling the readmission rate.[29] Without examining the sequence of admissions, the 1-day reduction would be considered a cost effective outcome, especially if an episode is defined as the time from admission to discharge.[30] In addition to variability in patient response to treatments, there are competing risks that result in different choices of treatment made either by the physician or the patient.[31][26] Treatment variability is very characteristic of psychiatric treatment, even more so than for physician medicine.[32]

Another consideration is the treatment continuum itself, defined by compliance with and the continuity of care, especially to determine the effectiveness of disease management.[33, 34] We need to create a definition of compliance with care, and to rank compliance with treatment. We also need to ensure that all treatments (including prescribed medications) are included in the continuum, and are used to define episodes of care.

Relying on claims data, which is combined into one database from multiple sources, a date of care is included in each claim. However, if a patient is treated in the hospital, there can be several different physicians giving different types of care. It becomes a major challenge to relate these together into one episode. At a minimum, claims from the same episode should have the same diagnosis related group (DRG) code. These DRG codes are used universally in the USA for billing and for provider payment. However, this DRG code may be entered inaccurately. Claims for medications may not contain this code at all. Each claim will have a service date. We start by creating a clustering for each patient based upon date and DRG codes. Not every patient claim will be clustered successfully. From there, predictive modeling will be used for the unclustered values to predict membership into each cluster.[35, 36]

To examine the sequence of episodes, we will define a time series with multiple time endpoints. The initial time point will define the initial treatment and beginning of a chronic problem. The additional time points will be defined as either the end of the episode, or a change in condition, where the chronic illness gets better, or worse. We will use both fixed and dynamic regressors to investigate the patient outcomes. These regressors can represent a different medication, or a decision to perform surgery, or a change from outpatient to inpatient status. They can also represent a new, ongoing treatment. The fixed regressors will represent patient demographic information, and the initial severity of the patient's condition. The time series will be transactional in nature as the changes in treatment will not necessarily occur at fixed intervals. We will start by defining a time series for each patient, and then consolidating them into a series of outcomes. Once we have the likelihood of various outcomes defined by the time series, we can create a decision tree to look at the probability of each outcome given treatment choices.

In addition, it will be important to detect outliers either as they occur, or before they occur in terms of both cost and outcomes. Therefore, the claims data can also be considered streaming data, with changes in treatments indicative of future outcomes that can be costly either to payer or patient.

## 2.3. Data Preprocessing

Much of the data collected in databases nowadays is incomplete and noisy. This may be deliberate as, for example, when a patient refuses to provide an accurate date of birth or accidental as due to input error. Also, there is always the danger that data may be old or redundant. Thus, it is essential to researchers to base their analysis on what is described as "*clean data*". Cleaning data or preprocessing the data prior to mining is designed to eliminate the following anomalies:

1. Missing field values.
2. Outliers.
3. Obsolete and/or redundant data.
4. Data in clear contradiction of common sense or well established industry norms.
5. Data in inconsistent state or format.

It is estimated the 50-60% of researchers' time is spent in data preprocessing to create databases suitable for data mining. Thus, it is no surprise that data preparation is an integral phase of the data mining process as a whole. Defining episode groupers is a large part of the preprocessing of claims data.

It is also the case that data preprocessing requires an understanding of the data and of the statistical analysis that is necessary to manipulate the data in order to remove any anomalies.[37-39] Another issue in preprocessing is the need to define the observational unit. For example, the dataset might focus on individual claims from one inpatient hospital stay. However, there would be separate claims for the hospital, the physicians, the medications prescribed on discharge, and any home health care required. In order to examine the entire cost of

one visit, the observational unit must be changed from claim to inpatient process.

## 2.4. Data Fusion

Data fusion has been a trend in the field of imaging, text and signal analyses, and it is a combination of many disciplines. Communication and data management technologies focus on the organization, storage, preservation, and distribution of data. Mathematics, computer science, and artificial intelligence all contribute to the development of automatic and principled methods for combining, restructuring and summarizing diverse, incomplete and conflicting information. Data fusion covers an entire process: data gathering from multiple sources, data format conversion, data combination, conflict resolution, data summarization and distribution.[40, 41] The process takes input from heterogeneous sources and produces a coherent representation. Although multi-sensor data fusion is still not regarded as a formal professional discipline, tremendous progress has been made. The success of data fusion, and later data mining, depends as much on the adoption of appropriate methodologies and processes as it does on the availability of suitable data and the use of appropriate technology.

Medical data fusion is an emerging field which has recently experienced a tremendous increase in innovation. Progress and advances in medical imaging, medical signals, and an unstructured text fusion, have an immediate impact on commercial products and clinical practice. Today, various data modalities with completely different capabilities are available for diagnosis, intervention, surgery, or monitoring.[42] In multi-modal data registration, data of different modalities are transformed into a single coordinate system. Physicians get simultaneous access to the patient's data. In our project of episode groupers, we are working with claims from different sources but of similar format.

# 3. METHOD

## 3.1. Episode Groupers

We applied the proposed methodology to patients with blocked arteries, requiring either angioplasty, angioplasty with a stent, or bypass surgery. We wanted to determine sequential episodes to find the longevity of each procedure. In order to do this, we must first extract from domain knowledge the estimated cost of a procedure, and we assume that the appearance of such a cost in a claim will identify the start of a new episode. Subsequent claims that do not reach this defined threshold charge are then identified as follow up to the procedure. The data were de-identified according to HIPAA requirements. HIPAA is the Health Insurance Portability and Accountability Act of 1996 and stipulates how medical data can be accessed for research purposes.[43] The demographic information for each patient is very limited in such a de-identified dataset.

There are three major steps to the creation of a sequential episode grouper that can be generalized to any procedure once the threshold charges are identified.

Step 1.

We first isolate patient identifiers in the claims database that are related to the procedures under study. In this case, we want all claims indicating surgical procedures involving bypass or angioplasty. We identify these procedures through the use of the DRG codes. Again, DRG codes, or diagnosis related groups are a series of codes published by the Center for Medicare and Medicaid. They are now in universal use by all insurers in the USA for reimbursement to healthcare providers. Therefore, we can start by filtering a claims database to specific treatments as represented by DRG codes.

The DRG codes that define these procedures are often changed regularly, particularly with the introduction of a new procedure or medical device. Domain knowledge will identify the target DRG values. Table 1 gives the DRG codes for the comparison of angioplasty to bypass surgery. Table 2 gives the changes to DRG coding over the years after 2000.

**Table 1. DRG Codes for bypass and angioplasty**

| DRG Code | DRG Description |
|---|---|
| 106 | Coronary Bypass W PTCA |
| 547 | Coronary Bypass W/Cardiac Cath W/Major CV Dx |
| 548 | Coronary Bypass w/cardiac Cath W/O Major CV Dx |
| 549 | Coronary Bypass W/O Cardiac cath W/Major CV Dx |
| 550 | Coronary Bypass W/O Cardiac Cath W/O Major CV Dx |
| 555 | Percutaneous Cardiovascular Procedure W/ major CV Dx |
| 556 | Percutaneous Cardiovascular Procedure W/non-drug-eluting stent W/O major CV Dx |
| 557 | Percutaneous Cardiovascular Procedure W/drug-eluting stent W/major CV Dx |
| 558 | Percutaneous Cardiovascular procedure W/drug-eluting stent W/O major CV Dx |
| 518 | Percutaneous Cardiovascular proc w/o AMI w/o coronary artery stent implant |

**Table 2. Changes to DRG Coding**

| DRG Codes before 10/01/2005 | DRG codes after 10/01/2005 |
|---|---|
| 547 and 548 | 107 coronary bypass w/cardiac cath |
| 549 and 550 | 109 coronary bypass w/o major cath |
| 555 | 516 precutaneous cardiovascular proc w/ AMI |
| 556 | DRG 517 percutaneous cardiovascular proc`w/o AMI, w/coronary artery stent implant |
| 557 | DRG 526 percutaneous cardiovascular proc w/drug-eluting stent w/ AMI |
| 558 | 527 precutaneous cardiovascular proc w/drug-eluting stent w/o AMI. |

The SAS code to isolate these variables is

```
PROC SQL;
 CREATE TABLE
SASUSER.QUERY_FOR_HIGHINFECTION_0000
AS SELECT dataset FORMAT=$255.,
        dataset.variable1
FORMAT=BEST12.,
        highinfection.variable2,
        highinfection.variable3
       .
       .
       .
       dataset.variablen
WHERE dataset.drgcodes IN ("DRG1",
"DRG2",… "DRGn");
QUIT;
```

Step 2.

Once filtered, we use domain knowledge to determine the potential length of an episode, and the threshold cost of an episode. Fortunately, both angioplasty and bypass tend to have relatively short duration, with the probability of exceeding 30 days as an inpatient so small that it can be discarded. In addition, the two procedures have very high cost. Therefore, we use a time period of 30 days and a threshold charge of $20,000. Other problems will have more of a continuum of treatment, and the episode will be more difficult to define. In that way, we eliminate follow up visits to focus on the main procedures.

We next use the high performance forecasting procedure to bin the patient claims into 1-month intervals. Total charges of each episode are accumulated for each 1-month period for each patient. We then eliminate any summed claims that do not exceed our defined threshold amount. The code used to bin the values is

```
proc hpf data=sasuser.dataset
out=sasuser.episodegroup1 lead=0;
id claimdate interval=month
accumulate=total;
forecast _all_;
by  patientidentifier;
run;
data sasuser.episodegroup2;
set sasuser.episodegroup1;
where total_charge>20000;
run;
```

Before continuing with the definition of the episode grouper, once we have accumulated by patient identifier, we summarize to get the number of procedures per month using the code:

```
PROC MEANS
DATA=sasuser.episodegrouper2
       FW=12
       PRINTALLTYPES
       CHARTYPE
       QMETHOD=OS
       NWAY
       VARDEF=DF
              MEAN
              STD
              MIN
              MAX
              N
              Q1
              MEDIAN
              Q3     ;
       VAR list variables here;
       CLASS month combinedDRG /
       ORDER=UNFORMATTED ASCENDING;
OUTPUT OUT=SASUSER.groupersummary
       MEAN()=
       N()=
       MAX()=
       SUM()=

RUN;
```

It is clear that by May, 2003, shortly after approval of the eluting stent, that the number of regular stent procedures has declined dramatically; however, the number of bypass procedures declined only slightly, and started to increase by January, 2006. Therefore, it appears that the use of the eluting stent had little impact on the use of bypass surgery. However, it has dramatically reduced the use of the regular stent. This shift has increased costs; we look to see if there is a corresponding increase in patient benefit.

Step 3

We divide identifiers into two subsets; those with only one inpatient stay and those with more than one procedure. Once separated, we place a code of "1" on the first subset to represent censored data. Similarly, we place a code of "0" on the second subset to represent uncensored data. Censoring means that a possible second episode was not observed because the study period ended. The analysis differs from the standard survival analysis in that there can be multiple events occurring over time, with multiple recurrences.

PROC Transpose is used to shift the values so that each patient identifier has just one observation in the dataset.

```
proc transpose
data=sasuser.episodegroup2
out=sasuser.transposedata
prefix=procedure_;
var claimdate;
by patientidentifier;
run;
```

## 4.2. Survival Data Mining

The code in the next portion of the program is used to define a censoring variable for the second episode, assuming that the first episode is the initial time point. In addition, a second survival function is defined for those patients who have a third episode. If the value is censored, the final date recorded is at the end of the data reference period, in this case, December 31, 2004.

```
data sasuser.censor;
set sasuser.transposedata;
lastdate='31dec2004'd;
if (procedure_2 = '.') then censor=0;
else censor=1;
```

The next code listed here computes the difference in date between the first and second episodes, and between the second and third episodes, if these episodes exist. Otherwise, the difference is defined as censored. We can continue the pattern beyond the third episode if needed.

```
if (censor=1) then
time1=datdif(procedure_1,procedure_2,
'act/act');
else
time1=datdif(procedure_1,lastdate,'ac
t/act');
if (censor=1 and procedure_3='.')
then censor2=0;
if (censor=1 and procedure_3 ne '.')
then censor2=1;
if (censor2=1) then
time2=datdif(procedure_2,procedure_3,
'act/act');
if (censor2=0) then
time2=datdif(procedure_2,lastdate,'ac
t/act');
run;
```

As traditional survival analysis cannot be used, we turn toward survival data mining. The process is also called multivariate failure time, or recurrent events data. The model used here was developed by Wei, Lin, and Weissfeld. The technique has been developed primarily to examine the concept of customer churn, again where multiple end points exist.[44, 45] However, medical use of survival is still generally limited to one defined event, although some researchers are experimenting with the use of predictive modeling rather than survival analysis. [46-52] Nevertheless, in a progressive disease, the event markers of that progression should be considered. However, we first construct survival curves to visualize the recurrence outcome of the procedure by using the variables for time1 and time 2. We also limit the starting procedure, or stratify by starting procedure to compute comparisons. We can also use this code to determine the proportion of patients with only one procedure during the period of study. This code is

```
PROC LIFETEST DATA=sasuser.censor
      ALPHA=0.05

;
      STRATA DRG;
      TIME Time1 * Censor (0);
      Run;
```

Once we have identified the episodes, we want to use a model for recurrent events data, or survival data mining. The intensity model is given by

$$\lambda_Z(t)dt = E\{dN(t) \mid F_{t-}\} = \lambda_0 e^{\beta' Z(t)} dt$$

where $F_t$ represents all the information of the processes N and Z up to time t, $\lambda_0(t)$ is an arbitrary baseline intensity function, and $\beta$ is the vector of regression coefficients. The instantaneous intensity is the hazard rate. This model has two components: 1) the covariates assume multiplicative effects on the instantaneous rate of the counting process and 2) the influence of prior events on future recurrences is mediated through the time-dependent covariates. The hazard rate function is given by

$$d\mu_Z(t) = E\{dN(t) \mid Z(t)\} = e^{\beta' Z} \mu_0(t)$$

where $\mu_0(t)$ is an unknown continuous function and $\beta$ is the vector of regression parameters. We compute the estimates of the regression coefficients by solving the partial likelihood function. A subject with K events contributes K+1 observations to the input data set. The $k^{th}$ observation of the subject identifies the time interval from the $(k-1)^{th}$ event or time 0 (if k=1) to the $k^{th}$ event, k=1,2,…,K. The $(K+1)^{th}$ observation represents the time interval from the $K^{th}$ event to time of censorship. In order to define these values, we need to unstack the sasuser.censor dataset defined previously.

The following code is used to investigate multiple occurrences. The first step is to create the necessary time variables, that is, the values of k-1 and k in the dataset.

```
Data sasuser.dmsurvival;
Set sasuser.censor;
Newcensor=censor; drg=drg1;
newtime=time1; oldtime=0; output;
Newcensor=censor2; drg=drg2;
newtime=time2; oldtime=time1; output;
Run;
```

In this code, we want to determine whether there are statistically significant differences between DRG codes in terms of time to a repeat procedure. We use Cox Regression, also called a proportional hazards model with the phreg procedure (phreg for proportional hazards). [1]

```
Proc phreg data=sasuser.dmsurvival;
Model=(oldtime,newtime)*censor(0)=drg
;
Run;
```

The result is statistically significant ($p<0.0001$). There are differences in outcomes of the different procedures. The methodology can be applied to other procedures that have major inpatient treatments that have considerable cost and that can be used to define the start of a new episode in claims data. The methodology has also been applied successfully to an examination

of breast cancer treatments and to the use of physical rehabilitation as a means to avoid orthopedic surgery.

The SAS code outlined here has been generalized so that it can be used on any comparison of procedures for which a threshold amount can be used to define the start date of a new episode.

# 4. RESULTS

## 4.1. Use of Procedures

Before we examined sequential episodes of treatment, we first wanted to examine whether there was a shift in procedures from regular stents and bypass surgery to the eluting stent. Figure 1 gives the relationship of the two types of stents by time period. Figure 2 gives the relationship to bypass procedures. The shift from the regular stent to the eluting stent was immediate after approval; there was no shift away from bypass surgery.

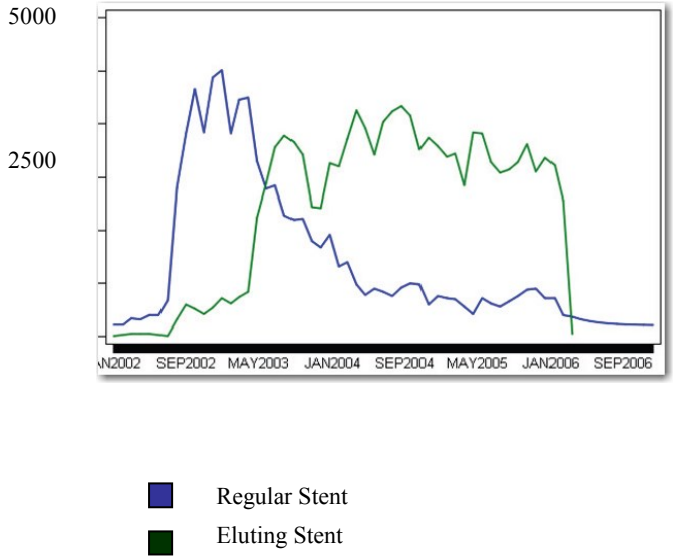



**Figure 1. Number of Procedures by Time**

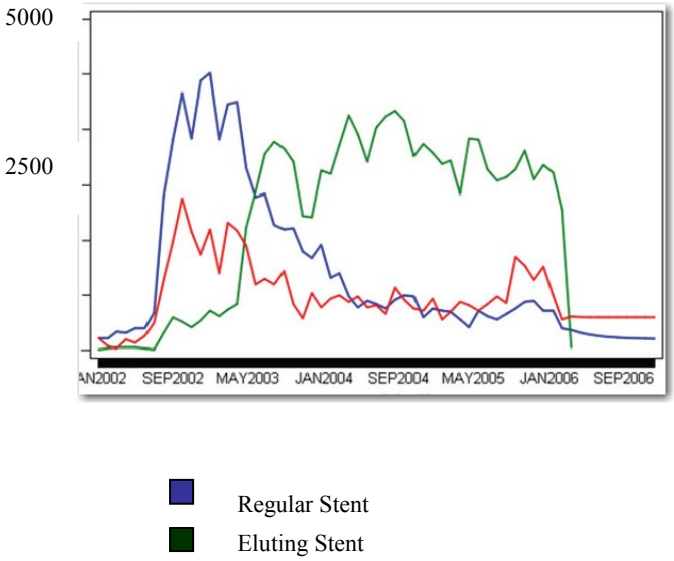



**Figure 2. Number of Procedures by Time With Bypass**

## 4.2. Recurrence of Procedures

Once the data are filtered by DRG, there remain approximately 270,000 total claims. Once binned, Table 3 gives the number of episodes by DRG code. We also specify that an episode has to have a minimum cost so that we exclude all follow up patient events that are related to the initial episode. There are almost 22,000 total episodes, of which 14,006 (approximately 2/3) are initial events without any recurrences. It also gives the percentage by DRG for repeat episodes.

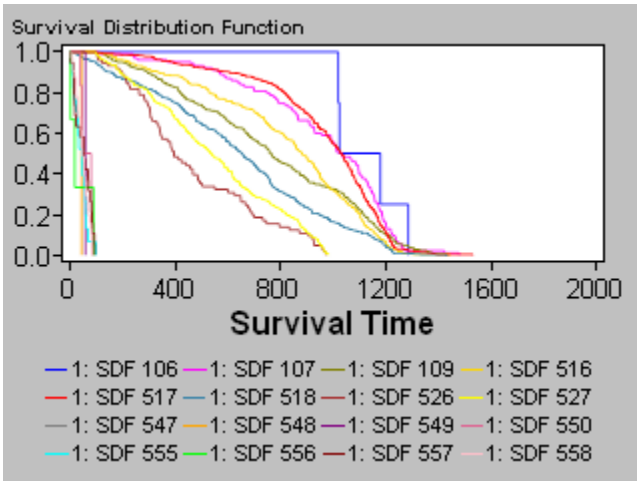
As expected, patients with DRG codes 517 and 518, angioplasty with and without a traditional stent have a high rate of repeats. However, patients with DRG code 527 with the newly developed drug-eluting stent also has a high rate of repeat episodes. This was unexpected. We construct survival functions for the second treatment episode, assuming the first episode as the initial time=0 point. Figure 3 gives the survival curve from the first episode to the second. Figure 4 gives the survival curve from the second episode to the third, for those who have a second episode.

**Table 3. Episodes by DRG**

| drg | Total | Failed | Percent Failed |
|---|---|---|---|
| 106 | 89 | 4 | 4.49 |
| 107 | 1628 | 95 | 5.84 |
| 109 | 1679 | 234 | 13.94 |
| 516 | 1249 | 260 | 20.82 |
| 517 | 1305 | 442 | 35.87 |
| 518 | 1776 | 434 | 24.44 |
| 526 | 888 | 68 | 7.66 |
| 527 | 2157 | 318 | 14.74 |
| 547 | 87 | 0 | 0 |
| 548 | 64 | 1 | 1.56 |
| 549 | 63 | 1 | 1.59 |
| 550 | 129 | 6 | 4.65 |
| 555 | 149 | 16 | 10.74 |
| 556 | 20 | 3 | 15.00 |
| 557 | 274 | 20 | 7.30 |
| 558 | 328 | 31 | 9.45 |

The survival curves are somewhat misleading because some codes are not commonly used and have been discontinued. Nevertheless, the curves show that bypass has a higher probability of a longer time to recurrence.

**Figure 3. Survival Curves for DRG Procedures for First Recurrence**




We combine the different DRG codes to indicate bypass (106, 107, 109, 547-550), regular stent (516, 517, 518, 555,556), or the eluting stent (526, 527, 557, 558). Table 4 gives the failure rate for the three major procedures from the first episode to the second; Table 5 gives the failure rate from the second to third procedure.. Similarly, Figures 5 and 6 give the survival curves when the treatments are reduced to just the three types.

**Figure 4. Survival Curves for DRG Procedures for Second Recurrence**

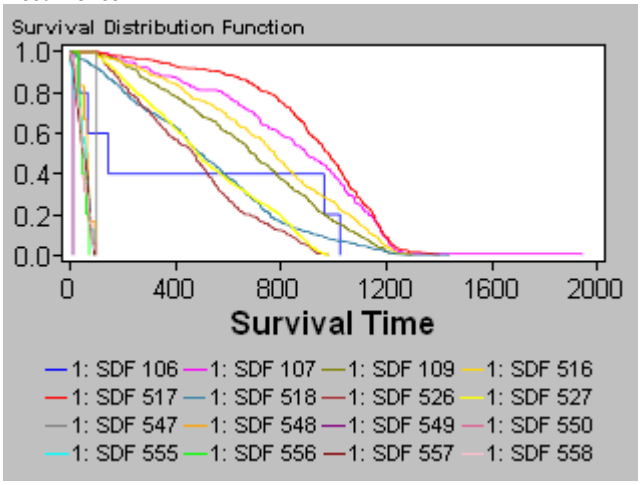



**Table 4. Failure Rate by Major Procedure for First Recurrence**

| Treatment Type | Total | Failed | Percent Censored |
|---|---|---|---|
| bypass | 3739 | 341 | 9.12 |
| Drug eluting stent | 3647 | 437 | 11.98 |
| stent | 4499 | 1155 | 25.67 |

The recurrence for the drug eluting stent appears to have a shorter time period even with the lower recurrence. It is more because of the more limited time period it is under study; that is, a shorter time to censoring.

**Figure 5. Survival Curves for Procedures for First Recurrence**

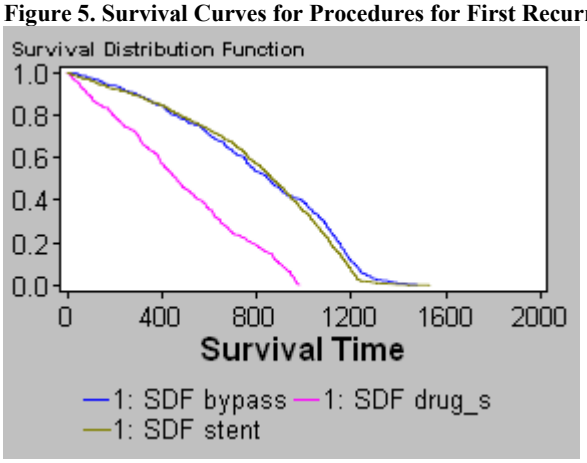
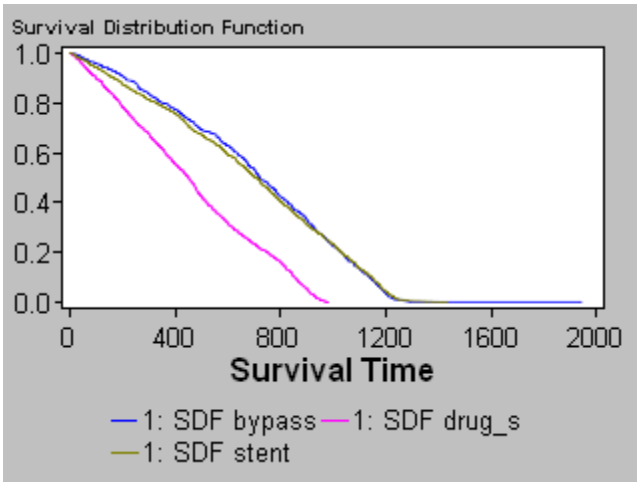



**Table 5. Failure Rate by Major Procedure for Second Recurrence**

| Treatment Type | Total | Failed | Percent Censored |
|---|---|---|---|
| Bypass | 3316 | 598 | 19.03 |
| Drug eluting stent | 3157 | 1473 | 46.66 |
| Stent | 3135 | 1405 | 44.82 |

However, for the second recurrence, the failure rate for the drug eluting stent is approaching that of the regular stent. Therefore, it appears that a patient who has a failure with the drug eluting stent should give very serious thought to shifting to a bypass procedure.

**Figure 6. Survival Curves for Procedures for Second Recurrence**




A total of six patients who had an initial angioplasty with a drug-eluting stent also failed, and had a bypass procedure. This process can be extended to any of the other initial procedures to compute the probabilities of second and third episodes. Once computed, the probabilities can be used to develop a decision tree.

The final step in this analysis will be to develop a Markov model outlining the potential risks of each of the physician choices as to procedure.

## 5. DISCUSSION

In a future study, we extend the definition of sequential episode grouper by applying it to the more complex issue of osteomyelitis in patients with diabetes. Patients with diabetes are at high risk of developing diabetic foot ulcers. If the ulcers get infected, especially with the bacteria, MRSA, the risk of developing osteomyelitis and subsequent amputation are also quite high. The longer it takes to heal the ulcers, the greater the probability of infection. Moreover, the choice of antibiotic treatment, and its duration are also directly related to success in healing. Because the treatment of foot ulcers and osteomyelitis are ongoing as the conditions, once started, often become chronic, we need to examine the totality of care. Unfortunately, in the past, each episode has been considered independent of other episodes. Therefore, the development of a sequential treatment pathway is invaluable to determine which pathways have a higher risk of amputation compared to others.

Because of the prevalence of MRSA, it is an enormous health concern. It is considered a "superbug" and is extremely difficult to eradicate.[53] Treatment costs are very high.[54] Inadequate use of antibiotics can lead to patient mortality, and the percentage of patients receiving inadequate antibiotics can be substantial.[55] One study found that out of 284 infections with MRSA, only 25 patients received appropriate antibiotics.[56] There are suggestions that in vitro susceptibilities to antibiotics can result in clinically unreliable treatments; antibiotic choice is related to recurrence of the infection.[57] Failure to eradiate the infection can lead to amputation, particularly for patients with diabetic foot ulcers.[58] While guidelines are available for treatment, they do not tend to commit to specific antibiotics even in the case of osteomyelitis.[59, 60]

New antibiotic treatments for MRSA are continually under development, and physicians need to keep up with current development in order to treat the infection effectively.[61] Linezolid was approved in 2001 for use with MRSA, and has been approved for use for skin infections and for pneumonia.[62] In clinical trials, linezolid has a higher cure rate and usually shorter hospitalization compared to the more standard vancomycin, resulting in lower overall treatment costs.[63-65] Linezolid is also used in combination with the antibiotics, rifampicin and vancomycin to fully eradicate the infection.[66] While vancomycin treatment has been the "gold standard", the failure rate is climbing close to 50% of infections resulting in mortality and amputations.[57, 67] There are some who suggest that antibiotics should be used without surgery and amputation as a first-line treatment, but where specific antibiotics are suggested but not exactly specified. Still others regard osteomyelitis as a surgical issue.[68] According to Jeffcoate and Lipsky, "The optimal approach to diagnosing and managing osteomyelitis of the foot in diabetes in unclear."[69]

Distinctions are made between treatment of community-acquired skin infections, where Septra is effective, versus nosocomial MRSA.[70] In other such infections, some antibiotics in common use are not effective.[71] Rifampicin and sodium fusidate are used to eliminate MRSA in carriers.[72] However, this distinction does not yet occur with osteomyelitis with MRSA, which still has treatment generally limited to vancomycin and Linezolid.[73]

Because of the continued resistance and difficulty in treatment, along with information that treatments are sometimes inadequate, the occurrence and treatment of MRSA should be carefully and continually monitored.[74] However follow up does tend to encourage longer treatment with antibiotics beyond the traditional 6 weeks for osteomyelitis.[75] Nevertheless, current physician practice tends to limit treatment to this traditional 6 weeks.[73, 76, 77] This limitation is problematic since most studies of Linezolid and osteomyelitis favor 12-15 weeks of treatment.[78, 79]

Since vancomycin is usually administered intravenously (IV), it is not always on a pharmacy drug formulary, but is readily available if needed to treat MRSA and exists in the hospital formulary. However, it requires hospitalization or home health services for administration. It lowers the patient's quality of life because of the need for a semi-permanent IV line. Linezolid, because it is not available in a generic equivalent, usually requires pre-authorization or is in the formulary third tier requiring the highest co-payment from the patient. Since drug benefits can be separated from hospital benefits in terms of insurance provider, there are often two different oversight agencies responsible for approving treatments.[80, 81] In particular, the British National Formulary allows the use of Linezolid for skin infections, but not for the treatment of osteomyelitis (access requires registration).[82] For this reason, it is used as a secondary treatment when vancomycin fails.[83]

An initial cost-effectiveness analysis when linezolid was first introduced suggests that linezolid is less costly as a primary rather than secondary treatment for soft tissue wounds.[84, 85] A cost effectiveness analysis for patients with diabetes and osteomyelitis was published in JAMA in 1995 using a Markov model.[86]

However, since 1995, new antibiotics have been developed that are effective in the treatment of osteomyelitis. Also, the cost model relied exclusively on data in the medical literature instead of using actual treatment data available. Moreover, the prevalence of MRSA was not as great.[87]

We have a dataset of the National Inpatient Sample (NIS) containing treatment information concerning osteomyelitis from the years 2000-2004; we discovered

- 117,000 records with a DRG of osteomyelitis or amputation in the lower limbs (DRG codes 113, 114, 213, 238 and 285)
- 48,932 records with primary or secondary diagnosis of osteomyelitis
- 29,136 cases of osteomyelitis were treated with amputation

These numbers indicate that the problem of osteomyelitis is substantial. Moreover, the primary method of treatment of osteomyelitis in the lower limbs appears to be amputation. However, there can be recurrence of the infection after amputation:

- 2788 of the 48,932 cases of osteomyelitis had previous amputation
- 4030 had an infection resistant to some antibiotics
- 3913 had MRSA

For treating the infection:

- 3176 had antibiotic infusions
- 10 had linezolid infusions

Because there is no possibility of follow up using the NIS database, it is not known whether antibiotic infusions were treated using home health. Of the patients with amputation,

- 96 received antibiotic infusions
- None received linezolid

It is clear from these summary results that amputation is the primary method of treatment of osteomyelitis in patients with diabetes. We intend to examine multiple events in sequence:

- Occurrence of foot ulcers
- Occurrence of osteomyelitis
- Amputation (toe, foot, leg below knee, leg above knee)
- Complications resulting from amputation
- Treatment with antibiotics

to determine whether amputation is progressive-from toes to feet to legs. Since wounds can take years to heal in the lower extremities, it becomes much more difficult to define the episodes of treatment. It is our intent to use a combination of cost and washout periods (when no treatment is provided) to define the treatment groups. Once defined, they will be defined sequentially, and survival data mining will be applied.

Physicians tend to be autonomous in their decision making, especially in the absence of treatment guidelines. Variability in decision making can lead to variability in patient outcomes. Only by comparing outcomes across physicians can optimal treatment pathways be discovered.

## 6. ACKNOWLEDGMENTS

We want to thank John Cerrito, PharmD and Glenn Lambert, MD, for their support in the development of this paper, which was supported in part by NIH grant #1R15RR017285-01A1, Data Mining to Enhance Medical Research of Clinical Data.

## 7. REFERENCES

1. Anonymous, *Guidant Patients and Families*. 2006, Guidant, Inc.
2. Anonymous, *Drug-eluting stent overview*. 2007, angioplasty.org.
3. Llera, L.-S.D.d.l., et al., *Sirolimus-eluting stents compared with standard stents in the treatment of patients with primary angioplasty.* American Heart Journal, 2007. **134**: p. 164.e1-164.e6.
4. Tsuchida, K., et al., *The clinical outcome of percutaneous treatment of bifurcation lesions in multivessel coronary artery disease with the sirolimus-eluting stent: insights from the Arterial Revascularization Therapies Study part II.* European Heart Journal, 2007. **28**: p. 433-442.
5. Katayama, T., et al., *Two cases of very late stent thrombosis after implantation of a sirolimus-eluting stent presenting as AMI.* European Journal of Cardio-Thoracic Surgery, 2007. **48**(3): p. 393-397.
6. Rosen, A. and A. Mayer-Oakes, *Episodes of care: theoretical frameworks versus current operational realities.* Joint Commission on Quality Improvement, 1999. **25**(3): p. 111-38.
7. Claus, P., et al., *Clinical care management and workflow by episodes.* Proceedings AMIA Annual Fall Symposium, 1997. **1997**: p. 91-5.
8. Forthman, M.T., H.G. Dove, and L.D. Wooster, *Episode treatment groups (ETGs): a patient classification system for measuring outcomes performance by episode of illness.* Top Health Information Management, 2000. **21**(2): p. 51-61.
9. Wan, G., et al., *Healthcare expenditure in patients treated with vaniafaxine or selective serotonin reuptake inhibitors for depression and anxiety.* Internation Journal of Clinical Practice, 2002. **56**(6): p. 434-9.
10. Kerr, E., et al., *Measuring antidepressant prescribing practice in a healthcare system using administrative data: implications for quality measurement and improvement.* The Joint Commission Journal on Quality Improvement, 2000. **265**(4): p. 203-16.
11. Thomas, J.W., *Should episode-based economic profiles be risk adjusted to account for differences in patients' health risks?* Health Research and Educational Trust, 2005. **April, 2006**: p. 581-590.
12. Currie, C.J., et al., *The financial costs of hospital care for people with diabetes who have single and multiple macrovascular complications.* diabetes Research and Clinical Practices, 2005. **67**: p. 144-151.

13. Bassin, E., *Episodes of care: a tool for measuring the impact of healthcare services on cost and quality.* Disease Management & Health Outcomes, 1999. **6**: p. 319-325.
14. Bonetto, C., M. Nose, and C. Barbui, *Generating psychotropic drug exposure data from computer-based medical records.* Computer Methods and Programs in Biomedicine, 2006. **83**: p. 120-124.
15. Anonymous, *Medicare Claims Processing Manual: Chapter 10, Home Health Agency Billing*. 2006, Health and Human Services.
16. Ritzwoller, D.P., et al., *The association of comorbidities, utilization and costs for patients identified with low back pain.* BMC Musculoskeletal Disorders, 2006. **7**: p. 1-10.
17. Thomas, J.W., *Economic profiling of physicians: does omission of pharmacy claims bias performance measurement?* American Journal of Managed Care, 2006. **12**: p. 341-351.
18. Hong, W., et al., *Medical-claims databases in the design of a health-outcomes comparison of quetiapine.* Schizophrenia Research, 1998. **32**(1): p. 51-58.
19. Anonymous, *episode treatment groups*. 2006. p. 1-8.
20. Claus, P.L., et al., *Clinical care management and workflow by episodes*. 1997.
21. Hall, D.L. and J. Llinas, *Handbook of Multisensor Data Fusion*. 2001, Cleveland: CRC.
22. Mehta, S., et al., *Determining an episode of care using claims data: diabetic foot ulcer.* Diabetes Care, 1999. **22**(7): p. 1110-1115.
23. Ling, X., et al., *Bacterial load predicts healing rate in neuropathic diabetic foot ulcers.* Diabetes Care, 2007. **30**(2): p. 378-380.
24. Sheehan, P., et al., *Percetn change in wound area of diabetic foot ulcers over a 4-week period is a robust predictor of complete healing in a 12-week prospective trial.* Plastic and Reconstructive Surgery, 2006. **117**(Suppl): p. 239S-244S.
25. Jude, E., et al., *Prospective randomized controlled study of Hydrofiber dressing containing ionic silver or calcium alginate dressings in non-ischaemic diabetic foot ulcers.* Diabetic Medicine, 2006. **24**: p. 280-288.
26. Jonsson, L., B. Bolinder, and J. Lundkvist, *Cost of hypoglycemia in patients with Type 2 diabetes in Sweden.* Value in Health, 2006. **9**(1): p. 193-198.
27. Peltokorpi, A. and J. Kujala, *Time-based analysis of total cost of patient episodes.* International Journal of Health Care Quality Assurance, 2006. **19**(2): p. 136-143.
28. Horn, S.D., *Quality, clinical practice improvement, and the episode of care.* Managed Care Quarterly, 2001. **9**(3): p. 10-24.
29. Koh, H. and S. Leong, *Data mining applications in the context of casemix.* Annals of the Academy of Medicine, Singapore, 2001. **30**(4 Suppl): p. 41-9.
30. Kujala, J., et al., *Time-based management of patient processes.* Journal of Health Organization and Management, 2006. **20**(6): p. 512-524.
31. Keen, J., J. Moore, and R. West, *Pathways, networks and choice in health care.* International Journal of Health Care Quality Assurance, 2006. **19**(1): p. 316-327.
32. Singh, S.P. and T. Grange, *Measuring pathways to carei n first-episode psychosis: a systematic review.* Schizophrenia Research, 2005. **81**: p. 75-82.
33. Greenberg, G.A. and R.A. rosenheck, *Continuity of care and clinical outcomes in a national health system.* Psychiatric Services, 2005. **56**(4): p. 427-433.
34. Solz, H. and K. Gilbert, *Health claims data as a strategy and tool in disease management.* Journal of Ambulatory Care Management, 2001. **24**(2): p. 69-85.
35. Xue, S. *A fault diagnosis system based on data fusion algorithm*. in *First international conference on innovative computing information and control*. 2006. Beijing, China.
36. Putten, P.v.d., J.N. Kok, and A. Gupta, *Data fusion through statistical matching*. 2002.
37. Zhu, X., X. Wu, and Q. Chen, *Bridging local and global data cleansing: identifying class noise in large, distributed data datasets.* Data Mining and Knowledge Discovery, 2006. **12**(2-3): p. 275.
38. Wong, K., et al., *A taxonomy of dirty data.* Data Mining and Knowledge Discovery, 2003. **7**: p. 81-99.
39. Hernandez, M.A. and S.J. Stolfo, *Real-world data is dirty: data cleansing and the merge/purge problem.* Data Mining and Knowledge Discovery, 1998. **2**: p. 9-17.
40. Makela, T., *Data registration and fusion for cardiac applications*. 2003, University of Helsinki: Helsinki.
41. Upstill, C., et al., *Infectious diseases: preparing for the future*. 2006, Foresight Science Reviews: United Kingdom.
42. Denzler, J. *Sensor data and information fusion in computer vision and medicine, Executive Summary*. in *Dagstuhl Seminar Proceedings*. 2007. Germany.
43. Anonymous, *HIPAA Overview*. 2007, Centers for Medicare & Medicaid Services.
44. Potts, W., *Survival Data Mining*. 2000.
45. Linoff, G.S., *Survival Data Mining for Customer Insight*. 2004, Intelligent Enterprise.
46. Xie, H., T.J. Chaussalet, and P.H. Millard, *A model-based approach to the analysis of patterns of length of stay in institutional long-term care.* IEEE Transactions on information technology in biomedicien, 2006. **10**(3): p. 512-518.
47. Shaw, B. and A.H. Marshall, *Modeling the health care costs of geriatric inpatients.* IEEE Transactions on information technology in biomedicien, 2006. **10**(3): p. 526-532.
48. Pinna, G., et al., *Determinant role of short-term heart rate variability in the prediction of mortality in patients with chronic heart failure.* IEEE Computers in Cardiology, 2000. **27**: p. 735-738.
49. Berzuini, C. and C. Larizza, *A unified approach for modeling longitudinal and failure time data, with application in medical monitoring.* IEEE Transactions on pattern analysis and machine intelligence, 1996. **16**(2): p. 109-123.
50. Eleuteri, A., et al. *Survival analysis and neural networks*. in *2003 Conference on Neural Networks*. 2003. Portland, Oregon.
51. Seker, H., et al. *An artificial neural network based feature evaluation index for the assessment of clinical factors in breast cancer survival analysis*. in *IEEE Canadian Conference on Electrical & Computer Engineering*. 2002. Winnipeg, Manitoba.
52. John, T.T. and P. Chen, *Lognormal selection with applications to lifetime data.* IEEE Transactions on reliability, 2006. **55**(1): p. 135-148.
53. Karaolis, D.K., et al., *c-di-GMP (2'-5'-Cyclic diguanylic acid) inhibits staphylococcus aureus cell-cell interactions and biofilm formation.* Antimicrobial agents and chemotherapy, 2005. **49**(3): p. 1029-1038.

54. Watters, K., T. O'Dwyer, and H. Rowley, *Cost and morbidity of MRSA in head and neck cancer patients: what are the consequences?* The Journal of Laryngology, 2004. **118**: p. 694-699.
55. Osmon, S., et al., *Hospital mortality for patients with bacteremia due to staphylococcus aureus or pseudomonas acruginosa.* Chest, 2004. **125**(2): p. 607-616.
56. Paydar, K.Z., et al., *Inappropriate antibiotic use in soft tissue infections.* Archives of Surgery, 2006. **141**(9): p. 850-6.
57. Tice, A.D., P.A. Hoaglund, and D.A. Shoultz, *Risk factors and treatment outcomes in osteomyelitis.* Journal of Antimicrobial Chemotherapy, 2003. **51**: p. 1261-1268.
58. Giurato, L. and L. Uccioli, *The diabetic foot: Charcot joint and osteomyelitis.* Nuclear Medicine Communications, 2006. **27**(9): p. 745-9.
59. Lipsky, B.A., et al., *Diagnosis and treatment of diabetic foot infections.* Clinical Infectious Diseases, 2004. **39**: p. 885-910.
60. Zgonis, T. and T.S. Roukis, *A systematic approach to diabetic foot infections.* Advances in Therapy, 2005. **22**(3): p. 244-62.
61. Jauregui, L.E., et al., *Randomized, double-blind comparison of once-weekly Dalbavancin versus twice-daily linezolid therapy for the treatment of complicated skin and skin structure infections.* Clinical Infectious Diseases, 2005. **41**: p. 1407-15.
62. Peppard, W.J. and J.A. Weigeit, *Role of linezolid in the treatment of complicated skin and soft tissue infections.* Expert Review of Anti-infective therapy, 2006. **4**(3): p. 357-366.
63. McKinnon, P.S., et al., *Impact of linezolid on economic outcomes and determinants of cost in a clinical trial evaluating patients with MRSA complicated skin and soft-tissue infections.* Annals of Pharmacotherapy, 2006. **40**(6): p. 1017-23.
64. Itani, K.M., et al., *Linezolid reduces length of stay and duration of intravenous treatment compared with vancomycin for complicated skin and soft tissue infections due to suspected or proven methicillin-resistant Staphylococcus aureus (MRSA).* International Journal of Antimicrobial Agents, 2005. **26**(6): p. 442-448.
65. Cunha, B., *Oral antibiotic treatment of MRSA infections.* Journal of Hospital Infection, 2005. **60**(1): p. 88-90.
66. Wenisch, C., et al., *A holistic approach to MRSA eradication in critically ill patients with MRSA pneumonia.* Infection Control and Hospital Epidemiology, 2006. **34**(3): p. 148-154.
67. Wunderink, R., *Stakes, treatment strategies and progression of MRSA nosocomial pneumonia, especially pneumonia due to mechcanical ventilation.* Presse medicale, 2004. **33**(12 pt 2): p. 255-9.
68. Henke, P.K., et al., *Osteomyelitis of the foot and toe in adults is a surgical disease.* Annals of Surgery, 2005. **241**(6): p. 885-894.
69. Jeffcoate, w.J. and B.A. Lipsky, *Controversies in diagnosing and managing osteomyelitis of the foot in diabetes.* Clinical Infectious Diseases, 2004. **39**(Supplement 2): p. S115-22.
70. Kaka, A.S., et al., *Bactericidal activity of orally available agents against methicillin-resistant Staphylococcus aureus.* Journal of Antimicrobial Chemotherapy, 2006. **58**: p. 680-683.
71. Jang, C.H., C.-H. Song, and P.-C. Wang, *Topical vancomycin for chronic suppurative otitis media with methicillin-resistant Staphylococcus aureur otorrhoea.* The Journal of Laryngology, 2004. **118**: p. 645-647.
72. Garske, L., et al., *Rifampicin and sodium fusidate reduces the requency of methicillin-resistant Staphylococcus aureus (MRSA) isolation in adults with cystic fibrosis and chronic MRSA infection.* Journal of Hospital Infection, 2003. **56**(3): p. 208-214.
73. Zahedi, H., *Data Mining Physician Decisions for MRSA*, in *Global Forum Proceedings*. 2006, SAS Institute: Cary, NC.
74. Goossens, H., *European status of resistance in nosocomial infections.* Chemotherapy, 2005. **28**: p. 177-181.
75. Salvana, J., et al., *Chronic osteomyelitis: results obtained by an integrated team approach to management.* Connecticut Medicine, 2005. **69**(4): p. 195-202.
76. Aneziokoro, C., et al., *The effectiveness and safety of oral linezolid for the primary and secondary treatment of osteomyelitis.* Journal of Chemotherapy, 2005. **17**(6): p. 643-50.
77. Rao, N., et al., *Successful treatment of chronic bone and joint infections with oral linezolid.* Clinical Orthopaedics and Related Research, 2004. **427**: p. 67-71.
78. Senneville, E., et al., *Effectiveness and tolerability of prolonged linezolid treatment for chronic osteomyelitis: a retrospective study.* Clinical Therapeutics, 2006. **28**(8): p. 1155-1163.
79. Birmingham, M.C., et al., *Linezolid for the treatment of multidrug-resistant, gram-positive infections: experience from a comparrionsate-use program.* Clinical Infectious Diseases, 2003. **36**(159-68).
80. Cerrito, J.C., *Drug Formularies*, P.B. Cerrito, Editor. 2006: Louisville, KY. p. Discussion of drug formularies and insurance approval.
81. Shield, C.B.C.B., *Rx Preferred Drug List*. 2006, CareFirst Blue Cross Blue Shield.
82. Anonymous, *British National Formulary*. 2006, British Health Service: London.
83. Harwood, P., et al., *Early experience with linezolid for infections in orthopaedics.* Injury 2006. **37**: p. 818-826.
84. Nathwani, D., *Economic impact and formulary positioning of linezolid: a new anti-gram-positive antimicrobial.* Journal of Hospital Infection, 2001. **49**(Supplement A): p. S33-41.
85. Vinken, A., et al., *Economic evaluation of linezolid, flucloxacillin and vancomycin in the empirical treatment of cellulitis in UK hospitals: a decision analytical model.* Journal of Hospital Infection, 2001. **49**(Supplement A): p. S13-24.
86. Eckman, M.H., et al., *Foot infections in diabetic patients: decision and cost-effectiveness analyses.* Jama, 1995. **273**(9): p. 712-720.
87. Dang, C., et al., *Methicillin-resistant Staphylococcus aureus in the diabetic foot clinic: a worsening problem. Diabetic Medicine.* 20, 2003(159-161).

# Weight Watchers:
# Four Years of Successful Data Mining

Tom Osborn
Thought Experiments
PO Box 164
Broadway, NSW. Australia
osborn@it.uts.edu.au

Starting with a member view and behaviour/demographic framework in 2003, we developed a consistent and disciplined "learn as you go" framework to Weight Watchers Winback campaigning (four times a year), their customer insights and informative testing.

During our initial year, propensity modelling with customised neural networks (crafted using Don Tveter's Backprop Pro) more than doubled campaign response volume compared to a control period. The control period (12 months) already targeted prospects with a basic RFM matrix.

Our second year saw a further rise to triple the control period using boosted decision trees (TreeNet from Salford Systems), and steady but lesser growth ever since. Part of this improvement was due to more accommodative handling by our models of missing values.

The current status is a cost per Winback response which is half that of the second best performing Weight Watcher's country, and two awards from the Australian Direct Marketing Association.

The predictor variables evolved over time as tests on explanatory power eventually settled on 18 variables. These variables also corresponded to meaningful customer behaviour metrics (and demographics and proximity) which helped tune marketing communication.

Our testing framework tested offers, "bring a friend" incentives, communication, and even envelop layout. Additional testing refined "touch strategies" (rules for resting prospects), holdouts (for discounting spontaneous rejoins) and sampling to discover niche prospects which were formerly overlooked. Regular address hygiene and other data quality initiatives rounded out the framework.

# Extreme Data Mining:
# Optimized Search Portfolio Management

Kenneth L. Reed
Xtreme Data Mining, LLC.
kreed@lowermybills.com

Online paid search marketing is relatively complex requiring maintenance of thousands of keywords. For paid search, cost is strongly related to position on the search page, usually exponentially distributed. Clicks are also strongly related to position, also exponentially distributed. Revenue depends on the monetization of the clicks and is usually somewhat fuzzy. Modern portfolio theory (MPT) is a useful approach to managing a stock portfolio in that it balances potential gain with risk (expressed as the market variance for each stock). An optimization of gain versus risk shows an "efficient frontier" which identifies the optimal portfolio mix relative to an acceptable risk. We adapted this approach to search keyword management. The adaptation requires at least three predictive models: (1) a cost mode predicting cost of each keyword given position, (2) a click model predicting the number of clicks from each keyword giving position, and (3) a revenue model predicting revenue from each individual keyword. This translates into hundreds or thousands of models feeding an optimization program. We developed a working prototype using KXEN models integrated with Crystal Ball, which executed thousands of Monte Carlo simulations to find the optimal solutions: where to set each keyword and what should the cost be to maximize profit. The model worked very well in a live test.

# Data Quality Models for High Volume Transaction Streams: A Case Study

Joseph Bugajski
Visa International
Foster City, CA USA
JBugajsk@visa.com

Robert L. Grossman
Open Data Group
River Forest IL USA
& National Center for Data Mining
University of Illinois at Chicago
Chicago IL USA
rlg1@opendatagroup.com

Chris Curry, David Locke & Steve Vejcik
Open Data Group
River Forest IL USA
{ccurry, dlocke, vejcik}@opendatagroup.com

## ABSTRACT
An important problem in data mining is detecting significant and actionable changes in large, complex data sets. Although there are a variety of change detection algorithms that have been developed, in practice it can be a problem to scale these algorithms to large data sets due to the heterogeneity of the data. In this paper, we describe a case study involving payment card data in which we built and monitored a separate change detection model for each cell in a multi-dimensional data cube. We describe a system that has been in operation for the past two years that builds and monitors over 15,000 separate baseline models and the process that is used for generating and investigating alerts using these baselines

## Categories and Subject Descriptors
G.3 [Probability and Statistics]: Statistical computing, statistical software

I.5.1 [Models]: Statistical Models

## General Terms
Algorithms

## Keywords
baselines, data quality, change detection, cubes of models

## 31. INTRODUCTION
It is an open and fundamental research problem to detect interesting and actionable changes in large, complex data sets. In this paper, we describe our experiences and the lessons learned over the past three years developing and operating a change detection system designed to identify data quality and interoperability problems for Visa International Service Association ("Visa"). The change detection system produces alerts that are further investigated by analysts.



The problem is difficult because of the following challenges:

1. Visa's data is high volume, heterogeneous and time varying. There are 6,800 payment transactions per second that must be monitored from millions of merchants located around the world that are processed over a payment network that connect over 20,000 member banks. There are significantly different patterns across regions, across merchants, during holidays and weekends, and for different types of cardholders. See Figure 1 for example.

2. Alerts arising from the change detection system generally require human examination. Because of this it is necessary to balance generating a meaningful number of alerts versus generating a manageable number of alerts. If too many alerts are generated, it is not practical to manage them. If too few alerts are generated, they are generally not meaningful.

In this paper, we describe our experiences using a methodology for addressing these challenges. The methodology we use is to build a very large number of very fine grained baselines, one for each cell in a multi-dimensional data cube. We call this approach Change Detection using Cubes of Models or CDCM.

For example, we built separate baseline models for each different type of merchant, for each different member bank, for each different field value, etc. In total, over 15,000 different baseline statistical models are monitored each month and used to generate alerts that are then investigated by analysts.

We believe that this paper makes the following contributions:

1. First, we have highlighted an important category of data mining problems that has not received adequate coverage within the data mining community and whose importance will continue to grow over time.

2. Second, we have introduced a new change detection algorithm called CDCM that is designed to scale to

large, complex data sets by building a separate change detection model for each cell in a data cube.

3. Third, we have introduced a software architecture that can reliably scale with tens of thousands of different individual statistical or data mining models.

4. Fourth, we have described how we dealt with some of the practical challenges that arise when deploying data mining and statistical models in operation.

Section 2 contains some background on payment card transactions. Section 3 describes the Change Detection using Cubes of Models (CDCM) algorithm that we introduce. Section 4 describes some typical alerts detected by the system we developed and deployed. Section 5 describes the program structure that we developed around the baseline models and monitor. Section 6 describes some of the issues arising when validating the models we developed. Section 7 describes the architecture of the system we developed. Section 8 describes the implementation. Section 9 describes some of the lessons learned. Section 10 describes related work. Section 11 is the summary and conclusion.

# 32. BACKGROUND ON PAYMENT CARD TRANSACTIONS

## 32.1 Processing a Transactions

We begin by providing some background on payment card transactions that will make this paper more self-contained. In this section, we define cardholder, merchant, acquiring bank, and issuing bank and describe the major steps involved when using a payment card.

1. A transaction begins when cardholder purchases an item at a merchant using a payment card. The payment card contains an account number that identifies the cardholder.

2. The merchant has a relationship with a bank called the acquiring bank, which agrees to process the payment card transactions for the merchant. The acquiring bank provides the merchant with a terminal or other system to accept the transaction and to process it.

3. The acquiring bank has a relation with a financial payment system, such as those operated by Visa and MasterCard. The transaction is processed by the acquiring bank and passed to the payment system. Visa operates a payment system called VisaNet.

4. The payment system processes the transaction and passes the transaction to the bank (the issuing bank) that issued the payment card to the cardholder. In other words, one of the essential roles of the payment system is to act as a hub or intermediary between the acquirer and the issuer.

5. The issuing bank processes the transaction and determines if there are sufficient funds for the purchase, if the card is valid, etc. If so, the transaction is authorized; the transaction can also be declined, or a message returned asking for additional information. The issuing bank also has a relationship with the cardholder or account holder. For example, with a credit card, the cardholder is periodically billed and with a debit card the appropriate account is debited. For the year ending March 31, 2007, there were over 1.59 billion Visa cards in circulation.

6. For each of these cases, the path is then reversed and the transaction is passed from the issuing bank to the payment system, from the payment system to the acquiring bank, and from the acquiring bank to the merchant.

Our problem was to use baselines and change detection algorithms to help detect data and interoperability problems at Visa [6]. Payment data arrives at Visa from millions of merchant locations worldwide. For the year ending March 31, 2007, total annual global card sales volume was over USD $4.8 trillion[12]. Payment data is processed through risk management rules set by over 20,000 individual member banks (issuing and acquiring banks). These rules determine if a payment authorization request from a merchant either is approved or rejected by the paying bank.

For this problem, we built separate baselines for a variety of data fields, for each member bank, and for thousands of merchants. Overall, over 15,000 separate baselines are currently used each month to monitor payment card transactions at Visa.

## 32.2 Baselines for Field Values

Note: The examples in this section are hypothetical and only used for the purposes of illustrating how to define baselines.

A payment card transaction typically includes a number of fields, such as information about the point of services (POS) environment, the merchant's type of business and location, the cardholder's identity, the transaction currency, the transaction amount, and bank routing information.

We begin with an informal description of baselines based upon a simplified example. In this simplified example, assume that one of the fields of interest describes characteristics of the point of service (POS). Specifically, we assume that this field can take the following (hypothetical) values: 00, 01, 02, 03, and 04.

For an observation period of a week, assume that the frequency of these values for a certain acquirer is given by the left handt table in Table 1. Later, during the monitoring, assume that distribution is instead given by the right hand table in Table 1. The observed distribution in Table 1 is similar, except the value 04 is six times more likely in the observed distribution compared to the baseline distribution, although in both cases the

[12] Data reflects all Visa programs except Interlink, PLUS, and commercial funds transfers in China as reported by member financial institutions globally and therefore may be subject to change.

values 02, 03 and 04 still as a whole contribute less than 3% of the distribution.

The challenge for detecting significant changes is that the distributions depend upon many factors, including the region, the season, the specific merchant, and the specific issuer and acquirer.

| Value | Percent |
|---|---|
| 00 | 76.94 |
| 01 | 21.60 |
| 02 | 0.99 |
| 03 | 0.27 |
| 04 | 0.20 |
| Total | 100.00 |

| Value | Percent |
|---|---|
| 00 | 76.94 |
| 01 | 21.67 |
| 02 | 0.90 |
| 03 | 0.25 |
| 04 | 1.24 |
| Total | 100.00 |

Table 1. The distribution on the left is the baseline distribution. The distribution on the right is the observed distribution. In this example, the value 04 is over 6x more likely in the observed distribution, although the two dominant values 00 and 01 still account for over 97% of the distribution.

# 33. CHANGE DETECTION USING CUBES OF MODELS (CDCM)

In this section, we describe a methodology called Change Detection using Cubes of Models or CDCM that is designed to detect changes in large, complex data sets.

## 33.1 Change Detection Models

Change detection models are a standard approach for detecting deviations from baselines [3].

We first describe the cumulative sum or CUSUM change detection algorithm [3]. We assume that we know the mean and variance of a distribution representing normal behavior, as well as the mean and variance of another distribution representing behavior that is not normal.

More explicitly, assume that we have two Gaussian distributions with mean $m_i$ and variance $s^2_i$ for i equals 0 and 1

$$f_i(x) = \frac{1}{\sqrt{2\pi\mu_i}} \exp \frac{-(x-\mu_i)^2}{2\sigma_i}$$

The log odds ration is then given by

$$g(x) = \log \frac{f_1(x)}{f_0(x)}$$

and we can define a CUSUM score $Z_n$ as follows [3]:

$$Z_0 = 0$$

$$Z_n = \max\{0, Z_{n-1} + g(x_n)\}$$

An alert is issued where the score $Z_n$ exceeds a threshold.

Quite often the statistical distribution of the anomalous distribution is not known. In this case, if the change is reflected in the mean of the observations and the standard deviation is the same pre- and post-change, the generalized likelihood ratio or GLR algorithm can be used [3]:

$$G_k = \frac{1}{2\sigma^2} \max_{1 \le j \le k} \frac{1}{k-j+1} \left[ \sum_{i=j}^{k} (x_i - \mu_0) \right]^2, \qquad k > 1$$

where $m_0$ is the mean of the normal distribution and s is the standard deviation of the both the normal and abnormal distributions, which are assumed to be Gaussian. Here k is fixed and determines the size of the window used to compute the score. Again, the detection procedure is to announced a change at the first up-crossing of a threshold by the GLR score.

## 33.2 Cubes of Models

The basic idea of the CDCM algorithm is that for each cell in a multi-dimensional data cube, we estimate a separate change detection model.

For the purposes here, we can define a *data cube* as usual, namely a multi-dimensional representation of data in which the cells contain measures (or facts) and the edges represent data *dimensions* which are used for reporting the data.

We define a *cube of models* as a data cube in which each cell is associated with a baseline model. See Figure 2.

## 33.3 Learning and Scoring Change Detection Models

In our model, we assume that there are a stream of events, which in our case are transactions, and each event can be assigned to one or more cells in cube of models. For example, for each project, each transaction is assigned to the appropriate cell(s), as determined by one of six regions, by one of over 800 Merchant Category Codes (MCCs), by one of 8 terminal types, etc. We also assume that various derived data attributes are computed from the event data to form feature vectors, which are sometimes called profiles in this context. The profiles contain state information and derived data and are used as inputs to the models as usual.

**Estimating baseline models.** To learn the baseline models, we take a collection of event data and process it as follows.

1. First, we assign each event to one or more cells in the data cube as appropriate.
2. Second, we transform and aggregate the events and compute the required profiles for each cell using the event data.
3. Third, for each cell, we use the resulting profiles to estimate the parameters for the baseline model, and output the baseline model.

**Scoring baseline models.** To score a stream of event data, we proceed as follows.

1. First, we retrieve the appropriate XML file describing the segmentation.
2. Next, we assign each event to one or more cells in the data cube as appropriate.
3. We then access the profile associated with each cell, and update the profiles using the new event.
4. We then use the profile as the input to the appropriate baseline model and compute a score.
5. Next, we process the resulting score using the appropriate rules to determine whether an alert should be produced.
6. Next, we apply XSLT transformations to the score to produce a report describing the alert.
7. Finally, if an alert is produced, we pass the alert to the required application or analyst.

# 34. SOME TYPICAL ALERTS

## 34.1 Summary

Since the data interoperability program began approximately three years ago, Visa and its trading partners have fixed 70 data interoperability issues. An improvement in annual card sales volume realized thereby is about $2 billion, compared with global annual card sales volume of $4.8 trillion. These "fixed Alerts" comprise 25% of data interoperability issues presently being investigated.

In this section, we describe four typical alerts that have been generated by the CDCM system. Currently, we compute alerts each month and re-estimate baselines several times a year.

It is important to remember when reading the case studies in this section that the issues identified by these alerts represent both a very small fraction of the transactions and a very small fraction of the total purchase dollars.

## 34.2 Dimensions of Cube

For the Alerts that we describe below, we used the following dimensions to define a data cube:

1. The geographical region, specifically the US, Canada, Europe, Latin America, Asia Pacific, Middle East/Africa, and other.
2. The field value or combination of values being monitored.
3. The time period, for example monthly, weekly, daily, hourly, etc.
4. The type of baseline report, for example a report focused on declines or a report describing the mixture of business for a merchant.

Today (June, 2007), for each of 324 field values times 7 regions times 1 time period times 3 report types, we estimate a separate baseline, which gives 324 x 7 x 1 x 3 = 6816. In addition, for 623 field values times 7 regions times 1 time period times 2 report types, we estimate a separate baseline, which gives an additional 623 x 7 x1 x 2 = 8726 separate baseline models. So in total, we are currently estimating 15,542 (=6816+876).

Actually, the description above is a simplified version of what actually takes place. For example, the 6816 baselines mentioned arise from 324 x 7 = 2272 different field values, but the 2272 different field values are not spread uniformly across the 7 regions as indicated, although the total is correct.

## 34.3 Incorrect Merchant Category Code

In this example, an airline was coding some of its transactions using a Merchant Category Code (MCC) B instead of the preferred MCC A, which coincided with a lower approval rate for the airline's payment authorization requests. Lower authorization approval rates have been shown in work on other alerts to be associated with a loss of purchase value for Visa and its Member banks. These factors led to the production of a baseline alert that was followed by an analyst's investigation. Once the analyst confirmed the issue, a conference call was arranged with a person responsible for the relationship with the acquiring bank for the airline. As a result of this call, the airline installed a fix that lead to improved authorization approval performance and increased annual purchase volume.

## 34.4 Testing of Counterfeit Cards

In this example, the decline rate for a large bank was essentially the same month to month but the baseline model identified a particular category of transactions (specified by a combination of five fields) for which the decline rate sharply peaked in September 2006 compared to an earlier baseline period. One way of thinking about this, is that for this bank, most of the 50,000+ or so baselines were normal for September, but one was not. When investigated, this particular baseline was elevated due to unauthorized testing of Visa accounts, a practice that sometimes is associated with underlying criminal activity; e.g., validation of active card accounts using illegally obtained card account data. Visa and the bank moved swiftly correct the problem following further investigation.

## 34.5 Incorrect Use of Merchant City Name

In this example, a European merchant's transactions were coded incorrectly; i.e., incorrect information was contained in the data field that is used to encode the name of the city where the transaction occurred. This also was associated with a lower approval rate for payments from this merchant. The lower approval rate was detected by a baseline alert that monitored decline levels for each MCC for each acquirer. After

investigation and communications with the acquirer, the merchant corrected the problem.

## 34.6 Incorrect Coding of Recurring Payments

A recurring payment is an arrangement agreed between a cardholder and merchant whereby a merchant periodically submits payment authorization requests on behalf of their customer for continual use of a product or service. Examples include monthly payments for mobile phones, internet, or satellite television services. Recurring payments are coded specially to indicate that the cardholder previously requested this Visa service of the merchant. In this case, a Middle Eastern merchant incorrectly coded payments as recurring and the decline rate was higher than it should have been had the transaction been coded otherwise. In addition, after examination by an analyst, it became clear that the merchant name was also inconsistently coded, compounding the problem. This alert was detected using a MCC-baseline.

# 35. PROGRAM STRUCTURE

In this section, we describe the structure of the program created to monitor and improve data quality. See Figure 3.

**Strategic Objective.** Critical to the success of the program was identifying the strategic objective of the program. After much discussion, the following objective was agreed to: identify and ameliorate data quality and data interoperability issues in order to maintain and improve 1) the approval of valid transactions; 2) the disapproval of invalid or fraudulent transactions; and 3) the correct coding of transactional data. The first two objectives increase the satisfaction of card holders and member banks, while the third objective lowers the overall cost and increases the efficiency of transaction processing. The success of the program was then tracked by a dashboard that monitored the additional dollars processed and the savings resulting from direct actions of the program.

**Governance.** The Visa Data Interoperability Program was established in 2004 by the global council of the CIO's of Visa's operating units. The program is governed by a council of business executives and technical experts who set the rules, procedures and processes of the program.

**Monitoring.** A system supporting data cubes of baseline models was designed and developed using the ideas described above to monitor transactional payment data. For the purposes here, we call the system the Monitor. The Monitor receives daily samples of tens of millions of authorization messages and clearing transactions from a central ETL facility inside VisaNet. Statistically significant deviations from baselines that are associated with high business value generate what are called Baseline Threshold Alerts.

**Screening of Alerts.** Baseline Threshold Alerts are screened by an analyst to produce what are called Baseline Candidate Alerts. Candidate alerts are then analyzed by program analysts and other subject matter experts to understand the issues that led to the candidate alert and to more carefully estimate the business value involved. If the program team believes that an issue is valid and sufficiently valuable that the cost of repair may be recovered through recaptured revenue or lower processing costs, and, furthermore, they believe that the issue is sufficiently clear that it may be explained accurately, they send a Program Alert to the customer relationship manager at the Visa operating region that is closest to the source of the problem. This may be a third party processor, an acquiring bank, or a VisaNet technical group.

**Investigations.** The customer relationship manager works with the program analysts to explain the problem identified by the Program Alert to the bank or merchant and to work with them to estimate the cost required to fix the problem. The program team meanwhile reviews measurements to determine when and if the problem is resolved. If the data measurements indicate a resolution, then the business that effected the change is contacted once again to validate recovery of revenue or loss avoidance.

**Reference Model.** The governance council adopted technical standards for how alerts are generated and business process standards for how alerts are investigated. These rules are recorded in a Reference Model that is maintained by the Program and updated at least twice each year.

**Standards.** Over time, we developed an XML representation for baseline models and for segmentation. We worked with the PMML Working Group and this work has not contributed to the PMML Baseline and Change Detection Model, which is currently in a RFC status. Over the long term, this should reduce the total costs of the system by enabling the use of third party tools that support this RFC draft standard.

# 36. VALIDATION

Fraud models are relatively easy to validate using detection rates and false positive rates. Response models are relatively easy to validate using lift curves.

On the other hand, the change detection models we used are somewhat harder to validate. Broadly speaking, change detection models are similar to association rules in that for a given threshold and support, an association algorithm will find all corresponding association rules. Of the ones found, some may have business significance, while others may not. An analyst is needed to tell the difference.

The change detection models we used for this project are similar in the sense that for a given segmentation and threshold, all corresponding changes are detected by the algorithm, but not all have business significance. Again, an analyst is needed to separate those with business significance from those without. The role of the baseline models is simply to produce a useful flow of alerts for investigation. The segmentation process is critical so that the alerts are both meaningful (the segment is no so broad that the alerts are without business significance) and manageable (the segment is not too narrow, producing too many alerts).

As the program matured, we have introduced several rules that have resulted in more useful alerts. First, as described above, we introduced a quick manual screening process before Baseline Candidate Alerts are generated. Second, we have introduced various thresholds, which depend upon the issue, region, and other factors, involving number of transactions, the type of

transactions, and the dollar amount, etc. that are used in part to determine Baseline Candidate Alerts.

# 37. SYSTEM ARCHITECTURE

In this section, we describe the system architecture that we developed for this project. See Figure 4. We believe that this architecture is quite general and would prove useful for a variety of projects.

## 37.1 Predictive Model Markup Language

Metadata for the system was stored using the Predictive Model Markup Language or PMML [5]. PMML was used to specify the data attributes (using the PMML data dictionary), the attributes used as inputs for each model (using PMML mining attributes), the specification for derived functions (using the PMML transformation dictionary and local transformations), and the specifications of model segments (using a PMML extension we helped define).

Additional metadata about the system was stored in an emerging part of PMML called the PMML deployment package. Using the PMML deployment package, we specified such things as the source and location of various data feeds, as well as how the data should be pre-processed and post-processed.

## 37.2 Baseline Producers and Consumers

The main two system components are a baseline producer and a baseline consumer.

The baseline producer extracts transactions data from a project data mart, computes derived attributes and state information from the transactions (which are sometimes called profiles), and then uses the information to estimate the parameters of a separate baseline model for each segment. The baseline models for each segment are saved as an PMML.

The baseline consumer first reads a PMML file specifying the segmentation and the model parameters for each segment. The baseline consumer than processes a stream of transaction data and produces a stream of scores. Sometimes a PMML consumer is called a scoring engine since it processes input data using a PMML file to produce a stream of output scores.

## 37.3 XSLT-based Report Processing

An important part of the architecture turned out to be a basic module that took XML files of scores produced by the baseline Consumer and transformed them using definitions for reports specified using XSLT. In this way, we could relatively easily change the report formats. Being able to experiment quickly and easily with different report formats turning out to be an important for the project, as we describe in more detail below in the section on lessons learned.

## 37.4 Metadata Repository

We stored the various metadata for the project, including the PMML files, in a repository. This was important since it was a project requirement that we be able to retrieve the specification for any baseline that had been computed by the project. Using the repository, which was backed up automatically, also satisfied a business continuity requirement of the project.

## 37.5 Defining Cubes of Models Using Segmentation

As we mentioned, the success of the approach was critically dependent on being able to define and manage easily ten of thousands of different segments. To do this we developed a PMML extension to define how data should be divided in order to define different segments, each of which was used to estimate a separate model.

After several refinements over the past two years, we defined several different ways to define segments. For this project, we mainly used the following mechanisms for defining segments:

- Regular Partitions. With a regular partition, a field name, the left end point, the right end point, and the number of partitions is specified. Regular partitions in two or more dimensions can be defined by specifying the required data for each field independently.
- Explicit Partitions. With an explicit partition, the field name, the left end point, and the right end point are given for each interval in the partition. Note that with explicit partitions, the intervals may be overlapping. Again, multi-dimensional partitions are defined by defining each dimension independently.
- Implicit Partitions. With an implicit partition, a field name is provided and then each unique value of the field is used to define a distinct partition. For example, assume that city is a field in a data set that is identified as an implicit partition field. In this case, a separate model would be created for each city.

# 38. IMPLEMENTATION

## 38.1 Augustus

In part motivated by this project, we developed an open source PMML-compliant data mining system called Augustus. Augustus, which is available through Source Forge (www.sourceforge.net/projects/augustus), includes a baseline producer and a baseline consumer.

Augustus is written in Python. The current version of Augustus is 0.2.6. The Augustus kernel contains a data management component called UniTable for Universal Table. A UniTable is broadly similar to a R data frame. Data is arranged in columns, the columns may be of different types, but all columns must have the same number of rows.

Derived columns can be easily added to a UniTable and many columns operations can be vectorized. UniTable is based on the Python numarray library.

Augustus includes a library for working with PMML and also a library for importing data. Augustus currently is distributed with two models: a baseline model and a tree mode

## 38.2 Scalability of Augustus

Transactions on VisaNet can peak at over 6,800 transactions per second. For this reason, an important requirement of this project is that our Augustus-based scoring must also be able to

process events at this speed. Table 1 are some sample benchmarks for scoring showing that scoring is independent of the number of events scored, which is an important requirement for our project.

Another important requirement for our implementation is that scoring of events and building of baselines be independent of the number of segments. The same table shows that Augustus satisfies this requirement also.

| Number of Events | Number of Explicit Segments | Events/sec |
|---|---|---|
| 1,000,000 | 10 | 14,880 |
| 1,000,000 | 100 | 15,216 |
| 1,000,000 | 1,000 | 14,808 |
| 1,000,000 | 10,000 | 13,790 |
| | | |
| 2,000,000 | 10 | 15,100 |
| 2,000,000 | 100 | 15,278 |
| 2,000,000 | 1,000 | 14,401 |
| 2,000,000 | 10,000 | 14,320 |
| | | |
| 4,000,000 | 10 | 15,198 |
| 4,000,000 | 100 | 15,251 |
| 4,000,000 | 1,000 | 15,137 |
| 4,000,000 | 10,000 | 14,845 |
| | | |
| 8,000,000 | 10 | 15,299 |
| 8,000,000 | 100 | 15,407 |
| 8,000,000 | 1,000 | 15,367 |
| 8,000,000 | 10,000 | 14,745 |

Table 2. The events per second processed by the Augustus scoring engine is approximately independent of the number of events scores and the number of segments.

## 38.3 PMML Baseline Models

Over the past few years, we have worked with the PMML Working Group to develop a PMML representation for baseline models, which was developed in part to support this project. This proposal has been modified by the PMML Working Group and the modified proposal is currently available as a PMML RFC (Request For Comment) [5]. The proposal for baseline models includes CUSUM models, GLR models, threshold break models, (a measure exceeds a threshold), and contingency table models.

We also developed, in part motivated by this project, a PMML extension for defining segmented models. Some of the different types of segments supported are described above. Having a easy to define segments turned out to be very useful for this project.

# 39. DISCUSSION AND LESSONS LEARNED

**Lesson 1.** The most important lesson we learned was that thus far it has been more fruitful to examine many individual baseline and change detection models, one for each different segment of the event stream, even if the these models are very simple, than to build a single, relatively complex model and apply it to the entire event stream.

**Lesson 2.** The time and effort required to get the alert format right is substantial. Although it was certainly expected, the business return on the project was dependent to a large degree on the ability to deliver to the analysts information in a format that they could readily use. After quite a bit of experimentation, a report format was developed that reported:

- What is the issue? This part of the report identifies the relevant business unit and the relevant business issue.
- Who has the issue? This part of the report identifies the relevant subsystem of VisaNet, the relevant attribute, and the relevant attribute value.
- What is the business opportunity? This part of the report identifies the daily business value associated with the issue and the statistical significance of the alert (Low, Medium High).
- What is the business impact? This part of the report describes the a business measure as currently measured, the historical measure of the business measure during the baseline period, and the number of transactions affected.

The final part of the report contains additional information, such as the alert ID, alert creation date, whether the alert is new, and whether the alert is associated with an issue that has been previously identified and now is being monitored for compliance.

One way to summarize the report is that the items in alerts gradually changed from those items related to the statistical models and how the alerts were generated to items directly related to how the alerts were investigated and how the business impact was estimated. The surprise for us was not that this transition had to be made, but rather the time and effort required to get it right.

**Lesson 3.** It turned out that some of the most important alerts we found were alerts that had low statistical significance. For each report, we include an estimate of the statistical significance of the alert (low, medium, high and very high) as well as an estimate of the business significance of the alert (in dollars). It turned out that after investigation, the alerts that generated by most dollars saved, were often the alerts with low statistical significance. For this reason, it was usually not a good idea to investigate alerts in the order of most statistically significant to least statistically significant. Rather, the analysts used a more

complex prioritization that thus far we have not tried to formalize.

**Lesson 4.** As a result of analysis of an alert, it was sometimes possible to create specialized baselines and reports that would look for similar problems in the future. We quickly learned that even a few specialized reports like this could easily occupy most of our available time. The lesson we learned was that it was important to devote some of our time to looking for new opportunities (think of this as a survey activity), since some of these turned out to be even more important than what we were currently doing.

## 40. RELATED WORK

There is a large amount of research on change detection algorithms per se. The monograph by Basseville and Nikiforov [3] is a good summary of this research field. In particular, the change detection algorithms that we use here, including CUSUMs, Generalized Likelihood Ratios, and related algorithms are covered in this reference.

The work described in this paper differs from classical change detection and contingency tables in that it uses a separate change detection model for each cell in a cube of models.

More recently, Ben-David, Gehrke and Kifer [4] introduced a non-parametric change detection algorithm that is designed for streams. The methods used here are parametric. In contrast to their approach which uses a single change detection model, we build a large number of models in order to handle complex, heterogeneous data, one for each cell in a multi-dimensional data cube.

The paper by Fawcett and Provost [7] has a similar goal – detecting unusual activity in large data sets – but uses a much different approach. Their approach is to introduce operating characteristic style measures in order to identify unusual behavior.

Guralnik and Srivastava [9] study event detection in time series data by using a new change detection algorithm they introduce, which involves iteratively deciding whether to split a time series interval to look for further changes.

In contrast to all these methods, our approach is to use relatively simple classical change detection algorithms, such as CUSUM and GLR, but to build thousands of them, one for each cell in a multi-dimensional data cube. As far as we are aware of, our paper is also one of the few papers in the data mining literature that presents a case study of change detection involving a system as large and heterogeneous as VisaNet.

## 41. SUMMARY AND CONCLUSION

In this paper, we have shared our experiences and some of the lessons learned over the past two years developing and operating a baseline and change detection system for Visa. Because of the complex and highly heterogeneous nature of Visa's transactional data, we did not build a single change detection model, but rather over 15,000 individual change detection models. Indeed we built a separate change detection model for each cell in a multi-dimensional data cube. This is an example of we have been calling Change Detection using Cubes of Models or CDCM.

Overall, the approach seems to work quite well. Indeed, substantial business value is being generated using this methodology, and thus far we have not been able to achieve the same performance using a single baseline or change detection model.

To summarize, we have demonstrated through this case study that change detection using data cubes of models (CDCM) is an effective framework for computing changes on large, complex data sets.

## 42. REFERENCES


[18] Alan Agresti, An Introduction to Categorical Data Analysis, John Wiley and Sons, Inc., New York, 1996.

[19] The Augustus open source data mining system can be downloaded from www.sourceforge.net/projects/augustus.

[20] M. Basseville and I. V. Nikiforov. Detection of Abrupt Changes: Theory and Application. Prentice Hall, 1993.

[21] Shai Ben-David, Johannes Gehrke, Daniel Kifer, Detecting Change in Data Streams, Proceedings of 2004 VLDB Conference, 2004.

[22] The Predictive Model Markup Language, Data Mining Group Version 3.1, retrieved from www.dmg.org on January 10, 2007.

[23] Joseph Bugajski, Robert Grossman, Eric Sumner, Tao Zhang, A Methodology for Establishing Information Quality Baselines for Complex, Distributed Systems, 10th International Conference on Information Quality (ICIQ), 2005.

[24] Tom Fawcett and Foster Provost, Activity monitoring: noticing interesting changes in behavior, KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 53--62, ACM Press, New York, 1999.

[25] Robert L. Grossman, PMML Models for Detecting Changes, Proceedings of the KDD-05 Workshop on Data Mining Standards, Services and Platforms (DM-SSP 05), ACM Press, New York, 2005, pages 6-15.

[26] Valery Guralnik and Jaideep Srivastava, Event detection from time series data, Proceedings of the Fifth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining, ACM Press, New York, NY, 33-42, 1999

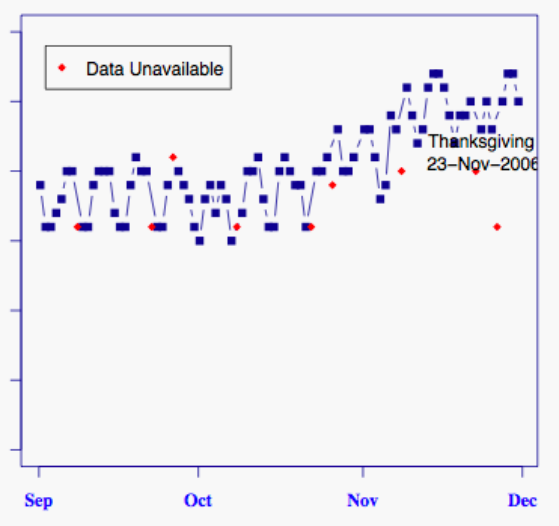



Figure 1. This is an example of one of the measures monitored in this project. This graph shows how the ratio of declined transactions varies for one of the Merchant Category Codes (MCC) monitored. Note the daily, weekly and monthly variation in the data. This variation, which is typical of the measures tracked, is one of the reasons detecting changes in this data is challenging. The vertical axis scale has been omitted due to confidentiality reasons.

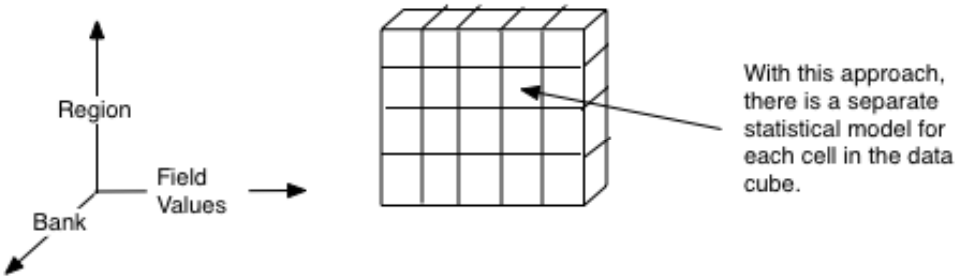



Figure 2. The basic idea with change detection using cubes of models orCDCM is that there is a separate change detection model for each cell in a multi-dimensional data cube. In the work described here we estimated and maintained over 15,000 different baseline statistical models and monitored them monthly.

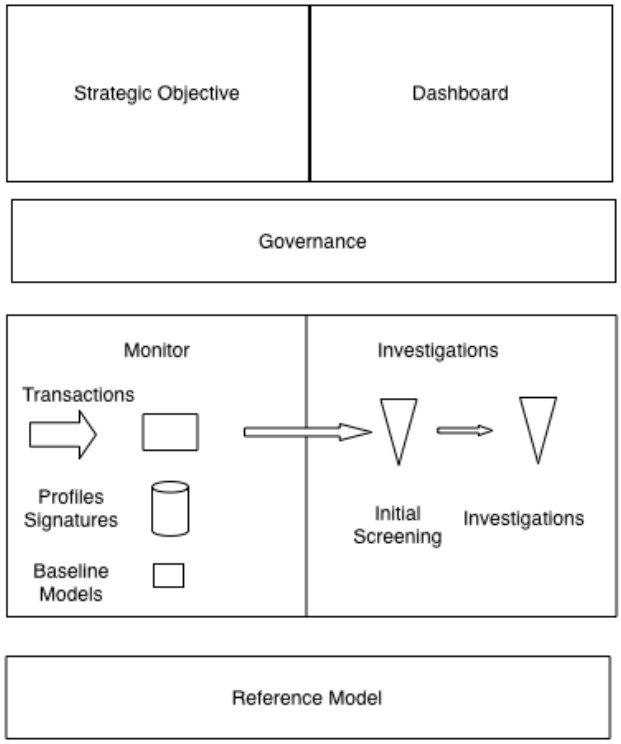



Figure 3. This figure summarizes some of the key components of the program and process set up to detect data quality and data interoperability problems. Note that the baseline model and monitor are just two of the components.

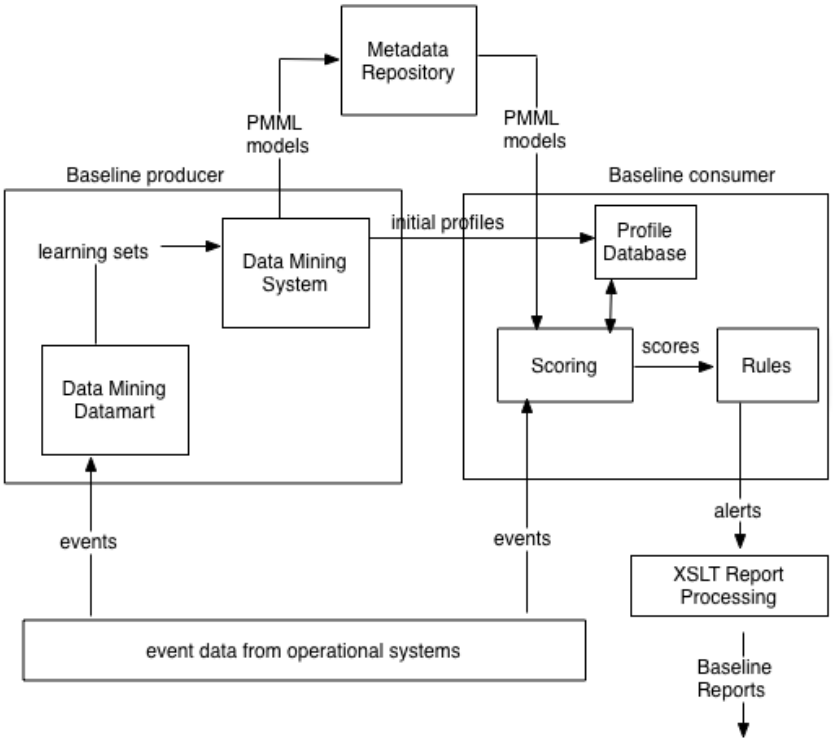



Figure 4. This figure shows the architecture we used to compute the baseline alerts