Identification, Validation and Measurement of Derived Factors in the Detection and Ranking of VoIP and Mobile Communication Fraud

Ivan Zasarsky
Deloitte & Touche LLP

Abstract

Mobile and VoIP communications are increasingly vulnerable to identity fraud as a breeder crime for various frauds. This is especially the case in the use of PSTN, IP and hybrid networks for voice over Internet protocol communications (VoIP).

Past research is based mainly on the interception and comparison of originating and terminating call numbers. Current systems are proprietary, inflexible and do not identify factors in a standard or replicatable manner, suitable for analysis or scientific method. Moreover the analytical techniques are based on fixed line communication network topologies which are unsuitable for comparison due to alternate standards, disparate technologies and use patterns.

The objective of this research is to advance the knowledge of the factors which represent events manifesting fraud within a mobile and VoIP communications environment. The researcher developed cardinal data sets to be used in the evaluation and experimentation of counter measures.

The researcher determined a highly accurate method to measure, count and analyze the frequency of events in a telecom switch environment so as to develop predictive models of these behaviors. Moreover, the researcher found that the use of multivariate analysis and the engagement of grid based ranking by way of matrix techniques improved the number of events detected and reduced the number of false positives as compared to the baseline.

Introduction

The Rationale and Significant Need for This Study

The objective of this analysis is to advance the knowledge of the factors which represent events manifesting fraud within a mobile communications environment. The researcher developed cardinal data sets to be used in the evaluation and experimentation of counter measures. The researcher sought to determine a highly accurate method to measure, count and analyze the frequency of events in a telecom switch environment so as to develop predictive models of these behaviors. The risk of not establishing these new methods, tools and processes could mean that entire IP communication networks will be exposed to process

and transit plans, acts and proceeds of crime on a national and international basis.

Cellular communications are composed of several elements, each having a technological vulnerability to the threat of exploitation by a fraudulent act. Several events types have been identified as being able to be represented by electronic data via intercepts from a telecom switch. These include “cloning,” “SIM card cracking,” “and subscription fraud.” The researcher sought to isolate a subset of core elements and derive variables to capture call information at the switch level. In obtaining a large sample of calls the researcher extracted and stored the identified attributes in a relational database of some 1 million Call Data Records and employed a random sample by inserting new records from a master set captured over a 13 month period.

Profile generation and the application of SQL procedures to extract key variables was done by methods as defined in figure 1.

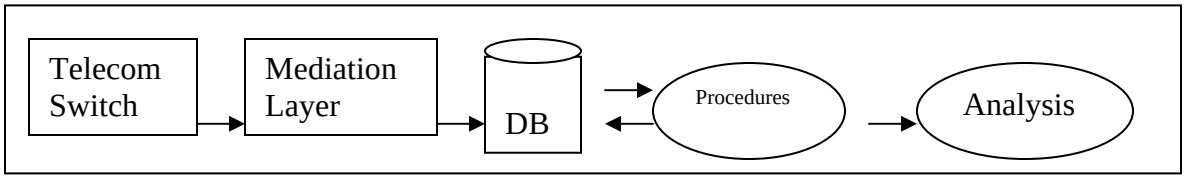

Figure 1: Data capture methodology

Telecommunication switches utilize standards to represent the data of a call to effect the origin, completion and termination of that call. The call is represented using a specified binary string of data which encompasses some dozens of attributes. The data is available for extraction in its native format; however creating meaningful information from this data requires the mapping and matching of the data to data segments by way of a mediation layer. Once in the format which is able to complete the next stage of transition to being loaded into a relational database, this data was placed into a specified data model capable of representing classes and structures useful for the detection of event representing fraud. Finally the researcher created particular stored procedures in sequential query language (SQL) in order to apply the mathematical methods, techniques and tools to accurately and effectively detect fraud. This is represented by example in Figure 2.

The researcher set thresholds of detected events after initial experimentation so as to determine acceptable numbers of events. These thresholds will be were established by way of comparative analysis with existing baselines and expert interpretations from current norms as understood by the ITU (International Telephone Union). As such the researcher was able to compare the effectiveness as a ratio of observable events when compared to pre-existing baselines. The researcher then tested the probabilities of predictive behavior of

the system by utilizing a tool designed to display the simple frequencies of the variables as identified above.

The researcher holds that the independent variables are the A- number and B-number of the caller and its connection party. Dependent variables include type of service, time, location and activity levels. A test bank of pooled data was maintained as a baseline, once validated by expected result and expert review. Thus the experimental observations were compared to the derive probabilities of the baseline.

The subjects were defined cellular telephone customers, namely users of portable telephones subscribing to an automatic public telephone service using cellular technology that provides access to the PSTN (Public Service Telephone Network).The researcher gained access via the inferred representation of these subjects as cellular telephone subscribers using captured records and interactions of A and B numbers via CDRs.

Analysis was done using SQL based representation of statistical functions bearing frequency and probability in the format of risk scoring and expressed as a complex deviation calculating a summative score called Damage. Various logarithmic charts and graphical display of said results are dynamically expressed so as to represent the attempts to evaluate the feasibility of establishing empirical relationships as observed in potentially causal events.

```
CREATE PROCEDURE ROAM_RISK
AS
        select      number,
                    property,
                    service,
                    thishour,
                    today,
                    thisweek,
                    thismonth,
                    lastmonth,
                    credit,
                    warninglimit,
                    alertlimit,
                    terminatelimit,
                    manager,
                    status,
                    damage,
                    firstcall,
                    modified,
                    'name'=(select distinct operator from networks b  (NOLOCK) where b.network = left(a.number,5)),
                    'DailyUnits'=(select distinct dailyunits from networks b (NOLOCK) where b.network = left(a.number,5)),
                    'caselink'=' '
        from
            roam_units a (NOLOCK)
                    where ((damage > 400 and status <> 'OK')  or (thismonth > warninglimit and status <> 'OK') )
                    order by left(a.number,5),damage DESC
GO
```

Figure 2: Example of a stored procedure to extract roaming events from CDRs.

Study Design

The challenge in the isolation of factors lending themselves to accurate and effective representation of fraud in mobile telephony is described in the existing literature. The methods of singular representation of data sets in static environments have lent themselves to both neural and probabilistic methods of detection of events represented by those data sets. In an especially creative and advanced work, the problem of dynamically representing a model to encompass an agnostic data set bearing no specified or prototypical scenarios has been met with a limited but promising approach. In more recent works the dynamic detection and representation of fraud in mobile topologies has met with a trade-off between highly accurate but computationally prohibitive methods or indicative and comparatively inaccurate methods.

The researcher seeks to provide a means by which the main classes and corresponding data sets represent specified aggregations of events capable of representing fraudulent behavior in mobile and VoIP usage. Moreover, doing so in a computationally efficient manner so as to maintain high degrees of accuracy and low time segment thresholds below 1 minute sampling would represent an

advance in the knowledge of fraudulent events, in the representation and application of computationally advanced frameworks.

Parametric Portfolio Revaluation

Parametric portfolio revaluation maps simulated changes in risk factors in simulated changes in risk portfolio value. Parametric portfolio revaluation differs from full revaluation in that changes in portfolio value are represented by a smaller number of parameters. These are in turn derived from a small number of exact recalculations of each profile in the portfolio. Two types of parametric portfolio representation are Taylor series expansion and Factor Sensitivity Grids. Comprehensive VaR shall use Taylor series expansions as the valuation method for standard scenarios and portfolios, and shall use factor sensitivity grids for specific products and portfolios for which Taylor Series approximations may be inadequate.

Factor Grid sensitivity model

Existing value at risk systems utilize internally developed models that account for fat tails and jumps in risk factors. They represent improvements over the standard (normal) methodology that substantially underestimates value at risk and hence cannot detect aggregates in a computationally effective manner, based on the dynamically changing data within the data input streams. Moreover, these models are not able to reproduce different distributional shapes and proper kurtosis for all risk factors and aggregation levels. Given these restrictions, difference in methodologies and lack of global correlations, a comprehensive risk measure across all risk types and markets cannot be achieved within existing systems.

Hypothesis

Given that there exists a relationship of specified variables within data sets and the derivative calculations of the probability that aggregations of these sets indicate a fraudulent event may be ranked by scoring said results as a high value event. The ability to computationally derive such a result will be significantly improved by way of applying highly effective mathematical filters so as to computationally sift through large and time sensitive data sets having their source in PSTN telephony, IP switches, specialized routers and statefull firewalls. In particular the researcher seeks the application of multivariate engines and the use of nodes to rapidly represent changes in the factors and corresponding data sets enable a high degree of correlation may be achieved using less computationally intensive methods.

Operationalization

Fraud detection is based on the calling activity of mobile phone subscribers. The problem of fraud detection is to discover dishonest intentions of the subscriber, which clearly can not be directly observed. Acknowledging that the intentions of the mobile phone subscribers are reflected in the calling behavior and thus in the observed calling data, the use of call data as a basis for modeling is well justified. Conventionally, the calling activity is recorded for the purpose of billing in call records, which store attributes of calls, such as the identity of the subscriber (IMSI, International Mobile Subscriber Identity), time of the call, and duration of the call. In all, dozens of attributes are stored for each call. In the context of GSM networks, the standard about administration of subscriber related events and call data in a digital cellular telecommunications system can be found in European Telecommunications Standards Institute. Similar standards and protocols exist for the establishment, transmission and quality of service related to the use of VoIP via RTP over UDP protocols. The ability to track and identify fraud related events have had little or no study to date. Moreover the use of type of service identifiers (TOS) within these protocols and standards are not present in several proprietary methods. The researcher therefore needs to restrict the operationalization to SIP and H.323 like transmission and simple interoperation with PSTN networks. (RFC 3225)

Definition of Fraud

Many definitions in the literature exist, where the intention of the subscriber plays a central role. Johnson (1996) defines fraud as any transmission of voice or data across a telecommunications network where the intent of the sender is to avoid or reduce legitimate call charges. In similar vein, Davis and Goyal (1993) define fraud as obtaining unbillable services and undeserved fees. According to Johnson, the serious fraudster sees himself as an entrepreneur, admittedly utilizing illegal methods, but motivated and directed by essentially the same issues of cost, marketing, pricing, and network design and operations as any legitimate network operator. Hoath (1998) considers fraud as attractive from the fraudsters' point of view, since detection risk is low, no special equipment is needed, and the product in question is easily converted to cash. Although the term fraud has a particular meaning in legislation, this established term is used broadly to mean misuse, dishonest intention or improper conduct without implying any legal consequences.

Motivation for Fraud Detection

Following the definition of fraud, it is easy to state the losses caused by fraud as the primary motivation for fraud detection. In fact, the telecommunications industry suffers losses in the order of billions of US dollars annually due to fraud in its networks (Davis and Goyal, 1993; Johnson ,1996; Parker, 1996; O'Shea,

1997; Pequeno, 1997; Hoath, 1998). In addition to financial losses, fraud may cause distress, loss of service, and loss of customer confidence (Hoath, 1998). The financial losses account for about 2 percent to 6 percent of the total revenue of network operators, thus playing a significant role in total earnings. However, as noted by Barson et al. (1996), it is difficult to provide precise estimates, since some fraud may be never detected, and the operators are reluctant to reveal figures on fraud losses. Since the operators are facing increasing competition and losses have been on the rise (Parker, 1996), fraud has gone from being a problem carriers were willing to tolerate to being one that dominates the front pages of both trade and general press (O'Shea, 1997). Johnson also affirms that network operators see call selling as a growing concern.

Development of Fraud

Historically, earlier types of fraud used technological means to acquire free access. Cloning of mobile phones by creating copies of mobile terminals with identification numbers from legitimate subscribers was used as a means of gaining free access (Davis and Goyal 1993). In the era of analog mobile terminals, identification numbers could be easily captured by eavesdropping with suitable receiver equipment in public places, where mobile phones were evidently used. One specific type of fraud, tumbling, was quite prevalent in the United States (Davis and Goyal 1993). It exploited deficiencies in the validation of subscriber identity when a mobile phone subscription was used outside of the subscriber's home area. The fraudster kept tumbling (switching between) captured identification numbers to gain access. Davis and Goyal state that the tumbling and cloning fraud have been serious threats to operators' revenues. First fraud detection systems examined whether two instances of one subscription were used at the same time (overlapping calls detection mechanism) or at locations far apart in temporal proximity (velocity trap). Both the overlapping calls and the velocity trap try to detect the existence of two mobile phones with identical identification codes, clearly evidencing cloning. As a countermeasure to these fraud types, technological improvements were introduced.

However, new forms of fraud came into existence. A few years later, O'Shea (1997) reports the so-called subscription fraud to be the trendiest and the fastest-growing type of fraud. In similar spirit, Hoath (1998) characterizes subscription fraud as being probably the most significant and prevalent worldwide telecommunications fraud type. In subscription fraud, a fraudster obtains a subscription (possibly with false identification) and starts a fraudulent activity with no intention of paying the bill. It is indeed non-technical in nature and by call selling, the entrepreneur-minded fraudster can generate significant revenues for a minimal investment in a very short period of time (Johnson 1996). Mechanisms of the first generation soon became inadequate. More advanced detection mechanisms must be based on the behavioral modeling of calling activity.

Data Collection

In order to develop models of normal and fraudulent behavior and to be able to assess the diagnostic accuracy of the models, call data exhibiting both kinds of behavior is needed. Gathering normal call data is relatively easy as this mode dominates the population, but collecting fraudulent call data is more problematic. Fraudulent call data is relatively rare and the data collection involving human labor is expensive. In addition, the processing and storing of data is subject to restrictions due to legislation on privacy of data.

Procedures in data collecting differ both in the way they are conducted and in the way the data is grouped in the normal and fraudulent modes. The manner of collecting fraud data for development of a fraud detection system is described. Data labeled as fraudulent is a sample from a mixture of normal and fraudulent data, the mixing coefficients being unknown and changing in time. Therefore, call data is labeled to classes fraud and normal on a subscriber basis. No geographical information about the calls was available. However, the use of derived factors may be present in specific protocols.

Research Design/ Data and Methods

Methodology

A multivariate MobyF-VaR model was used, which is a generalization of the SV-VaR Monte Carlo simulation methodology implemented in VaR Systems. The MobyF-VaR model allows for:

- fully integrated VaR measure across all risk types and mobile markets
- global correlation structure of the risk factor returns; appropriate shapes, kurtosis, and heavy tails for different risk factors and aggregation levels;
- multivariate stochastic volatility correlated across the mobile markets and risk factors;
- proper scaling.

These properties of the MobyF-VaR Model make risk estimates within the value at risk system more accurate, consistent across different fraud scenarios and aggregation levels, more reliable, and robust. The use of principal component decomposition in the MobyF-VaR model improves the Monte Carlo Engine performance more than ten times.

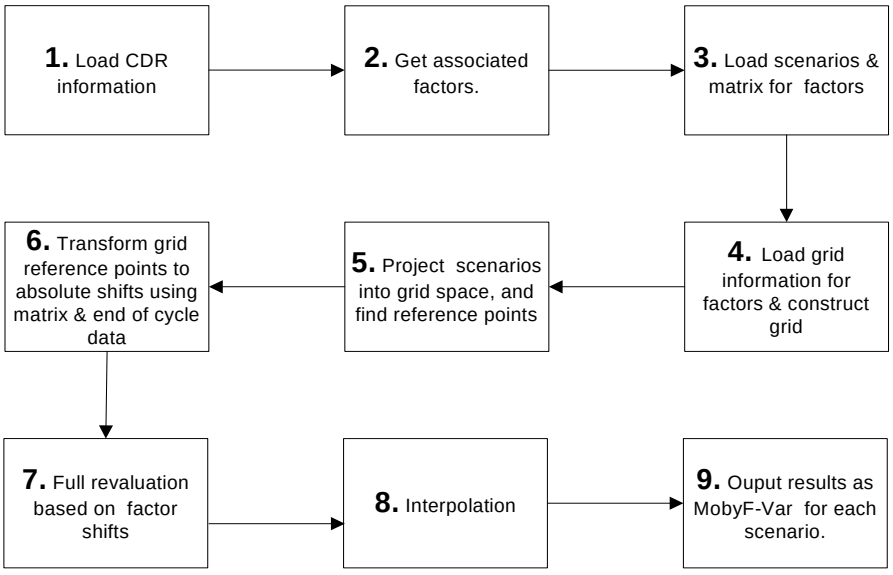

Figure 3: MobyF-VaR Process Flow

1.-Loading call information

Prerequisites:
Defining CDR information requires the creation of an MOBY portfolio, and populating the appropriate risk engine tables. The mechanism for doing that is the same as defining risk engine transaction / CDR information.

Description:
CDR information is loaded into memory. The CDRs are then divided into different groups based on the A-number and value score (VS). The VS ranges that are used for classifying these CDRs are specified in the grid index file. Note that the reason for dividing the CDRs into the different groups is that each group has its own associated risk factors and consequently, its own grid.

2.-Getting risk factors

Prerequisites:
Through the risk engine framework two associations must be defined. Firstly, the product+subproduct to curve associations must be defined for fraud aggregates, and secondly, the risk factor definitions have to be associated with the curves. These assignments will be done through the ERASE risk framework GUI.

Every group of fraud aggregates will specify the order of appearance of risk factors in a scenario. The ordered list of risk factors is obtained from the grid control file.

Finally, calibration risk factors will have to be supplied in a file. This file contains the mappings: call type + CDR range + curve type => risk factor names and tenor information.

Description:
Once the CDRs are classified, each CDR can be matched with its corresponding curves and from there, the corresponding risk factors. For non-calibration risk factors, this is done by the risk engine. For calibration risk factors, the partial revaluation engine maintains a separate mapping, which reflects the structure of the input file.

3.-Loading scenarios and matrix

Prerequisites:
Scenario files have to be prepared using the simulation engine. These scenario files are then used for preparing a CDR scenario file which contains only a few risk factors (these are related to the original risk factors through a linear transformation). The CDR scenario file is needed for every grid, and is prepared with the verify system program which executes before the run. A grid is defined for each call type currency and range. Similarly a transformation matrix is required for every grid, and is contained in the grid control file.

Description:
The partial revaluation engine loads from two types of scenario files. One contains CDR, or grid scenarios – typically having a small number of components such as 3 or 4, and used for populating the grid. One such file exists for grid definition. The other scenario file contains the full list of fraud scenarios, used for computing full revaluations. The two files are expected to contain scenarios in the same order, so that the 1st scenario in the CDR file corresponds to the 1st fraud scenario, etc. The CDR file is used for classifying scenarios, while the fraud scenarios file is used at the very least to retrieve scored values for risk ranking. The CDR matrix, like the fraud scenario file - is associated with a grid. It is loaded from the grid control file. Fraud scenario files and the CDR matrix are loaded one by one as each set of CDRs are processed.

4.-Constructing grid

Prerequisites:
A grid control file should be defined for every CDR and fraud combination.

Description
Based on the grid control file a grid is created with the appropriate number of dimensions and partitions along each dimension. The partitions do not have to be identical in every dimension, nor do they have to be evenly spaced along a dimension.

The grid is made up of logical partitions called cells. For instance, a 2-Dimensional grid will have rectangular cells. The vertices of the cells are referred to as nodes.

5.-Classifying scenarios using grid

Description:
As grid scenarios are read in, they are inserted into the corresponding grid cell. Recall that each grid dimension corresponds to a CDR risk factor. A grid cell represents a range of values for each of the CDR risk factors. Classification of a fraud scenario thus amounts to finding the range it falls into for each of its risk factors.

Once all the scenarios are classified into cells, they are divided into two groups. One group consists of scenarios which are contained within sufficiently populated cells. These scenarios will not be evaluated directly, but will instead be valued by interpolating full valuations done at the cell nodes. The other group of scenarios which resides in cells that are not sufficiently populated, will be valued using full revaluation. A reasonable criterion for a "sufficiently populated" cell is that the number of scenarios exceeds the number of nodes in the cells (Example: for a 2-Dimensional grid that number is 4).

6.-Transforming grid points to shifts

Prerequisites:
A risk ranking file has to be created and configured as a control parameter. The risk ranking is produced by exporting the last risk ranking table from the CDR database.

Description:
Using the transformation matrix, the grid node position is transformed to a fraud scenario. Grid scenarios need not be transformed to fraud scenarios since the fraud scenarios are already given in a file. When transforming a scenario then, this step is skipped. The resulting fraud scenario is expressed as relative shifts. Since the full revaluation engine requires absolute shifts, it is necessary to take into account the risk ranking in the following manner.

First, the risk factor names which correspond to the fraud scenario are retrieved from the grid control file. Using these names, the corresponding risk rankings are retrieved from the closing rate file. Then the absolute shift is computed: Absolute shift = relative shift x risk ranking x 10000 (where the relative shift and closing rates are given as a fraction of a whole (not of a percent).

7.-Performing full revaluation

Description:
Using the risk engine framework, the CDRs are valued for each set of risk factor shifts (either scenario, or grid node). The result is a list of values, one for each call.

8.-Interpolating results from full revaluations

Description:
For scenarios in populated cells, the values at the scenarios are computed using interpolation by first calculating the list of value scores (corresponding to the list of CDRs) at each of the nodes. The node values are used with multi-linear interpolation to find the scenario present values.

9.-Saving VS results

Description
Once the value scores are found for the portfolio, they are saved temporarily. Eventually, when the all scenarios are computed, the output is arranged by call and is written according to the format specified.

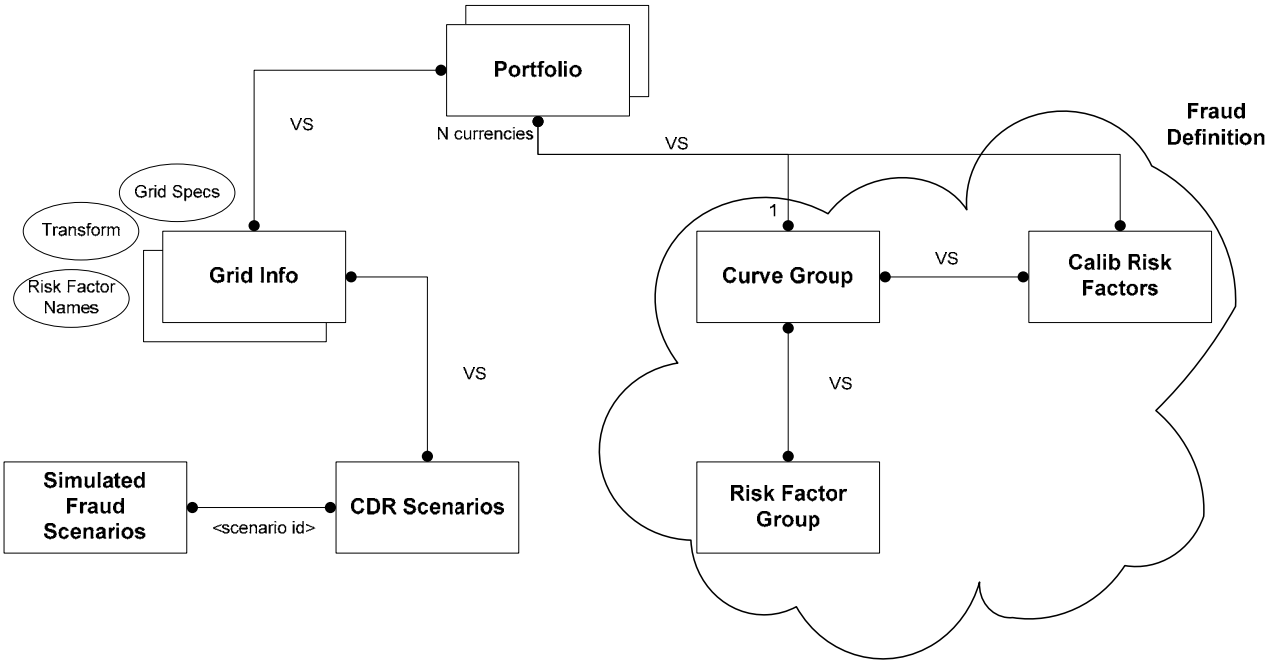

Figure 4: MobyF-Var Flow Diagram

Factor Sensitivity Grids

Factor sensitivity is the change in portfolio value resulting from a specified change in a fraud risk factor. A factor sensitivity grid is a set of factor sensitivities corresponding to a set of discrete changes in specified fraud factors. The efficiency of doing Monte Carlo simulation with factor sensitivity grids is that they can be created from a comparatively small number of exact revaluations. The exact revaluations are used to construct an interpolation table to approximate the change in portfolio value resulting from an arbitrary change in risk factors. The grid can be used to transform quickly and accurately risk factor simulations into simulated portfolio value changes. A one-dimensional factor sensitivity grid is a set of changes corresponding to discrete changes in a single risk factor.

Grid Concepts

Conceptually a grid is a partitioning of space into smaller rectangular cells. It is specified using cells and vertices (nodes). A vertex/node describes a spatial location in the risk factor space and is used for interpolation inside cells that it belongs to. Finally, a scenario which corresponds to a set of risk factor perturbations is associated with a spatial location (i.e. it is inside one of the cells). Some of the key design decisions in the grid API are based on the following observations:

- The number of cells in a grid may be very large. For example a 5-dimensional grid with partitions 7x20x7x20x15 has 294,000 cells. Since most of these cells are empty, it does not make sense for the grid to keep all possible cells, but rather only cells which are interesting (i.e. contain scenarios). The grid thus does a lazy cell insertion.

- Computing VSs is expensive and should be avoided. VSs are calculated typically at the grid nodes or for remote scenarios. Since again, for a large grid there may be many nodes, the researcher wants to compute the value at a node only when necessary. That is, only when the researcher decide to do interpolation. Moreover, once the researcher computes the VSs at a node, this result should be re-used for any interpolation which involves this node. In other words, node objects should be shared between cells. Nodes and cells must be uniquely identifiable.

- This observation is based on the previous two. In order to know whether a cell is inserted already, it must be identifiable; similarly if cells share nodes, then nodes must somehow be globally identifiable. This introduces the need for unique designations or Coordinates.

- Interpolation should be smart. The researcher does not necessarily want to guess when to use interpolation and when to use full evaluation. It is

better to give cells dynamic criteria for interpolation. Similarly, other design decisions are motivated by such observations.

Grid Characteristics

- Has 1 or more dimensions.
- Every dimension is associated with a risk factor and a standard deviation.
- The grid's resolution and range can be specified for each dimension separately.
- Maintains nodes and cells in an efficient way. Cells are added only when there is a scenario to put in the cell. Node information is shared between cells.
- Can classify scenarios into cells.
- Can identify what cells are populated and find cells with minimal population.
- Cells and nodes can be accessed by specifying their coordinates.

Cell Characteristics

- Has a list of nodes indexed by relative coordinates.
- Has a collection of scenarios.
- Can linearly interpolate the value at a scenario based on node values.
- Has coordinates which in each grid dimension have the range (0, # partitions - 1).

Node Characteristics

- Has spatial location.
- Hold values used for interpolation. Those values are in fact lists of VSs.
- Has coordinates which in each grid dimension have the range (0, # partitions).

Coordinates Characteristics

- Contains a coordinate for each dimension of the grid.
- Can be compared to other coordinates, added, and subtracted.

Scenario Characteristics

- Represents location in grid space.
- Has an id which represents the scenario sequence number in the input file.

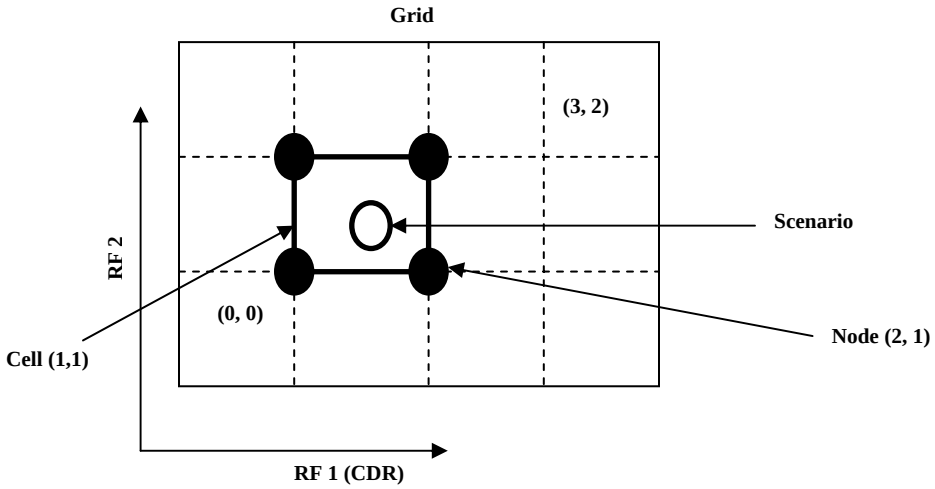

Figure 5: Grid representation example

Monte Carlo Simulation Engine

Monte Carlo Simulation Engine generates two sets of globally correlated risk factor scenarios with correlated multivariate stochastic variance. These scenarios represent random short term and historically relative changes of the corresponding underlying and implied volatilities. The globally correlated Monte Carlo scenarios for each holding period are stored in two Monte Carlo Scenario Files. These scenarios are regenerated by the System after each re-calibration and updating of the simulation model parameters by fraud risk.

Principal Component Analysis

Elements of Principal Component Analysis (PCA) to be used in MobyF-VaR :

- Construction of the risk factor covariance matrix.
- Covariance matrix diagonalization and eigenstructure.
- Principal component analysis.
- Truncation of market state vectors in principal component representation.
- Grid interpolation in a principal component representation.

Bhansali and Kokoska (2001) describe the motivation and essential mathematical background required for PCA. This section describes the creation of the covariance matrix from a set of risk factor observations.

Assume a set of *(n+1)* observations on *p* risk factors. Assume that the observations have occurred at times $\{t_0, t_1, \Lambda\ , t_n\}$ and are ordered so that $t_0 > t_1 > \Lambda\ > t_n$. That is, the observations are ordered with t_0 the most recent

observation. Assume all risk factors are observed simultaneously. Label the risk factors by $\{X_1, \Lambda\ , X_p\}$ and denote the j^{th} observation of the i^{th} risk factor by $x_{i,j}$.

Each risk factor therefore has the associated time series

$$\underset{\rho}{x}_1 = \begin{pmatrix} x_{1,0} \\ x_{1,1} \\ M \\ x_{1,n} \end{pmatrix}, \quad \underset{\rho}{x}_2 = \begin{pmatrix} x_{2,0} \\ x_{2,1} \\ M \\ x_{2,n} \end{pmatrix}, \quad \Lambda \ \underset{\rho}{x}_p = \begin{pmatrix} x_{p,0} \\ x_{p,1} \\ M \\ x_{p,n} \end{pmatrix}. \tag{2.1}$$

The *one-day relative return* for each risk factor is computed as

$$\chi_{i,j} \equiv \frac{x_{i,j} - x_{i,j-1}}{x_{i,j-1}}, \quad i = 1,..., p; \ j = 1,..., n. \tag{2.2}$$

A set of *one-day logarithmic returns* is similarly computed as

$$\phi_{i,j} \equiv \ln \frac{x_{i,j}}{x_{i,j-1}}, \quad i = 1,..., p; \ j = 1,..., n. \tag{2.3}$$

Unless otherwise mentioned the researcher shall always use relative returns. In the usual notation denote the mean and standard deviation of the time series of one-day relative returns by

$$\mu_i = \frac{1}{n} \sum_{j=1}^{n} \chi_{i,j},$$

$$\sigma_i^2 = \frac{1}{n-1} \sum_{j=1}^{n} \left(\chi_{i,j} - \mu_i\right)^2. \tag{2.4}$$

The researcher now defines a time series of mean-subtracted, variance-normalized one-day relative returns by

$$\xi_{i,j} \equiv \frac{\chi_{i,j} - \mu_i}{\sigma_i}, \quad i = 1,..., p; \ j = 1,..., n. \tag{2.5}$$

To obtain the time series

$$\underset{\xi_1}{\rho} = \begin{pmatrix} \xi_{1,1} \\ \xi_{1,2} \\ M \\ \xi_{1,n} \end{pmatrix}, \quad \underset{\xi_2}{\rho} = \begin{pmatrix} \xi_{2,1} \\ \xi_{2,2} \\ M \\ \xi_{2,n} \end{pmatrix}, \quad \Lambda \quad \underset{\xi_p}{\rho} = \begin{pmatrix} \xi_{p,1} \\ \xi_{p,2} \\ M \\ \xi_{p,n} \end{pmatrix}. \tag{2.6}$$

This is the representation of the events the researcher shall generally work with. Geometrically this corresponds to a translation and scaling of the set of *n p*-dimensional vectors which translates the set to its center of mass and scales the dispersion to enforce unit variance.

This allows the definition of the *(nxp)* matrix of mean-subtracted variance-normalized relative one-day returns by

$$\Xi = \begin{pmatrix} \xi_{1,1} & \Lambda & \xi_{1,p} \\ M & O & M \\ \xi_{n,1} & \Lambda & \xi_{n,p} \end{pmatrix}. \tag{2.7}$$

The covariance matrix is the *(pxp)* matrix defined by (ex, Bhansali [[11] pg. 243)

$$\Sigma = \frac{1}{n} \Xi^T \Xi \tag{2.8}$$

With elements

$$\Sigma_{ij} = \frac{1}{n} \sum_{k=1}^{n} \xi_{k,i} \xi_{k,j} \, . \tag{2.9}$$

The covariance matrix is real and symmetric since

$$\begin{aligned} \Sigma_{ij} &= \frac{1}{n} \sum_{k=1}^{n} \xi_{k,i} \xi_{k,j} \\ &= \frac{1}{n} \sum_{k=1}^{n} \xi_{k,j} \xi_{k,i} \, . \\ &= \Sigma_{ji} \end{aligned} \tag{2.10}$$

And hence $\Sigma = \Sigma^T$.

Covariance Matrix Diagonalization and Eigenstructure

A well-known result from linear algebra states that under sufficiently general conditions the eigenvalue decomposition

$$\Sigma = U\Lambda U^T \tag{2.11}$$

of the real symmetric matrix Σ exists, where

$$\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & O & \\ 0 & & \lambda_p \end{pmatrix} \tag{2.12}$$

is a diagonal matrix of real eigenvalues arranged in order of decreasing size $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$. The columns of the matrix U are unit eigenvectors belonging to the corresponding eigenvalues. That is, the researcher can write

$$U = \begin{pmatrix} u_{11} & \Lambda & u_{1p} \\ M & O & M \\ u_{p1} & \Lambda & u_{pp} \end{pmatrix} = \left(\hat{u}_1, \hat{u}_2, ..., \hat{u}_p\right) \tag{2.13}$$

where

$$U\hat{u}_i = \lambda_i \hat{u}_i . \tag{2.14}$$

A standard result is that the largest eigenvector corresponds to the variance of the first principal component and is the direction that accounts for the largest variance in the set of *p* risk factors.

Principal Component Analysis

The PCA representation allows data to be decomposed into a low-dimensional subspace that captures the bulk of the observed variance, and the complementary subspace that accounts for a relatively small amount of the total variance. In this way high-dimensional problems can be reduced to low-dimensionality, increasing computational efficiency with only a small penalty in explaining observed variance. Bhansali discusses the strengths and weaknesses of PCA in strategic risk management systems.

It will be convenient to establish the transformation rules between the original market state representation and the principal component representation.

Consider an arbitrary fraud state vector $\underset{\rho}{x} = (x_1,...,x_p)^T$ in the original fraud state space. The transformation from the fraud state representation to the principal component representation is a direct application of the eigenvector matrix U:

$$\underset{\rho}{y} = U^T \underset{\rho}{x} \tag{2.15}$$

or

$$y_i = \sum_{j=1}^{p} u_{ji} x_j,\ i = 1,...,p. \tag{2.16}$$

Equation (2.15) follows from

$$\begin{aligned} E\left[\underset{\rho}{y}\underset{\rho}{y}^T\right] &= E\left[U^T \underset{\rho}{x}\underset{\rho}{x}^T U\right] \\ &= U^T E\left[\underset{\rho}{x}\underset{\rho}{x}^T\right] U \\ &= U^T \Sigma U \\ &= U^T \left(U \Lambda U^T\right) U \\ &= \Lambda \end{aligned} \tag{2.17}$$

Since the columns of U are eigenvectors forming an orthogonormal basis in risk space, the market state vector can be represented in this basis as

$$\underset{\rho}{y} = \left(\hat{u}_1, \hat{u}_2, \Lambda\ , \hat{u}_p\right)^T \underset{\rho}{x}. \tag{2.18}$$

Define the cutoff dimension $\tilde{p} \le p$ as the truncation point in the PCA. The cutoff dimension can be calculated according to several criteria. For example it may be chosen to explain a prescribed percentage of the observed variance, k. Then $\tilde{p}$ is the smallest integer such that

$$\frac{\sum_{i=1}^{\tilde{p}} \lambda_i^2}{\sum_{i=1}^{p} \lambda_i^2} \ge k. \tag{2.19}$$

An alternative approach exploits the fact that the original time series has unit variance by construction (A2.5). Hence any principal component belonging to an eigenvalue of size less than one has less predictive power than the original market risk factor. Consequently it can be omitted without unduly compromising the variance of the original data.

PCA is based on the hypothesis that the dynamics with the truncated $\tilde{p}$-dimensional state vectors in the PC representation are a valid representation of the original dynamics. Choosing a low truncation dimension makes the analysis much simpler computationally.

Denote by $\tilde{\Sigma}$ the $\tilde{p}$-dimensional covariance matrix generated from:

$$\tilde{\Lambda} = \begin{pmatrix} \tilde{\lambda}_1 & & 0 \\ & \mathrm{O} & \\ 0 & & \tilde{\lambda}_p \end{pmatrix} \tag{2.20}$$

where

$$\tilde{\lambda}_j = \begin{cases} \lambda_j, & j \geq n - \tilde{p} \\ 0, & j < n - \tilde{p} \end{cases} \tag{2.21}$$

The modified covariance matrix is built in two steps. The first recovers

$$\tilde{\Sigma}^{(1)} = U\tilde{\Lambda}U^T \tag{2.22}$$

This matrix will not in general have the proper values along the diagonal. Elements must be rescaled to give unit values along the diagonal of the corresponding correlation matrix. This procedure is discussed in Best and results in $\tilde{\Sigma}$.

Description of Results

The use of multivariate analysis and the engagement of grid based ranking by way of matrix techniques provided an improvement in both the number of events detected and the accuracy by way of a reduction of false positives as compared to the baseline.

The efficiencies provided by the algorithm and the consequent reduction in computational cycles improved performance of the overall system in a reduction of the time to process. The methods and algorithms were evaluated using an industry tool for fraud mitigation, known as ERASE. This tool is a framework composed of a data repository and baseline rules engine. The researcher operationalized this environment so as to employ known industry parameters, from which comparisons could be made. Tests were performed on a hardware and software environment.

Fraud Scenarios	Baseline ERASE	% Detection	% false positive	MobyF-VaR	% Detection	% false positive
Cloning	22	0.011	n/a	26	0.013	n/a
Time Variance	117	0.058	n/a	192	0.096	n/a
Link Analysis	147	0.074	n/a	239	0.119	n/a
Shark Attacks	72	0.036	n/a	86	0.043	n/a
Subscription	376	0.188	n/a	483	0.241	n/a
Totals	734		0.072	1026		0.042

Figure 6: MobyF-VaR approach using CDR data comprising 200,000 records in a baseline comparison of event detection.

Fraud Scenarios	Time to Process Baseline ERASE	Time to Process MobyF-VaR
Mixed	2.1 hours	0.43 hours

Figure 7: MobyF-VaR approach using CDR data comprising 200,000 records in a baseline comparison of Time to Process

Using this sample and specified evaluation of the methods, the researcher found indications of improvement over pre-existing approaches to fraud analytics. The entire spectrum of fraud in mobile telecommunications has by no means been exposed to the approach.

The use of computational analysis is not novel, however the unique application of grid and multivariate techniques in a sequential method whilst supported by a generic framework proved to be positive in effect. Reduction of processor utilization infers the reduction in calculation and number of factors required to detect an event representative of fraud.

Comparative methods have only increased the demand for computational resources. The researcher holds that the two fold benefit of reduced number of required factors and the corresponding performance advances is an

improvement over comparative methods having only increased the demand for computational resources.

Prior to this effort the characteristics of a VoIP call were loosely defined and seldom represented by a standard presentation or format. Based upon the requirements set by the ITU, manufacturers of telecommunications switches provided a binary interface from which to derive a representation of the call within a CDR.

The GSM convention for CDR composition is amongst the most rigorous and can be consistently derived. The researcher sought to be able to systematically transpose and parse these binaries into corresponding data factors, sets and relationships. The researcher struggled to attain a consistent output amongst various telecommunication operators and realized that although the standard specification had been agreed, the interpretation of the data fields, codex representation and use were disparate.

The researcher hypothesized that by creating a mediation layer and composite parser the ability to extract data from switch based call transactions would be composed into CDR formats according to the GSM conventions of the ITU. The initial data sets were inconsistent and had little value in the specified factors sought. The researcher modified the parsers and established several hierarchies. This proved useful in enabling the selection of data sets comprising the various data factors. The researcher established J48 decision trees and representative models to ascertain the composition of the factors, relative weights and degree of variance.

Several data factors began to indicate a representation of VoIP events. The GSM call specification for CDR provided a particular identification; namely the capability to differentiate a GSM vs. GPRS call initialization. Typically a GSM call is a voice call at its origin and a GSM voice call at its termination. A GSM originating call could become a data call based on IP over GPRS. A specific identifier if used by the switch data would provide clear identification.

The researcher extracted binaries and parsed the data by specified data factors. A data set was composed of GSM originating calls having data services associated with these call events. The researcher recognized that 3 particular services might compose a VoIP call; namely an IP based service call, an i-mode based service call and an information service interconnect call. Basic analysis showed that in a population subset of 4886 events, 3290 were GSM originated and GSM terminated calls. 1076 events represented GSM based calls comprised of IP services based on the identifier of GPRS network subtype as an outcome data factor. This showed that the researcher could further segregate and bin the terminating services. The three possible outcomes showed 119 IP based terminations, 14 data service requests for data, and 2 i-mode gaming requests.

However, the researcher further ascertained that only initiations having a bi-directional movement of IP packets were of interest. This is simply because any initiated call having a data factor and quantifier for volume up, without a corresponding volume down, is either an upload of data via an unlikely channel, a non-terminated call (no answer), or a dropped call.

This data set and corresponding data factors were tested for kappa and correlation via J48. A kappa of 0.93 was attained, indicating the two factor correlation between the extracted data hierarchies and the measures derived.

Further Considerations

The researcher began to operationalize the findings as they did conform to a node based analysis. The researcher struggled to establish a causal relationship related to fraud. Typical mobile calls could indeed be characterized via CDR data factors alone. However VoIP events may require additional data points.

Consequent review of these findings indicates that additional efforts to refine the factors and their application as indices, to which both rules and probabilistic scenarios can be directed, may augment the efficiencies of fraud detection systems. This will afford advances in the cost / performance of similar systems and provide a generic methodology to support high performance fraud detection systems.

The researcher recognized that alternate data sets representing VoIP events would be further improved by upstream data. This would be collected at network IP switch level by employing active agents. The models and algorithms found to be operationally indicative of events representing potential fraud may be deployed as run time models within agents. The researcher holds that these additional data points could significantly improve the accuracy required in a real time environment. Moreover the potential of implementing these agents within handsets, smart phones and PDA's would certainly afford data at the source level.

The researcher will continue to evaluate additional scenarios for factor impact on performance and discovery of additional effects to increase segregation of call groups and their effect on the output accuracy and time to process. The researcher will continue to evaluate the use of agents and sensor arrays to improve the collection of data beyond the telecom switch level.

About the Author

Ivan Zasarsky leads the Forensic Database Analytics practice at Deloitte. Ivan joined Deloitte in December 2003 as a Technology Media and Telecommunications team member for R&D in financial services for some of the largest accounts. A winner of Deloitte Fast 50, he has advanced standards for wireless and database technologies on WAP Forum, OAG, Microsoft Leaders Group, Best on Oracle.

Prior to Deloitte, Ivan was instrumental in establishing bridgeheads for emerging technologies including, anti-money laundering. Ivan was involved in several global Analytics initiatives and was co-founder of Canadian/Dutch B2B, an E-commerce Company with both product and ASP offerings.

Ivan holds a Master of Science degree in Economic Crime Management from Utica College, and has published research in several areas, including technologies and methods for fraud mitigation in mobile and VoIP communications.

Appendix: Data and computer programs DISK ECM-001-2005

A CD-ROM is available from the author for the generation of results and use of generated data sets.

A construction of specialized parsers written in JAVA was particularly useful in the study of this field. A limited use license was granted for the purposes of research only.

A composite data set extracted from binaries of telecommunication switches is available with primary composition and analysis.

A composite model employing a J48 algorithm for these purposes is also available.

References

Bhansali, R.J., Kokoska, P.S. (2001). Prediction of long term memory over time series: An overview. University of Utah.

Bharat Bhargava, Yuhui Zhong, Yuhua Lu.(2003). *Fraud Formalization and Detection.* Center for Education and Research in Information Assurance and Security And Department of Computer Sciences Purdue University.

Boukerche Azzedine.(2000).Behavior-Based Intrusion Detection in Mobile Phone System. *Journal of Parallel and Distributed Computing 62*, 1476–1490.

Boukerche, Azzedine. and Notare M. S. M. A.(2000). Neural Fraud Detection in Mobile Phone Operations. *Lectures notes in computer science parallel and distributed processing, XV IPDPS/Bio3SP*,636–644.

Bowers, Simon.(2004, July 26).Suspected Fraudsters Move Into Mobile Phone Sales. *The Guardian.*

Burge, Peter.and J. Shawe-Taylor.(1997). Detecting cellular fraud using adaptive prototypes. *Proceedings of AAAI-97 Workshop on AI Approaches to Fraud Detection & Risk Management*, 9–13.

Burge Peter and J. Shawe-Taylor. (2000).*Unsupervised Neural Network Approach to Profiling the Behavior of Mobile Phone Users for Use in Fraud Detection.* Department of Computer Science, Royal Holloway University of London.

Clarke, R. V, & Kemper R., Wyckoff L. (2001).Controlling cell phone fraud in the US: Lessons for the UK foresight prevention initiative. *Security Journal Volume: 14 Issue:1*, 7 to 22.

Collins Michael.(2000).Telecommunications Fraud Part 3.*Computers and Security, 19,*141-148.

Commission of the European Communities.(2001, September).*Preventing fraud and counterfeiting of non-cash means of payment*.(Publication No.COM (2001)11 Final). Retrieved October17, 2005,from The European Parliament, The European Central Bank, The Economic and Social Committee and Europol. Online via Commission to the Council Access: http://www.securitymanagement.com/library/EU_Fraud0501.pdf

ESANN'1999 proceedings.(1999, April). A hybrid system for fraud detection in mobile communications. *European Symposium on Artificial Neural Networks Bruges (Belgium), 21-23*, 447-454.

Harrington, V. & Mayhew P.(2001, December). *Mobile Phone Theft* (Home Office Research Study 235). Home Office Research, Development and Statistics Directorate.

Hoath, P. (1998, January).Telecoms fraud, the gory details. *ComputerFraud & Security* 20(1), 10-14.

Hollmen Jaakko, and Tresp Volker.(2000).*Call-based Fraud Detection in Mobile Communication Networks using a Hierarchical Regime-Switching Model*. Helsinki University of Technology Lab of Computer and Information Science.

Hollmén Jaakko.(1997).*Novelty filter for fraud detection in mobile communications networks .* (Technical Report A48). Helsinki University of Technology, Laboratory of Computer and Information Science.

International Telecommunication Union. (2003). *WTI Statistics 2003.* Market, Economics and Finance Unit, Telecommunication Development Bureau.

Johnson, M. (1996, December).Cause and effect of telecoms fraud. *Telecommunication (International Edition) 30(12),* 80-84.

Joint Economic Committee United States Congress.(2002, May). *Security in the information age: New challenges, new strategies.* Joint Economic Committee.

Levin, Alex and Tchernitser, Alexander. (August 2001).*Multifactor stochastic variance models in risk management: Maximum entropy approach and Levy processes.* Bank of Montreal Working Paper, No. 07/2002.

Leyden, John.(2003, September 4).*Israeli Boffins Crack GSM Code.* The Register.

Lucyshyn W. & Richardson R.(2004).*CSI/FBI Computer Crime and Security Survey.* Computer Security Institute.

Miller, Christina.(2003, April).More Than a Cell Phone. *Law Enforcement Technology Volume: 30 Issue: 4,*82, 84 to 86.

Moreau,Y. and Vandewalle, J. (1997).Detection of mobile fraud using supervised networks: A first prototype. *ICANN 1997: 1065-1070.*

National White Collar Crime Center and the Federal Bureau of Investigation.(2003). *IFCC 2002 Internet Fraud Report.* (January 1, 2002—December 31, 2002).

O'Shea, D. (1997, January).Beating the bugs: Telecom fraud. *Tele-phony 232(3),* 24.

Open Mobile Alliance. (2004 April).*Wireless Identity Module.*(Candidate Version 1.2).

Parliament of Victoria Drugs and Crime Prevention.(2004).*Committee Inquiry into Fraud and Electronic Commerce Final Report.* Government Printer for the State of Victoria. (No. 55 Session 2003-2004).

Parliamentary Office of Science and Technology (1995).*Mobile Telephone Crime.* (POST Note 64 *extension 2840*).

Picoult, E.(1997, August).Calculating VaR at Risk with Monte Carlo Simulation. *Preprint Risk Magazine / CitiBank*.

Racal Research Ltd. (1988). Technical Information GSM System Security. (10-1617-01).

Rantala, R.R. & Edwards, T.J.(July 2000). *Effects of NIBRS on Crime Statistics.* (NCJ 178890).

Seminar on Economic and Market Analysis for Central and Eastern European countries (CEEC) and Baltic States. (2003, September). Czech Republic, Prague.

Smith, R.G.(1998). *Preventing Mobile Telephone Crime.* Australian Institute of Criminology

Smith, , R.J.(2000). *Session T39—Electronic Fraud.* Australian Society of Certified Practicing Accountants, Australian Institute of Criminology.

Taniguchi Michiaki, Haft Michael, Hollmén Jaako,Tresp Volker (1998). *Fraud detection in communication networks using neural and probabilistic methods*. Siemens AG Corporate Technology.

U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics. Technical Report. (2001). *Cybercrime against businesses: Pilot test results Computer Security Survey* (March 2004, NCJ 200639).