



2005-2006



# Discovering Telecom Fraud Situations through Mining Anomalous Behavior Patterns

**Ronnie Alves, Pedro Ferreira, Orlando Belo,  
João Lopes, Joel Ribeiro and Luis Cortesão**

Department of Informatics  
School of Engineering  
University of Minho  
PORTUGAL

- In this paper we tackle the problem of **superimposed fraud detection** in telecommunication systems. We propose two **anomaly detection methods** based on the concept of **signatures**.

Method 1 : based on a signature deviation-based;

Method 2 : based on dynamic clustering analysis;



- Scenario
- Orientation
- Signatures & Summaries
- Deviation-based Approach
- Dynamic Clustering Analysis
- Evaluation Study
- Final Remarks



- *Telecom companies generate in average ~ 2.5 millions of CDRs per week;*
- *The fraud detection process is still a laborious process performed essentially through manual inspection;*
- *Up to now, there is not any accurate database with previous detected cases of fraud;*



- How can **we** summarize the CDRs on **customer basis** without loss of information?
- How can **we** detect **anomalous situations** which could represent interesting situations?
- How can **we** **smooth the number of raised alarms (decrease False Positives)** in order to focus the analysis on the interesting alarms?
  
- *Our goal is to **detect deviate behaviors** in useful time, giving better basis to analysts to be more accurate in their decisions in the **establishment of potential fraud situations**.*

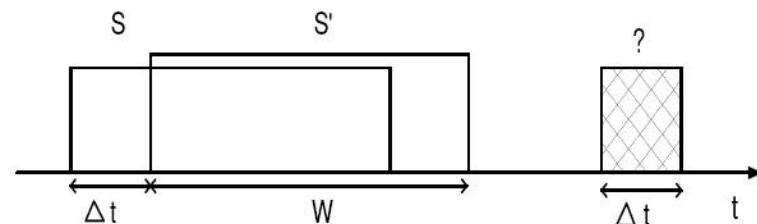


- What are signatures?
- What are summaries?
- What about the updating process?

| Description   | Type    |
|---|---------|
| Duration of Calls                                   | Complex |
| Number of Calls - Working days                      | Complex |
| Number of Calls - Weekends and Holidays             | Complex |
| Number of Calls - Working Time (8h-20h)             | Complex |
| Number of Calls - Night Time (20h-8h)               | Complex |
| Number of Calls to the Different national networks* | Simple  |
| Number of Calls as Caller (Origin)                  | Simple  |
| Number of Calls as Called (Destination)             | Simple  |
| Number of International Calls                       | Simple  |
| Number of Calls as Caller in Roaming                | Simple  |
| Number of Calls as Called in Roaming                | Simple  |

$$S_{t+1} = \beta \cdot S_t + (1 - \beta) \cdot c$$

- Signatures are typically calculated for a temporal period of one week; Summaries for a period of one day.



- Measuring similarity of simple variables

$$d(S_x, S_y) = e^{-\left\{ \frac{|S_x - S_y| \times B}{Amp} \right\}}$$

- Measuring similarity of complex variables

$$d(C_x, C_y) = d(M_x, M_y) \times \frac{|C_x \cap C_y|}{|C_x \cup C_y|}$$

- Calculating the distance among signatures

$$D(S, C) = \sqrt{\alpha_1 \cdot f_1(S_1, C_1)^2 + \dots + \alpha_n \cdot f_n(S_n, C_n)^2}$$

$$\mathbf{Dist(S, C) = MAX\{dist_1(S, C), dist_2(S, C), \dots, dist_m(S, C)\}}$$

- Raising alarms

$$Dist(S_t, C) > \xi$$



- Providing the similarity matrix **based on the customer signatures**

$$D(X, Y) = \sqrt{W_1 \cdot d_1(X_1, Y_1)^2 + \dots + W_n \cdot d_n(X_n, Y_n)^2}$$

- Comparison of signatures against cluster centroids

- Relative similarity

$$\Delta = \left\{ 1 - \frac{D(S_i, \text{SignCl}[S_i]_{t+1})}{D(S_i, \text{SignCl}[S_i]_t)} \right\} \times 100\%$$

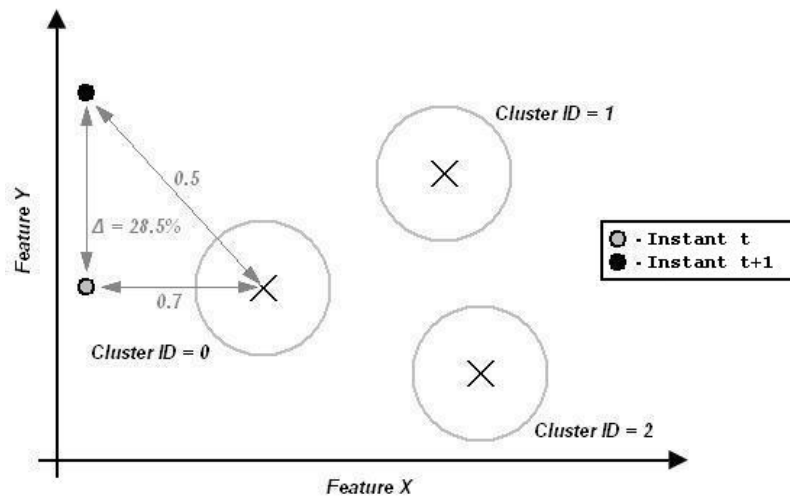
- Absolute similarity

$$D(X, Y) = \sqrt{W_1 \cdot d_1(X_1, Y_1)^2 + \dots + W_n \cdot d_n(X_n, Y_n)^2}$$

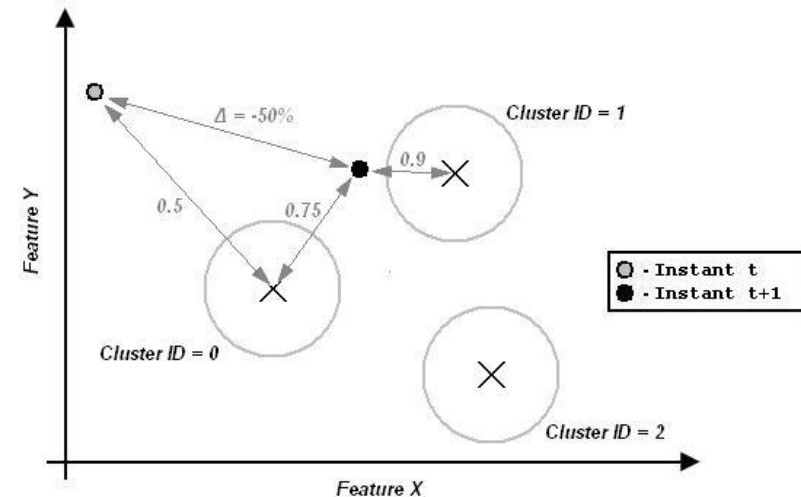


- Detecting changes on clusters membership
  - a signature  $S$  changes its cluster membership to cluster  $C_j$  in the instant  $t+1$ , if it belongs to cluster  $C_i$  in the instant  $t$ , in the instant  $t+1$  the distance  $D(S, C_j)$  is minimal concerning all clusters and  $D(S, C_j)_{t+1} < D(S, C_i)_t$

(+) variation



(-) variation

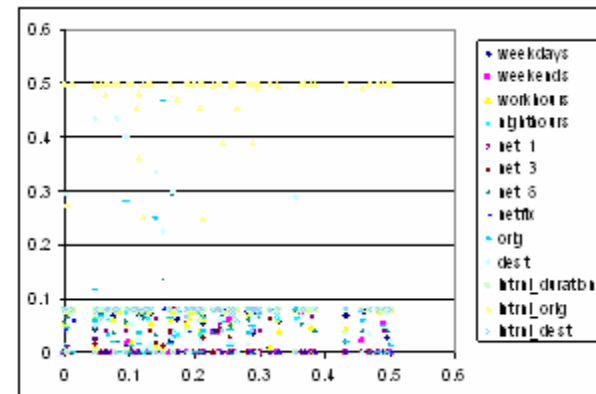
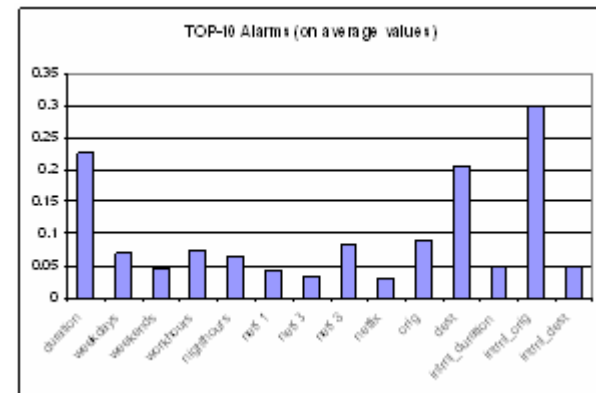


- The complete set of CDRs corresponds to approximately 2.5 millions of records
- 700 thousand of signatures processed per day
- the settings of our methods were guided by a small list of 12 customers (fraudsters in the referenced week)
- we worked on a subset of the previous data concerning to a sample distribution with approximately 5 thousands summaries per day



- Deviation-based results
  - imperative anomalous situations were raging from **2.76 up to 3.33**
  - The feature variable which has more impact over the distance calculation (**TOP-10 Alarms**) is the *international call*
  - *workhours* variable has great importance to the distance calculation (complete set of raised alarms)

| (€)/day | 0.8  | 1.0  | 1.2 | 1.6 | 2.0 |
|---------|------|------|-----|-----|-----|
| Tue     | 2141 | 649  | 139 | 50  | 25  |
| Wed     | 3029 | 1145 | 251 | 103 | 56  |
| Sat     | 1006 | 560  | 150 | 39  | 23  |



- Dynamic clustering results
  - the qualities of the clusters were maximized with 8 clusters

| Cluster | Tue | Wed | Sat |
|---------|-----|-----|-----|
| 1       | 3   | 9   | 1   |
| 2       | 9   | 7   | 123 |
| 3       | 3   | 12  | 71  |
| 4       | 5   | 17  | 16  |
| 5       | 23  | 21  | 22  |
| 6       | 20  | 31  | 40  |
| 7       | 8   | 11  | 26  |
| 8       | 52  | 72  | 0   |

(+) 909843678  
 3 2 F 0.886 0.0  
 4 8 V 0.829 8.91 (A)  
 5 5 V 0.871 -0.54  
 6 6 V 0.939 -7.75

(+) 909660610  
 2 1 F 0.895 0.0  
 3 8 V 0.84 8.86 (A)  
 4 7 V 0.863 2.29  
 5 3 V 0.929 -7.5

909892861  
 4 8 F 0.87 0.0  
 5 8 F 0.821 5.6  
 6 8 F 0.897 -9.31 (A)  
 7 7 V 0.946 -4.98



- Given the small list provided by the fraud analysts we can report a recall of **75%** (by the 1st method) and **91%** (by the 2nd method)
- The **relative similarity** measure (by the 2nd Method) provides a fine tuning by exploring signatures relative variation
- the **overlap rate** of both methods corresponds to approximately 62% for the whole sample used, and 66% for the blacklist
- next efforts will be the development of a **database of fraud cases**, as well as, an **induction rule engine** to help analyst on the evaluation of the raised alarms
- Graph mining (**fraud communities**) & Time series Analysis





2005-2006



## Discovering Telecom Fraud Situations through Mining Anomalous Behavior Patterns

**Ronnie Alves, Pedro Ferreira, Orlando Belo,  
João Lopes, Joel Ribeiro and Luis Cortesão**

Department of Informatics  
School of Engineering  
University of Minho  
PORTUGAL

websites

<http://www.di.uminho.pt/fratelo/>

<http://www.di.uminho.pt/~omb/>

<http://alfa.di.uminho.pt/~ronnie>

<http://alfa.di.uminho.pt/~pedrogabriel>



**???? Questions ????**