

# Synthetic logs generator for fraud detection in mobile transfer services

C. Gaber<sup>1,2</sup>, B. Hemery<sup>2</sup>, M. Achemlal<sup>1,2</sup>, M. Pasquet<sup>2</sup>, and P. Urien<sup>3</sup>

<sup>1</sup> Orange Labs, France Telecom, 42 rue des Coutures, BP 6243, F-14066 Caen, France,

<sup>2</sup> UNICAEN, ENSICAEN, CNRS, UMR 6072 GREYC, F-14032 Caen, France

<sup>3</sup> Telecom Paristech, UMR 5141, 37/39 rue Dareau 75014, Paris, France

{chrystel.gaber, mohammed.achemlal}@orange.com

Mobile payments become more and more popular and, thus, are very attractive targets for fraudsters. As those always find new ways to commit crimes and avoid detection, research in the field of fraud is always evolving. Yet, research in fraud detection is limited as publicly available transactional databases containing frauds and groundtruth are scarce [1]. The main cause is that stakeholders like banks are very reluctant to disclose information about frauds and their clients.

We address this issue by generating synthetic data of a mobile transaction system. As this work was done in the scope of the European FP7 project MASSIF, our model is based on the mobile-based transaction system and its users described by the MASSIF scenario providers in [2]. Our aim is to create a tool which can be used by the project contributors and the fraud detection community. Synthetic data are not commonly used in the field of fraud detection although there is a lack of test data. To our knowledge, only one method is used to generate synthetic data for the training and testing of fraud detection systems [3]. Compared to this method, ours enables to highlight specific characteristics of fraud detection algorithms because it is not necessarily set up with parameters driven from real data.

Our simulator simulates the mobile money transfer platform as well as the behavior of regular users and fraudsters. The regular users behavior is modeled as a set of habits whereas the fraudsters behavior is based on attack patterns. Only superimposed frauds were modeled [1]. A first prototype based on multi-agent models was implemented. We carried out a preliminary evaluation of the model and the data created with the implemented simulator. The period and amount parameters tend to be overfitted. This leads to less frequent transactions of higher value but a correct value of the total amount of money spent during the simulation. The comparison of synthetic and real data shows encouraging results and sets a landmark in the field of synthetic logs generation.

## References

1. Richard J. Bolton and David J. Hand. Unsupervised profiling methods for fraud detection. In *Conference on credit scoring and credit control*, 2001.
2. Mohammed Achemlal, Saïd Gharout, Chrystel Gaber, Marc Llanes, Elsa Prieto, Rodrigo Diaz, Luigi Coppelino, Antonio Sergio, Rosario Cristaldi, Andrew Hutchison, and Keiran Dennie. Scenario requirements. <http://www.massif-project.eu/>, March 2011.
3. E.L. Barse, H. Kvarnstrom, and E. Jonsson. Synthesizing test data for fraud detection systems. In *Computer Security Applications Conference, 2003. Proceedings. 19th Annual*, pages 384 – 394. dec. 2003.