B.2 Scientific description

Title: **Graphomaly** – software package for anomaly detection in graphs modeling financial transactions

B. 2.1 Project Scope and Objectives

General context and purpose

The value of fraudulent transactions in the EU was recently about 1.8 billion euro yearly only for card fraud. Many other financial criminal activities, like money laundering (e.g. through multiple or over-invoicing), corruption-related transfers, VAT evasion, identity theft, etc., affect individuals, banks and state activities like tax collection. As the number of transactions increases and criminal behavior becomes more sophisticated, fraud detection requires more attention and time from human experts employed by banks or state authorities. The need of performant automatic tools for at least selecting the most likely fraudulent activities, but aiming also to detect new types of ill-intentioned activities, is imperative. *Anomaly detection* is the general topic under which such a tool can be categorized in computer science.

Money transactions (payments, transfers, cash withdrawals, etc.) can be described by a vector of characteristics. Suspect transfers—the anomalies—may be the outliers in a given set of training vectors. However, treating transactions as independent vectors is an over-simplification, due to the intricacies of many types of criminal behavior. It is much more adequate to treat the transactions in their natural form, that of a *graph* whose nodes are the financial entities (individuals, firms, banks) and whose edges are transactions data (amount, time, payment mode, etc.). Graphs allow to model inter-dependencies, capture the relational nature of transactions and are also a more robust tool, as fraudsters usually do not have a global view of the graph.

So, the general umbrella for our project is *anomaly detection (AD) in graphs* and our purpose is to detect fraudulent financial activities by investigating graphs of financial transactions. Describing the financial transactions of a given institution (bank, state, cash-machine network) within a time-frame (a month, a trimester, a year) leads to a very large graph.

The goal: a library for anomaly detection in graphs modeling financial data

The main goal of this project is to create a toolbox, called Graphomaly, for anomaly detection in financial data modeled as graphs. Given a large dataset of transactions, the toolbox should be able to produce the subset of anomalous data in the following basic scenarios.

(i) Finding static patterns

Prior knowledge from financial entities about financial fraud are expressed as static transaction patterns and are represented as a small graph connecting different entities. Some examples are shown in Figure 1; for instance, a ring (B) or a clique (D) may be a sign of money laundering and directed multipartite graph (E) may show the money flow of an illegal network; note that the size of these structures can vary.

We want to look for these known static patterns within the existing transactions in order to identify possible frauds. Thus, the first problem that we focus on is identifying patterns, or sub-graphs, in a given graph.

Splitting the transactions into separate time-frames might break up such patterns, thus overlapping time-frames have to be taken into consideration.

(ii) Finding patterns without prior knowledge

Known financial fraud schemes often change when they start being detected by the authorities, turning this into a cat-and-mouse game where the static patterns from **(i)** are a reaction to the existing schemes identified in the market.

Becoming proactive implies looking at transaction data without the help of any prior knowledge such as known patterns or other inside information from the authorities or the institutions. Data labeled by experts are not available.

In machine learning this task is called unsupervised learning. Even with machine learning (ML) algorithms this is more challenging than **(i)** but, if successful, it has the benefit of providing new insight into present money-

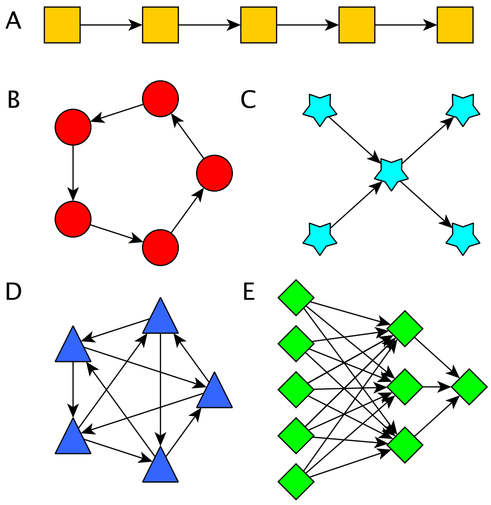


Figure 1: Static patterns of some financial fraud schemes [11]

laundering, tax evasion or other fraud schemes. Unlike normal ML tasks, the key is to over-fit on existing data in order to strongly reject anomalous behaviours.

Methods and technologies

To achieve our goal we employ several types of methods. We give here an overview of the most important.

The technique that is most specific to this project is *dictionary learning* (DL) [8] for sparse representations. This is a versatile tool that can be used for modeling diverse type of signals, including those related to graphs. DL can also be used for classification in different contexts (binary or multi-class, 1D or 2D signals). There are now several DL algorithms with well tested performance.

Besides some usual methods for data pre-processing, an important operation is *community detection*. Due to the large size of graphs of financial transactions, the dataset is usually split to help computing. A simple strategy is to use multiple short time-frames (a week, a month, a trimester), optionally with overlapping such that the risk of losing fraudulent behaviour that spans wider intervals is minimized; some data are, therefore, deliberately ignored. More appropriate is to take the entire large transaction graph and split it in a set of smaller sub-graphs through community detection, which preserves the full local structure and groups nodes that share significant similarities.

Finding static patterns is achieved through community detection and *classification*. We train a dictionary on data generated from the known patterns and use sparse modelling to perform classification on each of the previously obtained communities. Data generation and learning have to take into account and instill extra properties to the resulting dictionary, such as invariance to scaling (e.g. when the ring from Figure 1 has 10 nodes instead of 5) and noise tolerance (e.g. when a fraudulent entity also has legitimate activities such as paying bills or rent). To improve detection, we also use multiple methods from different fields such as deep learning (to train a classifier with similar properties as above) or support vector machines (SVMs) and tie their results together to produce the anomaly score.

Finding patterns without prior knowledge is based on the assumption that the given dataset is mostly made up of legitimate transactions and only a few (e.g. less than 1% or even 1‰) represent frauds. In this manner, we train the dictionary on the entire dataset and use sparse-modelling to identify the transactions that required denser representations or that lead to larger approximation errors than most. Again, to improve detection we use multiple unsupervised methods such as one-class SVMs, principal component analysis, and their deep learning variants, and tie their results together through voting methods.

All our programs will be written in Python, which provides mature implementations for most, if not all, of the standard methods mentioned above, which are made available through free, open-source, industry-proven toolboxes such as scikit-learn, numpy, networkx, and pytorch. However, at least in some cases, especially those related to DL, we aim to implement the best algorithms in an efficient manner.

State of the art

We further briefly survey the main directions for detecting anomalies in various structured networks under different assumptions.

Detection of anomalous static patterns. The large variety of anomalies types arising in networked environments have led to different research directions depending of applications at hand. In directed trading networks, blackhole and volcano patterns represents groups of nodes only with inlinks from the rest of nodes or only outlinks towards the rest nodes, respectively. These kinds of patterns, which often have fraudulent nature, are isolated in [20] through pruning (divide-et-impera) schemes based on structural features of blackholes and volcanoes. Subgraph patterns with structural particularities such as near-star, near-clique, heavy vicinity, dominant heavy links are labeled as anomalies in [1] through algorithms that separate egonets (subgraphs formed by the neighbors of a single node) in a weighted graph.

A supervised learning AML system is developed in [33], based on features involving information on network dynamics and party demographics, which extracts particular near-k-step neighborhoods and other communities from a transaction network. Subsequently, it applies a simple supervised learning method to detect anomalous structures. Similar coupling between network analysis tools and classification techniques is employed by [11]. A modularized anomaly detection hierarchical framework has been developed in [11] to detect static anomalous connected subgraphs, as enumerated in Fig. 1, with high average weights. For this purpose, particular community detection strategies are tailored based on 140 features (including Laplacian spectral information) and network comparison tests (such as NetEMD). Then, a classification via random forests or simple sum of individual (feature-based) scores is performed to highlight the anomalous subgraphs.

In [13], financial transactions are encoded in a weighted graph and, subsequently, low-rank decompositions with sparse terms of similarity and features matrices are computed to isolate pathological patterns following ring, merging and outlier point scenarios in financial fraud case studies. Another direction of algebraic arguments of matrix theory lead to spectral localisation. Spectral localisation is the phenomenon in which a large amount of the mass of an eigenvector is placed on a small number of its entries [11]. Related to Laplacian matrices, patterns of dominant components of their eigenvectors corresponds to patterns of nodes in the network with special properties, and thus constitute good candidates for the anomaly detection task (see [5,28]). In the series of papers [21,22,23,24] a set of schemes are developed in order to uncover anomalies using spectral features of the modularity matrix. Furthermore, the authors of [23] extend these methods to use the ℓ_1 norm for eigenvectors of sparse PCA, which performs well at the cost of being more computationally intensive. A different spectral approach, proposed in [35], is based on matrix factorization in bipartite networks and leverages the intuition that the nodes and edges which are badly represented by the factorization should be considered as anomalies.

Detection of anomalous patterns without prior knowledge. Several social network statistical metrics and clustering techniques are used in [4] to detect fraud in a factoring company. In [19] a hybrid anomaly detection approach is considered that employs clustering (Euclidean Adaptive Resonance Theory) to establish customer's normal behaviors and then uses statistical anomaly index to determine deviation of a particular transaction from the corresponding group behavior.

Evolutionary networks are considered in [3], where a community detection strategy is used to highlight anomalies based on the temporal quantitative evolution of network communities. In [12], the spectral embedding of individuals across different data sources is compared, declaring an anomaly if the embeddings deviate substantially.

The literature on pure deep learning for anomaly detection in networks is relatively reduced comparing with the previous approaches. However, we provide a few references with seemingly noticeable contributions. In [36], the authors learn the latent attributed network representation by using a number of network walks. The representation is obtained through maintaining the pairwise vertex-distance in the local walks and by hybridizing it with the hidden layer of deep auto-encoder, such that the resultant embedding is guaranteed to faithfully reconstruct the original network. Then, a dynamic clustering model is used to anomalous vertices or edges based on the learned vertex or edge representations. Moreover, leveraging a reservoir sampling strategy, any dynamic network change induces only easy updates on the learned representations. The anomaly detection problem in interactive attributed networks is approached, in [6], by allowing the system to proactively communicate with the human expert in making a limited number of queries about ground truth anomalies. The problem is formulated in the multi-armed bandit framework and after applying some basic clustering methods, it aims to maximize the true anomalous nodes presented to the human expert in the given number of queries. The results show certain improvements comparing with similar approaches.

Novelty and relevance

All banks have anti-fraud departments. Although some automated tools are used, they are usually rule-based, still have limited impact and human experts are often able to discover frauds only after significant delays and can discover only some types of fraud. A software package with modern methods can have significant impact: the savings resulting from the early discovery of frauds and from releasing the human experts from part of their work are much higher than the investment in acquiring such a package.

The novelty of our approach relies in the choice of the methods and in the overall desired functionality.

Sparse representations are linear combinations of a few vectors (named atoms) from an overcomplete basis (called dictionary). Modeling community sub-graphs with sparse representations, via a dictionary that is trained from examples through DL methods, was attempted only very recently and partially, by some of our team members (see section B2.2). Essentially, neither finding static patterns, nor unsupervised learning for AD in graphs have been solved using DL. Sparse representations can capture the significant features of a dataset. Due to overcompleteness, a trained dictionary implicitly covers some important properties like shift or rotation invariance, that typically make graph modeling difficult. It is also good at modeling the varying size of typical graph structures. While normal patterns are represented with only small errors, anomalies of all kinds are usually poorly modeled and, therefore, can be identified. (Note that nonlinear sparse representations are available through the kernel trick and special algorithms [26]). All the above support our claim that the main tools we propose are relevant and there is an important degree of novelty in their application to graphs and financial transactions.

The Graphomaly package that we aim to implement in Python will have an open architecture using a single internal data format for graphs (and conversion from/to other formats). The functions that implement the same operation (with several methods) will be completely interchangeable, in order to allow multiple choices for the user and full modularity. We aim to create a skeleton structure easy to use for further development. We are not aware of another product with the functionality and structure that we propose.

Main objectives of the Graphomaly project

O1 Python Toolbox

The toolbox will receive a large transactions dataset as input and provide as output the anomalous transactions. The main operations performed by the toolbox are as follows.

O1.1) Data pre-processing. The input data—transactions and their attributes—are first processed in order to obtain their graph and vectorized representations. To this end we will use tools such as: Python Data Analysis Library for graph data organization, preprocessing from scikit-learn, PCA, numpy.

Transaction information contains mixed categorical and numerical data. Label encoding is employed to transform the categorical variables. However, few of these categories are prohibitively large, with cardinality of more than 100 elements. Given the dataset dimensions, a dimensionality reduction step, such as PCA, is required in the case of these variables.

Feature selection and feature engineering are also necessary, as transaction attributes often contain information that is not readily operable by a machine learning algorithm. Date and time are such examples. Account opening and closing dates, for instance, are in themselves not meaningful, however the difference of these dates can indicate illegitimate behaviour.

O1.2) Community detection, which means splitting the large graph representation of the transactions into smaller sub-graphs that are numerically tractable. To this end we will use tools such as: the Louvain Community Detection Library and the Networkx library for creating and manipulating graphs and networks. Basic graph theory methods for finding cliques, egonets [1] or other fixed structures can be useful, as well as clustering methods.

O1.3) Anomaly detection for scenario (i). We will use dictionary learning methods to accommodate the graph structure of the transactions. We mention here: Laplacian structured dictionary learning where each dictionary atom is a vectorized Laplacian of common sub-graph transaction patterns, separable Laplacian classification where the transactions are matrices representing weighted Laplacians, and graph orthonormal blocks classification where each sub-dictionary represents the Laplacian of the static pattern.

O1.4) Anomaly detection for scenario (ii). We will adapt and use existing AD methods such as: One-class Support Vector Machine (OC-SVM), Support Vector Data Description (SVDD), Isolation Forrests (IF), Robust Principal Component Analysis (R-PCA), Gaussian Mixture Models (GMM) and their Deep Learning variants. We will also formulate an unsupervised dictionary learning method (used as a basis for the online version described in **O2**).

O1.5) Method combinations. The toolbox will be implemented in a modular structure, allowing the choice of the various preprocessing strategies and anomaly detection methods and thus providing multiple results. We will explore some of the combinations and attempt to find the most successful ones. Voting algorithms will also be employed for extracting the best results from several methods. Scoring methods will also be investigated. However, the main purpose is to offer flexibility to the potential user.

O2 Development of new algorithms

Most of the algorithms that we plan to use are already available. However, some adaptations are necessary for special situations that may occur in relation with anomaly detection in graphs. We want to develop and implement at least an algorithm for each of the two categories below.

O2.1) Online algorithms. Transaction data are typically analyzed in large batches, corresponding to a period of several months or even years, to detect abnormal behavior. Online algorithms give a quicker reaction, as they work continuously, adapting the learning results when small amounts of data (a day or less) are added to those already used. Fraudulent activities can be discovered in their incipient stages. We aim to provide fast unsupervised dictionary learning based methods for AD.

O2.2) Distributed algorithms. As the graphs modeling financial transactions can be huge, it is interesting to reformulate the AD algorithms in a distributed manner by splitting the graphs into (partially superposed) pieces and analyzing them in parallel, across a network of processing nodes.

O3 Testing

O3.1) Tests on simulated data. As a first validation step, we will test our algorithms on simulated data, obtained by adding known anomalous structures on random graphs or on graphs of normal transactions. Not only the detection efficiency can be tested, but also the behavior (especially execution time) of our methods on large graphs.

O3.2) Tests on public datasets. Further validation will be made on datasets that are publicly available. They cover only particular situations, are sometimes insufficiently explained and are not very large, but they are relevant for comparison with other methods.

O3.3) Tests on data provided by our external partner. BRD (see support letter in Figure 2) will provide anonymized data that include known fraudulent transactions. They will also analyze some of our results by investigating if the discovered anomalies are indeed suspicious. So, we expect partial validation from banking experts.

TRL discussion

The general problem that we aim to solve—anomaly detection on graphs of financial transactions—has been tackled with various methods but it is by no means satisfactorily solved. The techniques we propose, based on dictionary learning, are now relatively mature, but have been applied only scarcely to AD. Leaving aside the auxiliary tools and methods that are already implemented, we discuss here the TRL regarding the main aspects of our proposal. We claim that our starting position contains elements of both TRL 2 and 3 and that our results will belong to TRL 4.

TRL 2 (technology concept formulated). DL as a tool for detecting anomalies in graphs has been already investigated by part of our team members, as detailed in B2.2 (preliminary results). The optimization problem has already been formulated in both scenarios (known patterns and unsupervised AD) and brought to manageable size by community detection.

TRL 3 (experimental proof of concept). We have already MATLAB implementations of DL algorithms (see again B2.2) made by ourselves and well tested for various problems, mostly academic. Some DL algorithms were applied to graph AD but only on a very small number of datasets of rather limited size and without special concern for an efficient implementation. The first results are promising, hence the approach is sound, but there is not yet proof that a successful AD technology has been obtained.

TRL 4 (technology validated in lab). We aim to obtain a software package for graph AD that has several algorithmic options for each module and thus is able to provide results that cover a full spectrum of anomalies. The programs will be efficiently implemented and able to work with large datasets. They will be fully tested on synthetic data. Selected sets of real data will also be used for tests, including some provided by our external partner (BRD) and so significant validation will be demonstrated.

What is missing for a superior TRL? Although some elements of higher TRL are present in our validation, the number of real datasets that we will use and their cover of the (ideally) complete transactions graph will still be small compared to the full data seen by a bank or a consortium of banks. Online or distributed algorithms will be tested only in lab conditions, not in a quasi-real environment (with special hardware, in the latter case). Recognized data formats will be limited and no interface is planned for the non-expert user.

B. 2.2 Presentation of the concept of technology / productor or existing model which constitutes the starting point of the project

Preliminary results

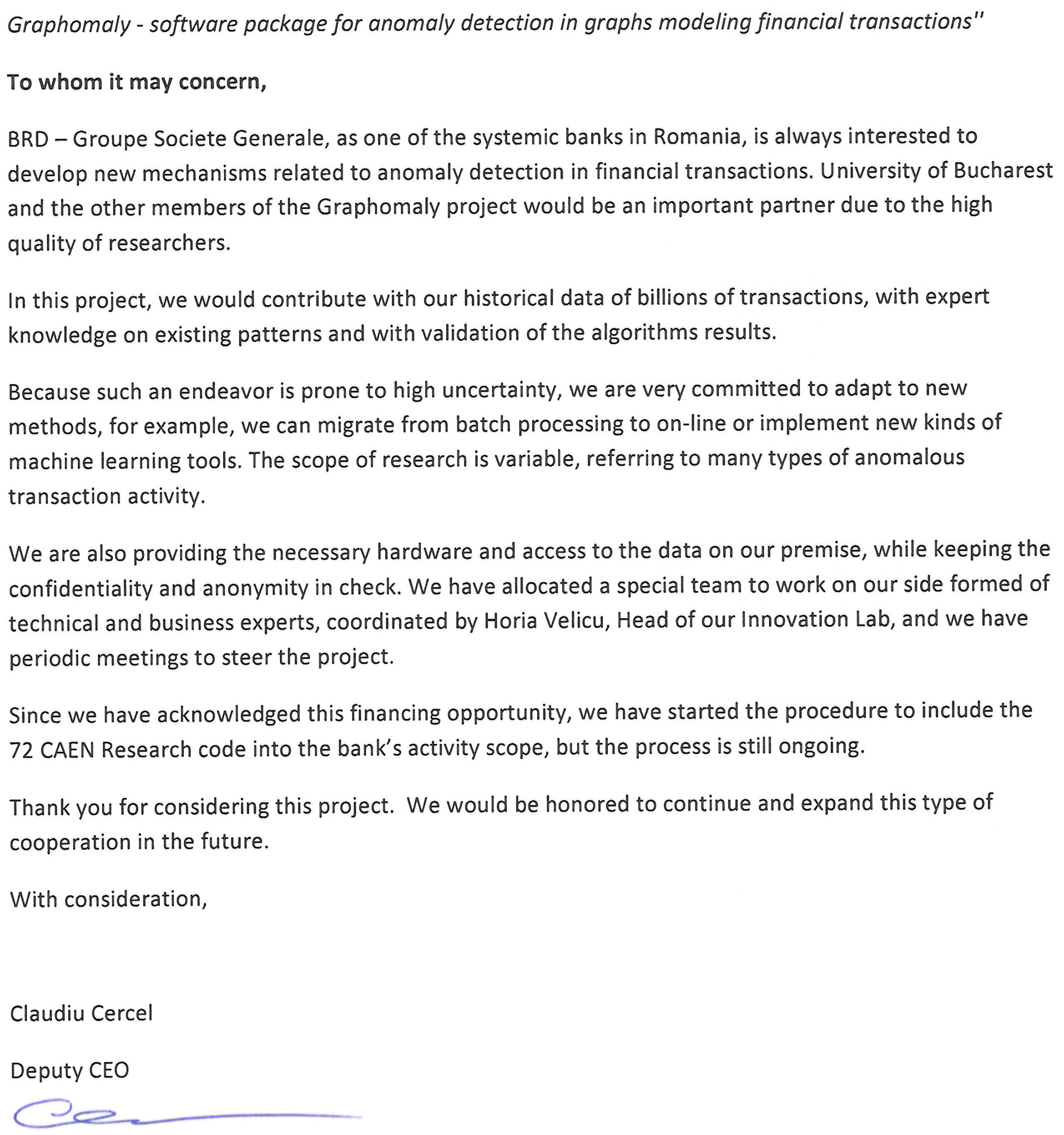
MATLAB programs for the book ”Dictionary Learning Algorithms and Applications” (Springer 2018, authors B. Dumitrescu and P. Irofti) [8] are already available at https://github.com/pirofti/dl-box. They include several DL algorithms, some developed by the authors, like regularized K-SVD [9]. DL-based classification methods are also present, in particular Discriminative K-SVD [37] and Label Consistent K-SVD [18]. Nonlinear DL using kernel methods is also mastered. Besides the book authors, several team members have already good DL expertise.

BRD Groupe Societe Generale started a Fellowships program in 2019 through the Data Science Research Center at UB-CS, founded with 100,000EUR. The program was equally split in two research directions, one of them focusing on Anti Money Laundering (AML) for banking transactions where Paul Irofti, Andrei Pătrașcu, and Andra Băltoiu were among the members that won an 8 month Scholarship (January to August). Within this program, the three formed a team and established the state of the art in anomaly detection for financial fraud, gained experienced with real data provided by BRD, adapted existing algorithms and proposed new ones, organized an anomaly-detection reading-group at UB-CS (https://sla.cs.unibuc.ro/index.php/events/category/seminar/seminar-securitate/) together with Horia Velicu (Head of Innovation Lab at BRD). This scholarship led to three papers that will be published soon and a set of software programs (Python scripts) implementing and adapting new and existing methods to financial data. The software was successfully used by BRD to verify that it identifies some known transaction frauds and to identify frauds in new transaction data that was then confirmed through manual inspection by BRD AML experts. We mention here that preliminary dictionary learning based results were able to identify two new static patterns that were used for transaction fraud and were unknown before.

Project team

The three partners cooperating at the implementation of this project are:

- University Politehnica of Bucharest, Department of Automatic Control and Systems Engineering (UPB-ACSE), coordinator (CO)
- University of Bucharest, Department of Computer Science (UB-CS), partner (P1)
- Tremend Software Consulting SRL (TSC), partner (P2). TSC currently delivers IT services to 4 major Romanian banks (Raiffeisen, ING, First Bank, BRD Groupe Societe Generale). In 2019, TSC signed a



Graphomaly - software package for anomaly detection in graphs modeling financial transactions"

To whom it may concern,

BRD – Groupe Societe Generale, as one of the systemic banks in Romania, is always interested to develop new mechanisms related to anomaly detection in financial transactions. University of Bucharest and the other members of the Graphomaly project would be an important partner due to the high quality of researchers.

In this project, we would contribute with our historical data of billions of transactions, with expert knowledge on existing patterns and with validation of the algorithms results.

Because such an endeavor is prone to high uncertainty, we are very committed to adapt to new methods, for example, we can migrate from batch processing to on-line or implement new kinds of machine learning tools. The scope of research is variable, referring to many types of anomalous transaction activity.

We are also providing the necessary hardware and access to the data on our premise, while keeping the confidentiality and anonymity in check. We have allocated a special team to work on our side formed of technical and business experts, coordinated by Horia Velicu, Head of our Innovation Lab, and we have periodic meetings to steer the project.

Since we have acknowledged this financing opportunity, we have started the procedure to include the 72 CAEN Research code into the bank's activity scope, but the process is still ongoing.

Thank you for considering this project. We would be honored to continue and expand this type of cooperation in the future.

With consideration,

Claudiu Cercel

Deputy CEO

Figure 2: BRD Group Societe Generale support letter

global partnership agreement with Mastercard for delivering IT services and particularly services related to ePayment solutions. Tremend also serves 2 major international financial institutions. In total, 25% of the revenues are generated in the financial-banking market.

In addition, BRD will act as external partner along the lines of the support letter from Figure 2. The only reason for which BRD is not part of this project is the lack of a research CAEN code (to be obtained in the near future).

The project team comprises the following researchers.

CO: **Bogdan Dumitrescu** (professor), Florin Stoican (professor), Denis Ilie Ablachim (PhD student since 2019)

P1: **Paul Irofti** (lecturer), Marius Popescu (associate professor), Andrei Pătraşcu (lecturer), Andra Băltoiu (PhD student aiming to defend thesis in 2020).

P2: **Bogdan Savu**, Ştefania Budulan, Florin Ilie

Bogdan Dumitrescu (born 1962) works at UPB-ACSE since 1990 and is a professor since 2003. ORCID: 0000-0003-4555-1714, ResearcherID: B-5839-2011, Scopus: 6603839944, Google Scholar: QEf1T4gAAAAJ, UEF-ID: U-1700-039W-5496. He has published more than 50 journal articles and 90 conference papers, including many articles in prestigious (Q1) journals, like IEEE Trans. on Signal Processing (9 articles), IEEE Signal Processing Letters (10), Signal Processing (9). He is the author of the books "Positive trigonometric polynomials and signal processing applications" (Springer 2007, 2nd ed. 2017) and "Dictionary Learning Algorithms and Applications" (Springer 2018, with Paul Irofti). He holds three international patents. Full publication list at http://www.schur.pub.ro/BD_PublicationList.html. Hirsch index: 14 WoS, 14 Scopus, 18 Google Scholar. Citations: >500 WoS, >700 Scopus, >1300 GS.

His current research interests are in numerical methods and optimization for signal processing, with focus on sparse representations and dictionary learning. Research grants: several grants from the Romanian National Science Council, including two IDEI (currently PCE) grants (Positivity in the analysis and synthesis of multidimensional systems 2007-2010, Sparse representations and signal processing applications 2011-2016). Industry contracts: Nokia Research Center and Microsoft Finland. He was FiDiPro (Finnish Distinguished Professor, a program through which top international researchers work half-time in Finnish universities) fellow at Tampere University of Technology in 2010-2013; he had many research stages there between 1998 and 2016.

B. Dumitrescu has well recognized contributions to dictionary learning [32],[9],[7] and sparse representations [10,27], which are core tools in this project.

Paul Irofti (born 1984) works at UB-CS since 2017. UEF-ID: U-1700-038X-5086, ORCID: 0000-0002-7541-4334, Google Scholar: yJZpIfgAAAAJ. He has published papers in prestigious journals, including one in IEEE Signal Processing Letters (Q1) [9], and at flagship conferences such as ICASSP [16] and EUSIPCO [14,15]. He is the author of book "Dictionary Learning Algorithms and Applications" (Springer 2018, with Bogdan Dumitrescu). He has been a member in 4 research projects, one in collaboration with industry, and in 2019 he was the winner of the BRD Groupe Societe Generale Fellowship for researching anomaly detection techniques for banking transactions. Full publication list at https://cs.unibuc.ro/~pirofti/publications.html.

Besides his academic career, he has 15 years of experience working in the industry, out of which 3 were spent as AntiMalware Engines Engineer at BitDefender and 9 as a security-consultant at various companies and start-ups. He has also been a core kernel developer of the security-orientated OpenBSD operating system since 2008. Between 2011–2013 he lead a start-up team that developed from the ground-up a new security technology that he pitched and sold to a security agency of the German government. In 2015 he consulted for a newly created Sillicon Valley start-up called BitFusion where he designed and prototyped their main product: a new language that virtually remote attached devices across the network (such as GPUs and CPUs). Bitfusion was aquired in 2019 by VMWare and this product will be available as part of their next release.

P. Irofti has algorithmic and hands-on experience with real-data for anomaly detection in financial transactions, is well connected in the security industry with a proven record of growing ideas into end-products, and has many well recognized results in dictionary learning (the core tool in this project).

Bogdan Savu (born 1977) works at TSC since 2018 as Director of Engineering. He has graduated from the University of Bucharest with a BSc in Computer Science. He has also got his CNAM Paris Master of Business Administration (MBA) diploma and Master in Management from the Bucharest Academy of Economic Studies. He has 20 years of experience in IT, Business Development and Management. His expertise include software development, architecture and system design and also management (strategy, leadership, risk management, financial and investment operations).

Florin Stoican (born 1984) works at UPB-ACSE since 2013 and is a professor since 2019. UEF-ID: U-1700-031Y-8324, ORCID: 0000-0002-4550-9113, Google Scholar: vcbJEuQAAAAJ. He obtained his PhD from SUPÉLEC, France (2011), had a postdoctoral fellowship at NTNU, Norway (2011-2012). He is the (co-author) of two books ("Mixed-Integer Representations in Control Design. Mathematical Foundations and Applications" Springer 2016, "Set-theoretic fault detection in multi-sensor control" Wiley 2013), the patent WO2018215910-A1, over 65 publications (47 of them ISI-indexed). He will contribute in tasks related to his field of expertise: model-based fault detection and isolation where he has previous results [17,34] connecting it to dictionary learning (the core tool in this project). Full publication list at http://florinstoican.com/os/academic/publications.

Marius Popescu (born 1967) works at UB-CS since 2004, where he is an associate professor. UEF-ID: U-1700-039W-6249, Google Scholar: UPWSjkAAAAAJ. Full publication list at http://fmi.unibuc.ro/ro/popescu_marius/. His domains of interest are: artificial intelligence, machine learning, computational linguistics, information retrieval, authorship identification, computer vision. Over 50 articles published at international peer-reviewed conferences and journals.

Some of his achievements in these fields include: coauthor of the method ENCOPLOT for plagiarism detection which won the first international competition in automatic plagiarism detection in 2009, followed by ranking 4th at PAN@CLEF 2010 and 2nd at PAN@CLEF 2011; a method for authorship analysis which obtained the best results in the author identification task at PAN@CLEF 2012; coauthor of learning systems that ranked on 2nd place in the Arabic Dialect Identification Shared Task of the VarDial Workshop of COLING 2016 and 1st place in all three tracks (essay, speech, fusion) of the Native Language Identification Shared Task of BEA-12 Workshop of EMNLP 2017.

M. Popescu, as Head of the Data Science Research Center at UB-CS and Machine Learning expert, will act as an expert consultant to the project members in regards to data processing and Deep Learning.

Andrei Pătraşcu (born 1987) works at UB-CS since 2017. UEF-ID: U-1700-032E-3292, ORCID: 0000-0002-9293-9386, Google Scholar: tndIB4oAAAAJ. He has published 9 journal articles and 10 conference papers. Most articles (7 out of 9) are published in top (Q1) journals such as Journal of Machine Learning Research [30], SIAM Journal on Optimization [25] and IEEE Transactions on Automatic Control [31]. He has been awarded with Best Paper Award from Journal of Global Optimization for paper [29] published in 2015. He won BRD Groupe Societe Generale Fellowship to develop anomaly detection methods for financial transactions. He has been member in 7 research grants since 2011. His current research interests are in the development of numerical optimization algorithms, endowed with computational complexity estimates, for nonlinear optimization with focus on machine learning and anomaly detection applications. Full publication list at `https://sites.google.com/site/andreipatrascuro/publications`.

Ștefania Budulan (born 1991) works at TSC since 2017, UEF-ID: U-1900-062K-9532, Google Scholar: gcUHzRUAAAAJ. With over 4 years of experience developing AI/ML solutions for various domains, including financial, audit and software security through biometric authentication, she has a strong interest towards creating viable state-of-the-art solutions for industrial use, thus reducing the gap between academic research and the industry reach and potential. At TSC, she worked on a complex unsupervised Anomaly Detector adapted for time-series, intended for detecting novelty/unusual events based on system metrics (e.g. CPU usage, memory consumption, etc.). The knowledge-gain, along with specific methods and algorithms can be successfully transferred onto this project.

She is currently enrolled in a PhD degree in Artificial Intelligence, starting in 2017, with a focus on natural language processing. During her academic route, she published a few conference papers, usually in close connection with a private company, including [2] presented at a rank A conference, researching the possibility of identifying fraudulent users based on their smartphones screen gestures.

Florin Ilie (born 1978) works at TSC since 2017. He has more than 20 years of experience as software specialist and several years of technical project management and enterprise programming. Florin graduated the Faculty of Mathematics at the University of Bucharest.

Andra Băltoiu (born 1985) is a research assistant at the Research Institute of University of Bucharest, UEF-ID: U-1800-055D-9454. Starting from 2016, she is a PhD student under the supervision of prof. Bogdan Dumitrescu.

She has 4 years of experience in data analysis and signal processing, gained as research scientist at the Institute of Space Science and the National Institute for Sport Research, between 2013 and 2017. Her work routinely involved modelling or handling anomalies in physiological signals. Currently, her research is focused on dictionary learning methods, which she successfully applied to the problem of anomaly detection in the context of malware identification. The article appeared in 2019 at EUSIPCO [14]. In 2019 she won the BRD Groupe Societe Generale Fellowship, during which she worked on anomaly detection methods for financial transactions. Her firsthand experience with financial data gained throughout the fellowship is relevant for devising the requirements for the software package.

Denis Ilie Ablachim (born 1994) started PhD in 2019 under the supervision of BD, after graduating master studies at UPB-ACSE in the same year. His research topic is dictionary learning with focus on kernel and nonlinear techniques and classification as main application.

B. 2.3 Method of project implementation

Activities and deliverables

Most of the objectives imply industrial research activities (e.g. for the efficient implementation of new algorithms) and also experimental development activities (e.g. for data manipulation, program testing). Objective **O2** requires also some fundamental research activities for the development of new algorithms, within the 10% threshold imposed by the financing authority. The list below shows the activities, deliverables and responsible persons; a person is denoted by team (CO,P1,P2) and initials; only the main actor from a team is mentioned in connection with an activity (or the whole team).

The main roles of the three partners are as follows. UPB-ACSE and UB-CS will be the main drive in the application of the DL methods and the development of the new algorithms. TSC will contribute to software architecture and will implement alternative methods based on their AI expertise. All partners will participate in testing on various datasets.

A1.1.1 Design of internal data formats. Implementation of conversion routines from/to other formats met in datasets. Deliverable: internal document. (CO-BD, P1-PI, P2-BS)

A1.1.2 Design of the package structure. Define main operations types and input/output arguments. Define part of the global parameters that are user tunable. Deliverable: internal document. (P1-PI, CO-BD, P2-BS)

A1.1.3 Choice of pre-processing tools and tests on selected data. Deliverable: code and documentation. (CO-FS,P1-AB, P2-FI)

A1.2.1 Selection and integration into the selected input/output format of Louvain, Girvan-Newman, clustering and other graph theory community detection algorithms. Testing. Deliverable: code. (CO-FS, CO-DI, P1-AB, P2-FI)

A1.3.1 Python implementation of relevant DL algorithms for which we already have MATLAB programs. In particular, sparse representation algorithms like OMP have to be very efficiently implemented, including their bulk versions that are instrumental in DL context. Deliverable: code, documentation. (P1-AB, CO-DI)

A1.3.2 Design and implementation of AD through DL using vectorized Laplacian graph representations. Several DL algorithms will be employed, among which AK-SVD and its regularized version and SGK. Deliverable: code, documentation. (P1-PI, CO-BD)

A1.3.3 Same as above, using separable Laplacian representations. Deliverable: code, documentation. (P1-AB)

A1.3.4 Write extended report, used as a basis for a scientific publication. Deliverable: submitted manuscript. (CO-BD, P1-PI)

A1.4.1 Adaptation and implementation of existing generic AD methods. One-class Support Vector Machine, Support Vector Data Description, Isolation Forrests, Robust Principal Component Analysis, Gaussian Mixture Models and their Deep Learning variants. Deliverable: code, documentation. (P2-SB, P1-MP, P1-AP, CO-DI)

A1.4.2 Design and implementation of unsupervised DL method for graph AD. Deliverable: code, documentation. (CO-BD, P1-PI)

A1.4.3 Write extended report and a scientific publication. Deliverable: submitted manuscript. (P2-SB, P1-MP, CO-BD)

A1.5.1 Implementation of scoring and voting methods. Deliverable: code, documentation. (CO-DI, P2-FI)

A1.5.2 Experimental optimization of combinations of methods with the purpose of finding the ”best” one. Deliverable: report. (CO-FS, P1-AP)

A1.5.3 Write extended report and a scientific publication. Deliverable: submitted manuscript. (CO-BD,P1-PI,P2-BS)

A2.1.1 Design of online graph AD based on unsupervised DL. Deliverable: MATLAB code and preliminary results. (P1-AP, CO-FS)

A2.1.2 Python implementation. Testing by comparison with batch version designed at A1.4.1. Deliverable: code and report. (CO-DI, P2-FI)

A2.2.1 Design of distributed graph AD based on unsupervised DL. Deliverable: sequential MATLAB code. (CO-BD,P1-PI)

A2.2.2 Python implementation using multiprocessing. Tests on synthetic data of moderate size. Deliverable: Python code and timing results on multicore processor. (P1-PI, P1-AB)

A2.2.3 Write scientific publication on A2 activities. Deliverable: submitted manuscript. (P1-PI, CO-BD)

A3.1.1 Design several types of anomalies and write programs that add a predefined mixture of them to a given graph (synthetic or real). Deliverable: Python code. (P2)

A3.1.2 Test execution time of designed AD algorithms on graphs of different sizes and draw performance/speed diagrams. Deliverable: report. (P2)

A3.2.1 Gather public datasets and convert to our internal format (in conjunction with A1.1.1) Deliverable: datasets, documentation. (CO,P2)

A3.2.2 Test performance for all implemented methods and gather all relevant results in synthetic form. Comparison with the state of the art. Deliverable: report (to be used also for publication). (P2,CO)

A3.3.1 Receive from BRD data with labeled fraudulent activities. Run Graphomaly training methods on these data and report rate of success in anomaly detection. Deliverable: report. (P1)

A3.3.2 Receive from BRD unlabeled data. Run Graphomaly methods on these data and report detected anomalies. BRD human experts further analyze the reported transactions and decide if the anomaly is indeed a fraudulent activity or a rather rare but innocent transaction. Deliverable: report. (P1)

Some examples of public databases for the A3.2 activities are
https://www.kaggle.com/c/ieee-fraud-detection/overview/evaluation,
https://www.kaggle.com/dileep070/anomaly-detection,
https://offshoreleaks.icij.org.

The first two contain anomalies; the third contains normal transactions where anomalies can be injected like in [13].

The Gantt chart corresponding to the above activities is shown in Figure 3, with different colors for the predominant type of an activity.

Project management and decision making

The main decisions regarding work tasks and budget will be made by the board composed of the leaders of the three partners. Decisions will be made based on the majority rule, but consensus will be sought in all important matters. The board will meet monthly (in person or via skype). Other team members will be invited (without voting rights) for presenting their opinion and also for the discussion of technical decisions regarding the current tasks. All researchers will be encouraged to state their viewpoint and to openly discuss any technical or nontechnical issues that might appear with their team leader or, in case of disagreement, with the project leader.

The project leader and partner leaders will make the executive decisions, will follow closely the progress of the work tasks and will give brief monthly reports of the current status.

All attempts will be made to solve possible conflicts in a friendly manner. Informal communication will be encouraged as well as participation at the AD reading group at UB-CS. The whole team (including a BRD representative) will be gathered for a kick-off meeting, then every about six months for technical meetings and finally for a closing event.

Dissemination of results and IPR rights

Since this research is financed mostly by the Romanian state, the software package will be made public (but not the datasets). The project will have a web site on which relevant information will be updated periodically.

We plan to publish at least four articles using the results obtained in this project, in a Q1/Q2 journal and three conferences (WoS indexed), at least one being at a quality level similar with ICASSP and EUSIPCO (flagship signal processing conferences). Authorship will be confined to the team members that have actually contributed to the work on which an article is based.

We will also try to approach representatives of the main Romanian banks, inform them about our project and its website and seek possible cooperation.

In case there will be enough technical novelty for a patent proposal to be submitted, the three partners will have equal institutional rights.

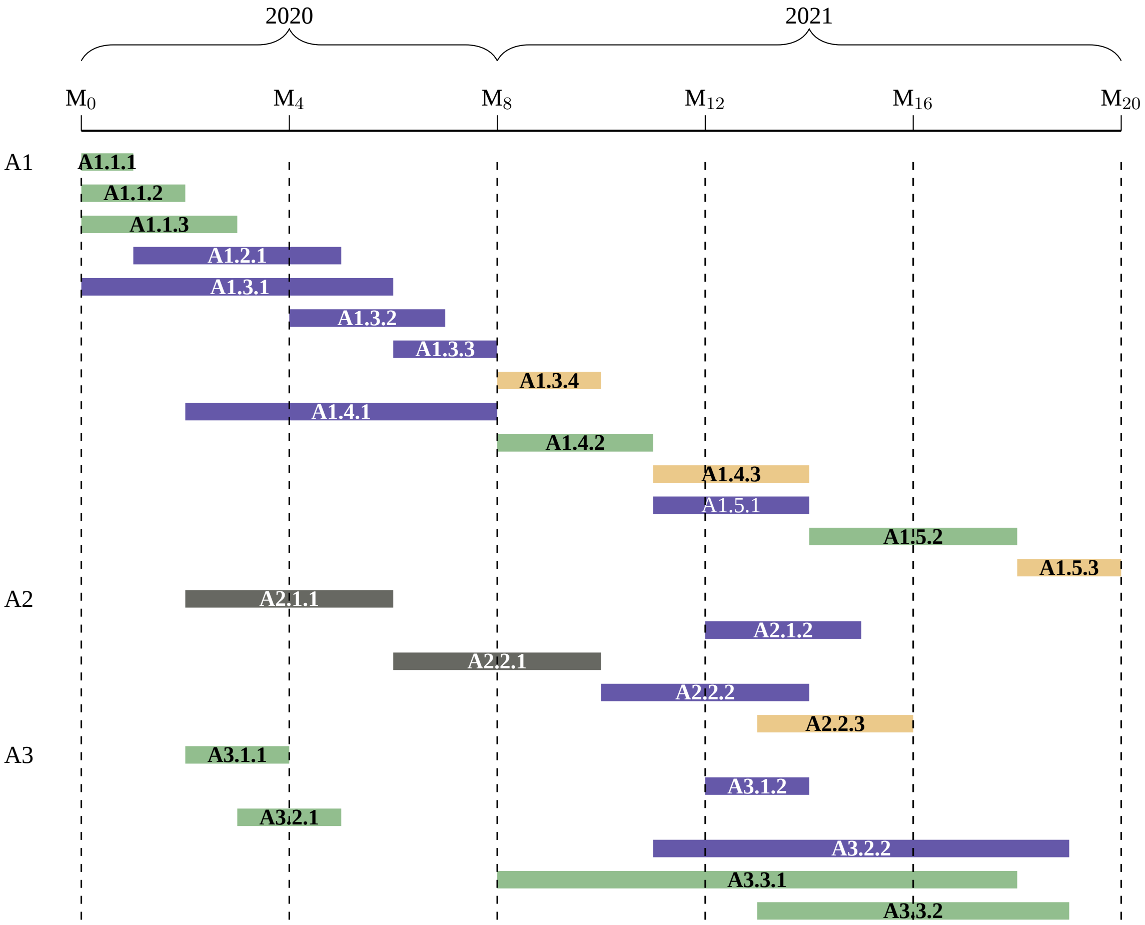


Figure 3: Gantt chart: **fundamental research**, **industrial research**, **development**, **writing**

.

Research infrastructure

Regarding infrastructure, the Graphomaly project needs only computing power for achieving its goals. General purpose computers or laptops are sufficient for the development of algorithms and for some of the activities like data conversion or pre-processing. Running learning algorithms for large graphs may need hours or even days in some cases, hence dedicated computers are necessary. Testing AD after learning is not so time consuming.

UPB-ACSE members work in room ED206, where they have a graphic station with a 6-core processor and GPU (powerful but rather old – bought in 2012) and other equipment like printer, scanner, copying machine. Department infrastructure can be seen at `https://erris.gov.ro/ACSE---UPB`; see also `http://acse.pub.ro`.

UB-CS members have permanent access to a DELL PowerEdge R530 with 2 Xeon CPUs with 20-cores each, 4 GPUs and 256GB system memory. Upon request, the members can also gain access to resources from the Research Institute of the University of Bucharest (`https://erris.gov.ro/ICUB`), the MOCALC research center for computational models, algorithms and cryptography, and the Human Language Technologies Research Center (`https://erris.gov.ro/HLTCenter`).

Tremend currently has the necessary infrastructure to carry out the project, using its own software development and engineering laboratory, located in the TN Office 2 office building (4th floor) in Bucharest. The laboratory is equipped with LAN and WiFi network, state-of-the-art hardware, DELL servers with high-performance processors and redundant backup. All workstations and servers are equipped with complete software packages required for the development.

Project budget

The total budget is 669 825 lei, out of which 600000 come from the public budget and 69825 from own contribution of TSC.

Allocated budget / costs (Lei)						
		Personnel costs	**Logistics**	**Travel**	**Indirect costs**	**Total**
Coordinator (CO)	**Public budget**	144 000	30 000	20 000	36 000	230 000
Partner 1	**Public budget**	144 000	25 000	19 140	40 785	228 925
Partner 2	**Public budget**	112 860	0	0	28 215	141 075
	Own contribution	55 860	0	0	13 965	69 825
Total		456 720	55 000	39140	118 965	669 825

Logistics. UPB-ACSE and UB-CS will increase their computing power by buying two computing systems with Xeon high performance processor (one for each team). They will also buy 2-3 laptops each for the team members, starting with the junior ones (depending on the price of the more powerful computers).

The remaining money will go to consumables, uninterruptible power sources (UPS), IEEE member fees (for getting lower registration fees at IEEE conferences) and a small reserve for unforeseen situations.

Travel. We plan four participations at conferences (two for UPB-ACSE, two for UB-CS) that are relevant in the AD, machine learning and signal processing fields.

Indirect costs. UPB-ACSE and TSC: 25% of personnel costs. UB-CS: 25% of direct costs minus logistic costs. In both cases, the threshold imposed by the financing entity is respected.

Justification of salary expenses. See table below. At TSC, Bogdan Savu and Ştefania Budulan will do industrial research (75% from public budget) and Florin Ilie development (50%).

Person	Position	Workload	Man-months	Salary (lei/month)	Total (public)	Total (own)
CO-BD	Project leader	20%	20	2400	48000	
CO-FS	Senior researcher	28%	15	2800	42000	
CO-DI	PhD student	50%	18	3000	54000	
P1-PI	Partner leader	20%	20	2400	48000	
P1-MP	Senior researcher	20%	3	2000	6000	
P1-AP	Researcher	28%	15	2200	33000	
P1-AB	Postdoc	50%	19	3000	57000	
P2-BS	Partner leader	20%	12	6080	54720	18240
P2-SB	Researcher	15%	12	3420	30780	10260
P2-FI	Researcher	20%	12	4560	27360	27360

Risk management

Technical risks:

- Incomplete training data. A bank may have little information about the other end of transactions with clients of another bank or located in another country. While all data regarding the sender are known, the receiving part of the link is obscured. So, part of the graph structure may be also missing. Methods to enrich such graphs or to aggregate information are available, but the graph thus generated may potentially contain artificial communities. This is an inherent risk; we will select the most adequate graph completion methods and evaluate them whenever possible.

- Delays due to technical difficulties: some activities are more complex than expected and require more time. We will focus more effort on these activities and try to identify early signs of potential delays.

- The anomaly detection performance is below expectations. In this case, we will try alternative methods (DL or others).

Management risks:

- Technical conflicts. Action: get opinions from other team members, even if not directly involved in that task; get informal opinions from expert colleagues. Debate and vote.

- Delays due to unbalanced workload: academic obligations (for UPB and UB) or superposition with work at other projects (Tremend). Action: active management and possible replanning of some activities. There is some room in the last six months, see the Gantt chart.

Human resources risks:

- Team members that leave. Action: attract master students to this research topic, starting immediately after submitting the Graphomaly proposal.

C Bibliography

[1] Akoglu, L., M. McGlohon, and C. Faloutsos.“Oddball: Spotting anomalies in weighted graphs”.In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*.2010,pp. 410–421.

[2] **Budulan, Ș.**, E. Burceanu, T. Rebedea, and C. Chiru.“Continuous user authentication using machine learning on touch dynamics”.In: *International Conference on Neural Information Processing*.2015,pp. 591–598.

[3] Chen, Z., W. Hendrix, and N.F. Samatova.“Community-based anomaly detection in evolutionary networks”.In: *Journal of Intelligent Information Systems* 39.1 (2012), pp. 59–85.

[4] Colladon, A.F. and E. Remondi.“Using social network analysis to prevent money laundering”.In: *Expert Systems with Applications* 67 (2017), pp. 49–58.

[5] Cucuringu, M., V. Blondel, and P. Van Dooren.“Extracting spatial information from networks with low order eigenvectors”.In: *Physical Review E* 87 (2013).

[6] Ding, Kaize, Jundong Li, and Huan Liu.“Interactive Anomaly Detection on Attributed Networks”.In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*.ACM. 2019,pp. 357–365.

[7] **Dumitrescu, B.** and C.D. Giurcăneanu.“Adaptive-Size Dictionary Learning Using Information Theoretic Criteria”.In: *Algorithms* 12.9 (Sept. 2019), pp. 1–13.

[8] **Dumitrescu, B.** and **P. Irofti**.*Dictionary Learning Algorithms and Applications*.Springer. 2018,pp. XIV, 284.ISBN: 978-3-319-78673-5.

[9] **Dumitrescu, B.** and **P. Irofti**.“Regularized K-SVD”.In: *IEEE Signal Proc. Letters* 24.3 (Mar. 2017), pp. 309–313.

[10] **Dumitrescu, B.**, A. Onose, P. Helin, and I. Tăbuş.“Greedy Sparse RLS”.In: *IEEE Trans. Signal Proc.* 60.5 (May 2012), pp. 2194–2207.

[11] Elliott, A., M. Cucuringu, M.M. Luaces, P. Reidy, and G. Reinert.“Anomaly detection in networks with application to financial transaction networks”.In: *Arxiv:1901.00402 [stat.ap]* (2018).

[12] Gao, J., N. Du, W. Fan, D. Turaga, S. Parthasarathy, and J. Han.“A multi-graph spectral framework for mining multi-source anomalies”.In: *Graph Embedding for Pattern Analysis, Springer* (2013), pp. 205–227.

[13] Huang, D., D. Mu, L. Yang, and X. Cai.“CoDetect: financial fraud detection with anomaly feature detection”.In: *IEEE Access* 6 (2018), pp. 19161–19174.

[14] **Irofti, P.** and **A. Băltoiu**.“Malware Identification with Dictionary Learning”.In: *27th European Signal Processing Conference*.2019,pp. 1–5.

[15] **Irofti, P.** and **B. Dumitrescu**.“GPU Parallel Implementation of the Approximate K-SVD Algorithm Using OpenCL”.In: *22nd European Signal Processing Conference*.2014,pp. 271–275.

[16] **Irofti, P.** and **B. Dumitrescu**.“Pairwise Approximate K-SVD”.In: *Acoustics Speech and Signal Processing (ICASSP), 2019 IEEE International Conference on*.2019,pp. 3677–3681.

[17] **Irofti, P.** and **F. Stoican**.“Dictionary Learning Strategies for Sensor Placement and Leakage Isolation in Water Networks”.In: *The 20th World Congress of the International Federation of Automatic Control*.2017,pp. 1589–1594.

[18] Jiang, Z., Z. Lin, and L.S. Davis.“Label Consistent K-SVD: Learning a Discriminative Dictionary for Recognition”.In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.11 (Nov. 2013), pp. 2651–2664.

[19] Larik, A.S. and S. Haider.“Clustering based anomalous transaction reporting”.In: *Procedia Computer Science* 3 (2011), pp. 606–610.

[20] Li, Z., H. Xiong, Y. Liu, and A. Zhou.“Detecting blackhole and volcano patterns in directed networks”.In: *2010 IEEE International Conference on Data Mining*.2010,pp. 294–303.

[21] Miller, B.A., N. Arcolano, and N.T. Bliss.“Efficient anomaly detection in dynamic, attributed graphs: Emerging phenomena and big data”.In: *IEEE International Conference on Intelligence and Security Informatics*.2013,pp. 179–184.

[22] Miller, B.A., M.S. Beard, and N.T. Bliss.“Eigenspace analysis for threat detection in social networks”.In: *Proceedings of the 14th International Conference on Information Fusion (FUSION)*.2011,pp. 1–7.

[23] Miller, B.A., M.S. Beard, P.J. Wolfe, and N.T. Bliss.“A spectral framework for anomalous subgraph detection”.In: *IEEE Transactions on Signal Processing* 63.16 (2015), pp. 4191–4206.

[24] Miller, B.A., N.T. Bliss, and P.J. Wolfe.“Toward signal processing theory for graphs and non-Euclidean data”.In: *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*.2010,pp. 5414–5417.

[25] Necoară, I., P. Richtarik, and **A. Pătrașcu**.“Randomized projection methods for convex feasibility problems: conditioning and convergence rates”.In: *SIAM Journal on Optimization* (2019).

[26] Nguyen, V.H., V.M. Patel, N.M. Nasrabadi, and R. Chellappa.“Design of Non-Linear Kernel Dictionaries for Object Recognition”.In: *IEEE Trans. Image Proc.* 22.12 (Dec. 2013), pp. 5123–5135.

[27] Onose, A. and **B. Dumitrescu**.“Adaptive Randomized Coordinate Descent for Sparse Systems: Lasso and Greedy Algorithms”.In: *IEEE Trans. Signal Proc.* 63.15 (Aug. 2015), pp. 4091–4101.

[28] Pastor-Satorras, R. and C. Castellano.“Distinct types of eigenvector localization in networks”.In: *Scientific Reports* 6 (2016).

[29] **Pătrașcu, A.** and I. Necoară.“Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization”.In: *Journal of Global Optimization* 61(1) (2015), pp. 19–46.

[30] **Pătrașcu, A.** and I. Necoară.“Nonasymptotic convergence of stochastic proximal point methods for constrained convex optimization.”In: *Journal of Machine Learning Research* 18 (2018), pp. 1–42.

[31] **Pătrașcu, A.** and I. Necoară.“Random coordinate descent methods for ℓ_0 regularized convex optimization”.In: *IEEE Transactions on Automatic Control* 60(7) (2015), pp. 1811–1824.

[32] Rusu, C. and **B. Dumitrescu**.“Stagewise K-SVD to Design Efficient Dictionaries for Sparse Representations”.In: *IEEE Signal Proc. Letters* 19.10 (Oct. 2012), pp. 631–634.

[33] Savage, D., Q. Wang, P. L. Chou, X. Zhang, and X. Yu.“Detection of money laundering groups using supervised learning in networks”.In: *Corr, abs/1608.00708* (2016).

[34] **Stoican, F.** and **P. Irofti**.“Aiding dictionary learning through multi-parametric sparse representation”.In: *Algorithms* 12.7 (2019), pp. 131.

[35] Tong, H. and C.-Y. Lin.“Non-negative residual matrix factorization with application to graph anomaly detection”.In: *Proceedings of the 2011 SIAM International Conference on Data Mining*.2011,pp. 143–153.

[36] Yu, Wenchao, Wei Cheng, Charu C. Aggarwal, Kai Zhang, Haifeng Chen, and Wei Wang.“NetWalk: A Flexible Deep Embedding Approach for Anomaly Detection in Dynamic Networks”.In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.ACM. 2018,pp. 2672–2681.

[37] Zhang, Q. and B. Li.“Discriminative K-SVD for Dictionary Learning in Face Recognition”.In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition*.2010,pp. 2691–2698.