Received December 27, 2019, accepted January 11, 2020, date of publication January 20, 2020, date of current version January 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2967974

Combining Network Visualization and Data Mining for Tax Risk Assessment

WALTER DIDIMO[1], **LUCA GRILLI**[1], **GIUSEPPE LIOTTA**[1], **LORENZO MENCONI**[2], **FABRIZIO MONTECCHIANI**[1], **AND DANIELE PAGLIUCA**[1,2]

[1]Dipartimento di Ingegneria, Università degli Studi di Perugia, 06125 Perugia, Italy
[2]Agenzia delle Entrate, 00147 Roma, Italy

Corresponding author: Walter Didimo (walter.didimo@unipg.it)

This work was supported in part by MIUR under Grant 20174LF3T8 "AHeAD: efficient Algorithms for HArnessing networked Data," and in part by the Dipartimento di Ingegneria dell'Università degli Studi di Perugia, Grant Ricerca di Base 2018, "Algoritmi e sistemi di analisi visuale di reti complesse e di grandi dimensioni."

ABSTRACT This paper presents a novel approach, called MALDIVE, to support tax administrations in the tax risk assessment for discovering tax evasion and tax avoidance. MALDIVE relies on a network model describing several kinds of relationships among taxpayers. Our approach suitably combines various data mining and visual analytics methods to support public officers in identifying risky taxpayers. MALDIVE consists of a 4-step pipeline: (*i*) A social network is built from the taxpayers data and several features of this network are extracted by computing both classical social network indexes and domain-specific indexes; (*ii*) an initial set of risky taxpayers is identified by applying machine learning algorithms; (*iii*) the set of risky taxpayers is possibly enlarged by means of an information diffusion strategy and the output is shown to the analyst through a network visualization system; (*iv*) a visual inspection of the network is performed by the analyst in order to validate and refine the set of risky taxpayers. We discuss the effectiveness of the MALDIVE approach through both quantitative analyses and case studies performed on real data in collaboration with the Italian Revenue Agency.

INDEX TERMS Tax risk assessment, tax evasion discovery, network visualization, data mining, human-computer interaction.

I. INTRODUCTION

Tax noncompliance is a serious economic problem for many countries. It consists of a range of activities, such as tax evasion and tax avoidance, that undermine the government's tax system. As a consequence, a fundamental goal is to reduce the so-called *tax gap*, that is, the difference between the tax amount that should be collected and the actually collected amount. For example, in the United States, the estimated tax gap for the period 2008-2010 was about 458 billion USD per year [1], while in Europe, the estimated VAT (Value Added Tax) gap for the year 2016 amounted to 147 billion Euros [2]. Among the European countries, Italy has a severe tax gap, which is estimated at over 97 billion Euros per year in the period 2013-2015 [3].

In order to deal with this phenomenon, many tax administrations are experimenting novel solutions that exploit advanced data analytics techniques [4]–[6]. In particular, the Italian Revenue Agency (Agenzia delle Entrate), abbreviated as AdE in the following, has put in place various strategies to reduce the tax gap. A key one is building an effective law enforcement policy based on an accurate assessment of the tax risk associated with the different taxpayers. Namely, the main purpose of this assessment is to identify those taxpayers who are more likely to be involved in relevant tax evasion activities; such taxpayers are then subject to fiscal audits.

A. OUR CONTRIBUTION

To support AdE tax officers in the tax risk assessment, we present a novel approach, called MALDIVE (MAtch, Learn, DIffuse, and VisualizE), that combines different data mining and data analytics methods, such as graph pattern matching, social network analysis, machine learning, information diffusion, and network visualization. This approach has been designed and implemented in collaboration with

The associate editor coordinating the review of this manuscript and approving it for publication was Alba Amato.

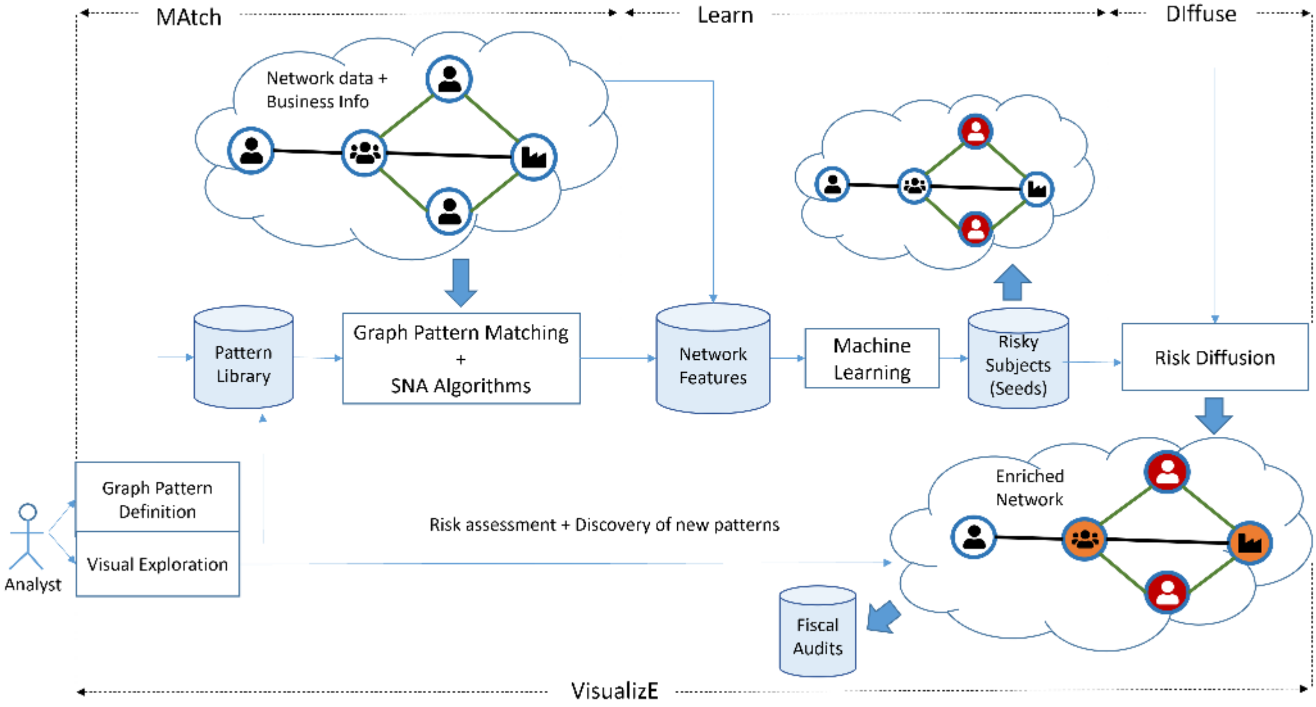


FIGURE 1. A high-level scheme of the proposed approach for the tax risk assessment of taxpayers. The MALDIVE approach follows a pipeline that combines various data mining and analytics methods in order to support a human decision-maker.

the AdE. We remark that MALDIVE is specifically conceived to keep the public administration decision-making process under control, by allowing public officers to better analyze and validate the results provided by automatic classification techniques. The approach is based on the following conceptual pipeline (refer to Fig. 1):

1) **MAtch.** We construct a *social network* where taxpayers are interconnected by various types of relationships, such as economic transactions, shareholdings, and corporate offices. The idea of modeling the taxpayers data by means of a suitable network is motivated by the fact that many real cases of tax evasion are implemented through the interaction of various subjects. Examples include carousel VAT frauds, exchange of false invoices, and tax avoidance implemented through transfer pricing between corporate groups (see, e.g., [7]–[10]). The MAtch phase is aimed to define suspicious graph patterns that represent risky schemes in the taxpayer social network. Such patterns are stored in a pattern library and *matched* in the social network in order to retrieve risky subjects. Based on this social network and on the matched results, we compute both classical social network analysis (SNA) indexes and domain-specific indexes to highlight the most relevant actors. We recall that the use of social network indexes in fraud-risk assessment has been positively evaluated in many papers (see, e.g., [11]–[14]).
2) **Learn.** Next, we make use of a tax risk forecasting model in which machine learning algorithms are trained not only on standard business features of the taxpayers but also considering their social network indexes computed in the previous phase. The forecasting model is trained on the basis of the outcome of previous fiscal audits and it turns out to be quite effective on identifying the most risky taxpayers. The output of the forecasting model is used to enrich the social network and is passed to the next phase.
3) **DIffuse.** Based on the general consensus that risky subjects can negatively influence the behavior of their business partners [15]–[19], we apply an information diffusion method to propagate the fiscal risk in the taxpayer social network. The diffusion process is based on a stochastic model that simulates the spread of an information over an underlying network (see, e.g., [20]). At the end of this phase, a fiscal risk score is assigned to the taxpayers of the network.
4) **VisualizE.** The social network enriched with the fiscal risk scores is the input of a network visualization interface. The purpose of this phase is to support the analyst in validating the fiscal risk scores assigned by the previous phases. This human validation activity is fundamental for the tax administration, which must have complete control over the taxpayer selection process. In fact, thanks to a visual exploration of the social network, the analyst can better assess the real risk profile of taxpayers, thus carrying out a more effective selection of tax audits [9], [10]. Namely, the analyst can find new risky graph patterns or false negative cases, as it will be

shown in our case studies. This information closes the loop of the system by enriching the pattern library and by improving the performance of the forecasting model.

The remainder of this paper is structured as follows. In Section II we review the main literature related to our research. Section III describes the MALDIVE approach in details and our implementation in a real-world system. Section IV discusses some case studies performed on real data to give evidence of the validity of our approach. Section V concludes the paper and proposes future research directions.

II. RELATED WORK

In this section we survey the main literature concerning the application of data mining and data analytics methods in the field of financial crime and fiscal risk analysis. We distinguish between three categories, namely network-based analysis, data mining and machine learning approaches, and visualization approaches; for each of these categories we also remark similarities and differences between the cited papers and our work.

A. NETWORK-BASED ANALYSIS

Public administrations are often asked to monitor the correctness of the behavior of economic and financial operators. This type of assessment can strongly benefit from network-based analysis, where operators are seen as actors that relate to each other within an interdependent system. The importance of analyzing economic transactions over a network of subjects rather than focusing on individual entities has been recently emphasized in [21]. Numerous papers highlight the relevance of SNA (Social Network Analysis) to detect suspicious cases of money laundering. For example: Didimo *et al.* [11] present a system equipped with several indexes to measure the centrality of each actor in the financial network; Drezewski *et al.* [12] analyze criminal group activities by means of SNA metrics that determine the roles of financial operators; Fronzetti Colladon and Remondi [13] show the importance of the centrality metrics to assess and predict risk profile in money laundering analyses.

A crucial aspect of tax risk analysis is to study the spread of fiscal behavior within the taxpayer social network. Recent studies use agent-based models to explain how the tax noncompliance behavior is influenced by the connections between taxpayers and then propagated within the network [16]–[18]. This line of research motivates the use of information diffusion models in our approach.

Network-based analysis are also applied in various economic and financial fields, as well as in crime detection systems, to detect anomalous relational patterns [10], [22]–[29]. The Belgian government uses a data mining technique to find corporate residence frauds [23]. Vlasselaer *et al.* [19] apply a network-based approach to detect frauds related to companies which intentionally go bankrupt in order to avoid paying taxes. Network-based methods to find a specific type of tax evasion performed by affiliated-transaction and involving medium and large companies are proposed in [10], [25], [26].

The scope of our work is different from the aforementioned papers. Firstly, we aim at identifying a wider set of tax evasion patterns and adopt a flexible approach to support the continuous evolution of tax evasion and financial fraud schemes. Secondly, our focus is more on small and medium-sized enterprises, which mainly characterize the economic environment in Europe and particularly in Italy (see, e.g., [30]).

B. DATA MINING AND MACHINE LEARNING APPROACHES

Several studies reveal that data mining and machine learning approaches can increase the audit selection performance against traditional methods, which require strong manual interventions by tax officers [31]–[33]. An association rule technique to identify VAT evasion tax reports is applied in [33]. Other rule-based approaches are proposed for detecting fraudulent VAT credit claims [34] and for the identification of frequent fraud patterns in the Brasilian fiscal environment [35]. Some tax administrations use clustering techniques to detect groups of taxpayers characterized by non compliance behavior [36], [37]. The Chile administration exploits neural networks, Bayesian networks, and decision trees to detect tax evasion performed with the use of fake invoices [8]. The Iranian tax administration adopts a hybrid intelligent system that combines multilayer perceptron neural networks, support vector machines, and logistic regression to detect corporate tax evasion [38]. Kim *et al.* [39] develop multi-class financial misstatement detection models for discovering fraudulent activities. Höglund [40] proposes a genetic algorithm-based decision support tool for predicting tax payment defaults.

In comparison with these papers, our approach allows tax officers to better validate and understand the classification results by visually exploring risky taxpayers and their relationships. Furthermore, to enhance the forecasting performance in our domain, unlike the above papers, we consider new features retrieved from a suitably defined taxpayer social network, such as the presence of a subject in one or more suspicious relational patterns.

C. VISUALIZATION APPROACHES

The importance of using the visual channel to identify and analyze economic and financial frauds is well described in the literature [41], [42]; see also the survey in [43]. Application examples of visualization approaches in the economic and financial fields include: systems for financial fraud and money laundering detection based on the analysis of banking data [11], [12], [44], [45]; visual analytics techniques for financial stability monitoring and fraud detection in financial markets [46], [47]; decision support systems for tax evasion discovery [9], [48]. In the specific domain of financial transaction analysis, few works propose hybrid approaches that combine visual analytics and machine learning techniques to support human decisions. An early work of Kirkland *et al.* [49] adopts artificial intelligence, visualization, pattern recognition, and data mining to support regulatory analysis, fraud detection alerts, and knowledge discovery; this approach however is not meant for user interaction. The more recent system EVA [50] integrates various techniques to detect fraudulent transactions within a financial institution. It offers a scoring mechanism based on data mining techniques and interactive visual analytics facilities to perform fraud validation, but it is not based on a network representation of the data.

Unlike the above works, our hybrid approach focuses on the fiscal domain and provides an interactive network visualization environment through which tax officers can assess the results provided by machine learning and information diffusion methods. The importance of combining data mining and visualization methods has also been confirmed in [51] for the analysis of crime data and in [52] for the detection of outliers.

III. THE MALDIVE APPROACH: DESIGN AND IMPLEMENTATION

In this section we present in detail our approach MALDIVE and its implementation in place at the AdE. We first describe the MAtch phase, i.e., how we construct the taxpayer social network from the raw data and the subsequent graph pattern matching step (Section III-A). Then, we illustrate the Learn phase and its corresponding fiscal risk forecasting model (Section III-B). Finally, we describe the DIffuse phase, which is based on a stochastic spreading process of risk scores (Section III-C), and the VisualizE phase, which allows human analysts to visually explore the taxpayer social network enriched with the fiscal risk scores (Section III-D).

A. THE MATCH PHASE

The data sources queried by the AdE officers are modeled as a unified *taxpayer social network* G. Each node v of G is a single taxpayer, which can be either an individual or a legal person, like a private company or a public institution. Many business attributes are associated with v, including the type of economic activity, the geographic location and territorial scope, the declared income, the amount of VAT credits/debts and of VAT refunded/paid, and the amount of economic exchange within the European Union. The aforementioned business and fiscal features mainly concern the data recovered from the tax register, tax returns, and financial statements. The edges of G are directed edges. An edge (u, v) can model different types of relationships between u and v. From the economic point of view, the main types of relationships considered in this work are *economic transactions*, *shareholdings*, and *corporate positions*. For an economic transaction (u, v), the source node u is the seller and the target node v is the buyer; the two main attributes for such an edge are the transaction amounts declared by the two subjects in the considered time window. For a shareholding (u, v), the source u is the shareholder and the target v is the participated company; the main attribute for this edge is the percentage of share. For a corporate position (u, v), the source u is the subject holder of corporate positions and the target v is the assigning company; the main attribute for this edge is the type of corporate position. For example, the taxpayer social network constructed from the data of the Tuscany region in Italy for the tax year 2014 consists of about 700,000 nodes and 1,800,000 edges. We remark that our model can be easily extended with more types of relationships, for example depending on the specific application domain and/or geographic context.

The MAtch phase allows analysts to define suspicious graph patterns that represent risky schemes in the taxpayer social network. Example of risky schemes are those that arise from carousel VAT frauds, exchange of false invoices, and tax avoidance implemented through transfer pricing between corporate groups. For instance, Fig. 2 highlights a suspicious pattern identified in a larger network, which we call `SuppliesFromAssociated`, that models the exchange of high transactions among taxpayers participated by the same owner. More formally, a graph pattern P is defined as a pair $\langle G_P, R_P \rangle$, where $G_P = (V_P, E_P)$ is a graph that describes the topology of P, and R_P is a set of rules on the nodes and the edges of G_P. An edge of E_P corresponds to a single edge of G or to a path whose length is within a desired range. This correspondence is established by a specific type of rules of R_P. Other types of rules in R_P are used to describe desired properties for node/edge business attributes of G_P; these properties can then be combined with logical operators AND, OR, NOT. Regular expressions should also be usable to specify classes of values for the node and edge attributes of G_P. For example, in the pattern of Fig. 2 graph G_P consists of three nodes and three directed edges. The set R_P of rules specifies that: (i) a node of G_P represents a person while the other two nodes represent companies AND (ii) the edges of G_P describe shareholdings of the person in the two companies (see the two green edges) and economic transactions above a desired threshold t between the two companies (see the weighted black edge).

The patterns defined in the MAtch phase are stored in a *pattern library*, which is one of the knowledge bases of our approach. For example, Table 1 reports the pattern library defined by tax experts of the AdE and used to extract specific network features for the fiscal year 2014. We refer the reader to [9] for further details about some of the risky fiscal patterns and their business rules.

Our specific implementation for the MAtch phase is built on top of the popular Neo4J[1] graph database, which is particularly effective and efficient in executing graph pattern matching tasks. The native query language of Neo4J, called *Cypher*, can be used to specify graph patterns to be searched in the network. In order to make the MAtch phase more user friendly, one can exploit a visual query language that allows analysts to manually draw a graph pattern and that automatically translates this pattern into an equivalent Cypher

[1] http://neo4j.com

TABLE 1. Patterns defined by the AdE tax experts and used to extract specific network features for the fiscal year 2014.

Pattern ID	Pattern name	Pattern Description
1	`Missing Trader Suppliers`	Economic transactions of high amount where the seller is a missing trader with serious tax irregularities.
2	`Purchase From Related`	High-value purchases from a participated or represented company.
3	`Risky Activity Suppliers`	High-value purchases from suppliers that carry out economic activities at risk.
4	`Intra Community Services`	High-value purchases of services from intra-community operators.
5	`Closed Intra Community`	Economic transactions of high amount with intra-community operators who closed the economic activity.
6	`Capital Movements`	Individuals with low income and high foreign capital movements linked to companies.
7	`Supplies From Associated`	Triangular pattern in which there are economic transactions of high amount between companies participated or represented by the same subjects.

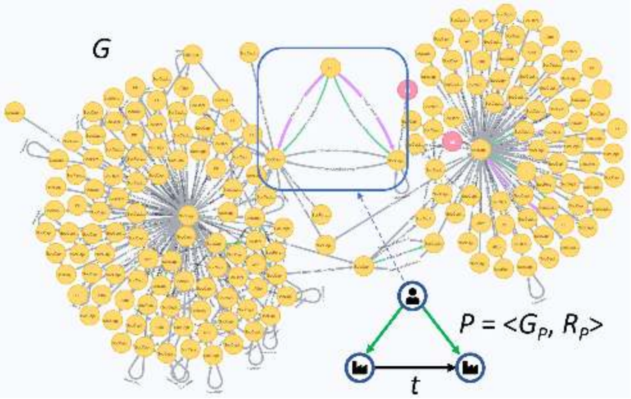


FIGURE 2. In the MAtch phase, tax officers encode risky relational schemes among taxpayers into suspicious graph patterns to be searched in the taxpayer network. The pattern in the figure represents a `SuppliesFromAssociated` scheme, consisting of an economic transaction (black edge) and two shareholding relationships (green edges).

query (see, e.g., [53], [54]). In our implementation we use the visual language described in [9], [48], which is specifically tailored to the fiscal domain.

B. THE LEARN PHASE

The goal of this phase is to predict the fiscal risk of the taxpayers. This can be accomplished by exploiting different supervised forecasting models and different types of features related to both the structural properties of the network and the node/edge attributes in the specific fiscal domain. In the following we describe different implementations of the Learn phase and we report the results of an experimental comparison between them. These results provide useful insights for the design of the Learn phase in application domains similar to ours.

We considered different supervised forecasting models trained on the samples of previous fiscal audits. We used a sample of $2,790$ fiscal audits for the fiscal year 2014. A fiscal audit is positive if there are Assessed Additional Taxes (AAT), while it is negative if no tax evasion or tax avoidance were detected and therefore there are no AAT. Given that tax administration has limited resources and there are administrative costs associated with the tax assessment process, the efforts are primarily focused on the most serious cases of tax evasion. Following this strategy, we used a threshold value for the amount of AAT to identify the taxpayers with greater risk. Each fiscal audit is assigned one of two possible labels (i.e., two classes):

- `(LOW_RISK)` Taxpayers with either negative or not relevant fiscal audits (i.e., whose AAT values are less than the chosen threshold).
- `(HIGH_RISK)` Taxpayers with relevant fiscal audits (i.e., whose AAT values are greater than or equal to the chosen threshold).

In order to effectively represent the taxpayer social network data, we make use of the following types of features:

1) BUSINESS AND FISCAL FEATURES

We used many business and fiscal attributes associated with taxpayers, including the type of economic activity, the geographic location and territorial scope, the declared income, the amount of VAT credits/debts and of VAT refunded/paid, and the amount of economic exchange within the European Union. The aforementioned business and fiscal features mainly concern the data recovered from the tax register, tax returns, and financial statements.

2) PATTERN FEATURES

The results retrieved in the MAtch phase are merged into a temporary Risky Relationship Network (RRN) so to compute the pattern centrality measures introduced by [9]. A central actor in the RNN is a taxpayer involved in many suspicious patterns. Thus, we used these measures to compute a set of pattern features that represent the centrality of each taxpayer in the RRN.

3) SNA FEATURES

This set of features should bring information about the relevance of a taxpayer in the whole social network. For each type of relation, we use as features several common measures of centrality so to identify the most relevant social actors [55]. Namely, for each taxpayer we consider the in-degree, the out-degree, the degree, the closeness, the betweenness, and the page rank [56] centrality measures. In addition, since some specific cases of tax evasion are carried out by operators who present occasional but high-value economic transactions, we consider for each taxpayer the set of weights of its incident economic edges and use as features the mean, the variance, the skewness, and the kurtosis of this set.

4) MACHINE LEARNING ALGORITHMS

We built four forecasting models for detecting the fiscal risk associated with each taxpayer. These models are logistic regression (LR), support vector machines (SVM), multi layer perceptron neural networks (MLP), and random forests (RF) [57]. We started with the data cleansing of the features vector, which is composed of 177 business/fiscal features and 56 structural (patterns and SNA) features. We transformed categorical features in dummy variables, for a total of 348 features. Considering the high number of features, we performed a selection using these methods: Features with a high percentage of missing values; low correlated features with AAT; collinear (highly correlated) features; features with zero importance in a tree-based model. We used a gradient boosting machine learning model [58] to analyze feature importance (see Fig. 3) and to filter out features with zero importance which are not used to split any nodes of the tree. The filtered features vector is composed of 129 business/fiscal features and 54 structural features. We normalized numeric feature values in 5 categories, using quantiles of the data distribution. The adopted method tends to spread out the most frequent values and it also reduces the impact of outliers. This method makes variables measured at different scales more directly comparable.

For reasons of confidentiality we are not authorized to show data related to the impact of taxpayers' business and fiscal features in the forecasting models. However, we show some analyses conducted on structural variables. Fig. 4 shows the Pearson correlation heat map computed for the network-based (patterns and SNA) features.

By looking at Fig. 4, we can see that the highest correlation value with the AAT is with a feature (patternId01value) related to pattern 1 (`MissingTraderSuppliers`) of Table 1, which is the most important structural feature as shown in Fig. 3. We recall that this pattern is based on the presence of economic relationships with missing traders subjects who have serious tax irregularities, such as omitted VAT payments or tax declarations.

The negative correlation of AAT with the centrality indexes related to the corporate positions and shareholdings is also interesting. Among these indexes, one can observe that in fact the "degree_corporateposition" is the second most important structural feature as shown in Fig. 3. This data highlights how in a company with many shareholders, directors, and corporate controllers, there is a smaller risk of serious tax violations.

In order to train the different models and in order to subsequently evaluate the performances of these models, we partitioned the whole sample of $2,790$ fiscal audits into two parts: A *training set*, consisting of 70% of the fiscal audits and a *test set*, containing the remaining 30% of audits. We used a 10-fold cross validation technique to reduce the problem of over fitting in the training phase and we used a grid search algorithm to find the best tuning parameters of the models. The results on the *test set* are based on the metrics defined in Table 2 and they are shown in Table 3.

TABLE 2. Description of the metrics used to compare the fiscal risk forecasting models.

Acronym	Metric	Description
TP	True Positive	Number of taxpayers with *high-risk* correctly classified in the *high-risk* class.
FP	False Positive	Number of taxpayers with *low-risk* wrongly classified in the *high-risk* class.
TN	True Negative	Number of taxpayers with *low-risk* correctly classified in the *low-risk* class.
FN	False Negative	Number of taxpayers with *high-risk* wrongly classified in the *low-risk* class.
ACC	Accuracy	$(TP + TN)/(TP + FP + FN + TN)$
AUC	Area Under the ROC Curve	Area Under the Receiver Operating Characteristic (ROC) Curve from prediction scores. ROC curves display the sensitivity versus the specificity. The higher the area, the better the model performs.
p	Precision	$TP/(TP + FP)$
r	Recall or Sensitivity	$TP/(TP + FN)$
F1	F1 score	$F1 = 2 * (p * r)/(p + r)$

The Random forest model obtained the best result in terms of Accuracy (74.29), AUCROC (74.29), Precision (75.42), and F1-score (73.66), while the best score for the Recall was achieved by the MLP model (75.63). The Logistic Regression model exhibited the worst performance for all metrics, still with values higher than 70%.

We conducted further experiments to compare the forecasting models subject to an additional constraint. Namely, we require that each model decides the class of a data item only if its membership probability to that class is higher than 80%; see Table 4. While this constraint leads to a lower response rate of the model, it is particularly suitable for tax administrations who have to monitor a large number of subjects and who must select the taxpayers with the greatest probability of being in the *high-risk* class. In particular, in the Italian economic context, the analysis of tax risk must take into consideration a large number of small businesses and professionals (for example, in the Tuscany region there are over 400,000 business operators).

Finally, on our specific dataset, we evaluated the impact of replacing our business features with others previously used in the fiscal domain, namely the set of 21 financial features proposed in the study of [38]. The results are reported in Table 5. As one can see, the values for the different metrics are significantly lower than those achieved with our business features.

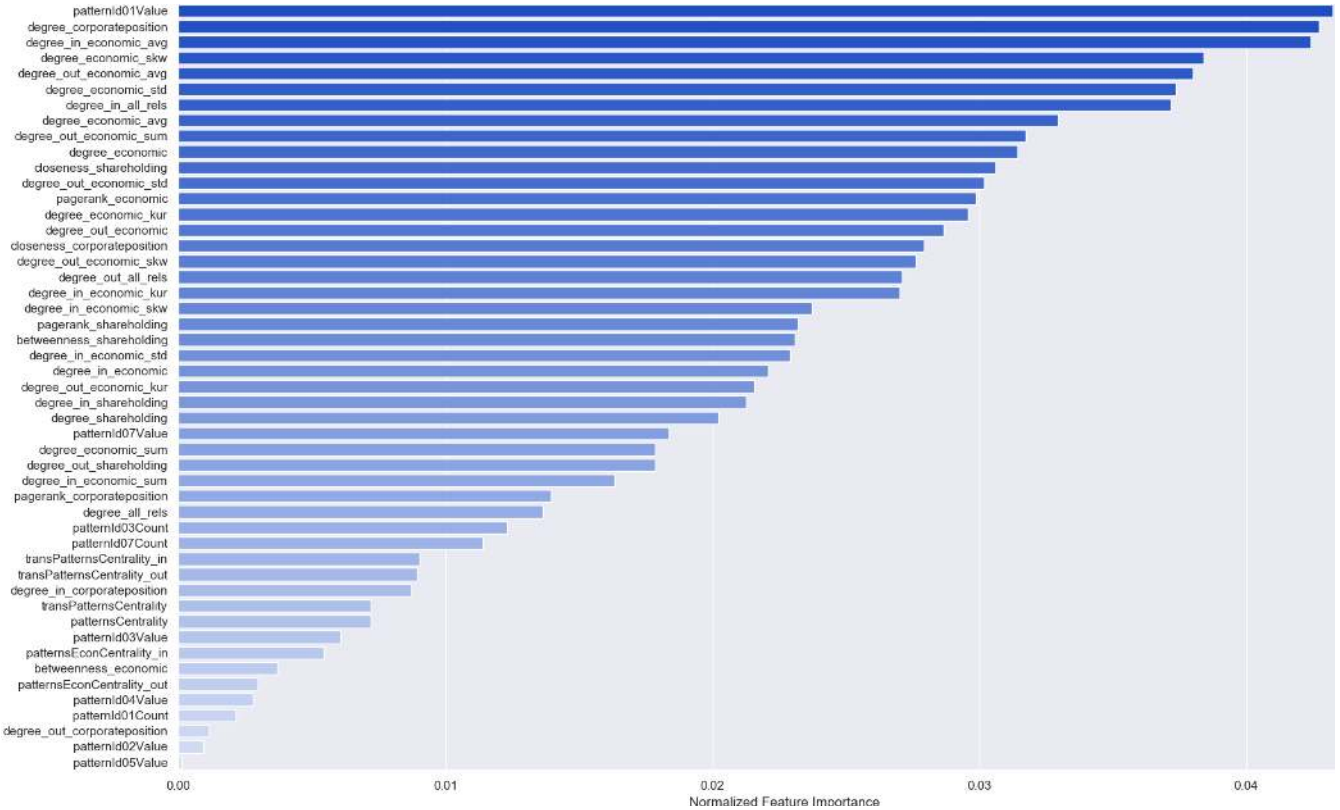


FIGURE 3. Analysis of the structural feature importance using gradient boosting.

TABLE 3. Experiments conducted to compare various forecasting models.

Model Description	Accuracy	AUCROC	Precision	Recall	F1-score
Random Forest	74.29	74.29	75.42	71.98	73.66
MLP Neural Network	72.24	72.25	70.79	75.63	73.13
SVM	71.44	71.45	71.08	72.21	71.64
Logistic Regression	70.31	70.31	70.05	70.84	70.44

TABLE 4. Performance and response rate of the models when the classification is made only if the membership probability is higher than 80%.

Model Description	Accuracy	AUCROC	Precision	Recall	F1-score	Response Rate
MLP Neural Network	80.79	80.80	82.16	78.93	80.51	63.37
RandomForest	86.54	86.71	90.22	83.42	86.68	43.12
SVM	86.73	86.85	90.52	86.07	88.24	24.00

TABLE 5. Performance obtained by replacing our business variables with those suggested in [38].

Model Description	Accuracy	AUCROC	Precision
Logistic Regression	58.77	58.56	58.95
MLP Neural Network	61.84	61.60	63.04
RandomForest	60.53	59.97	66.15
SVM	58.33	58.36	56.90

C. THE DIFFUSE PHASE

Starting from the core of *high-risk* taxpayers determined by the Learn phase, the goal of the DIffuse phase is to identify other taxpayers that are potential risky subjects. This is done by applying an information diffusion method to propagate part of the fiscal risk of the core taxpayers to other nodes of the network. There is, in fact, a general consensus that risky taxpayers may negatively influence the behavior of their business partners [16]–[19]. A diffusion process is modeled as a stochastic process that simulates the spread of an information over an underlying network (see, e.g., [20]).

More precisely, in our implementation of the Learn phase a subset of nodes of the taxpayer social network (possibly all nodes) is labeled as either *high-risk* or *low-risk*. The pseudo-code of our fiscal risk diffusion algorithm is given in

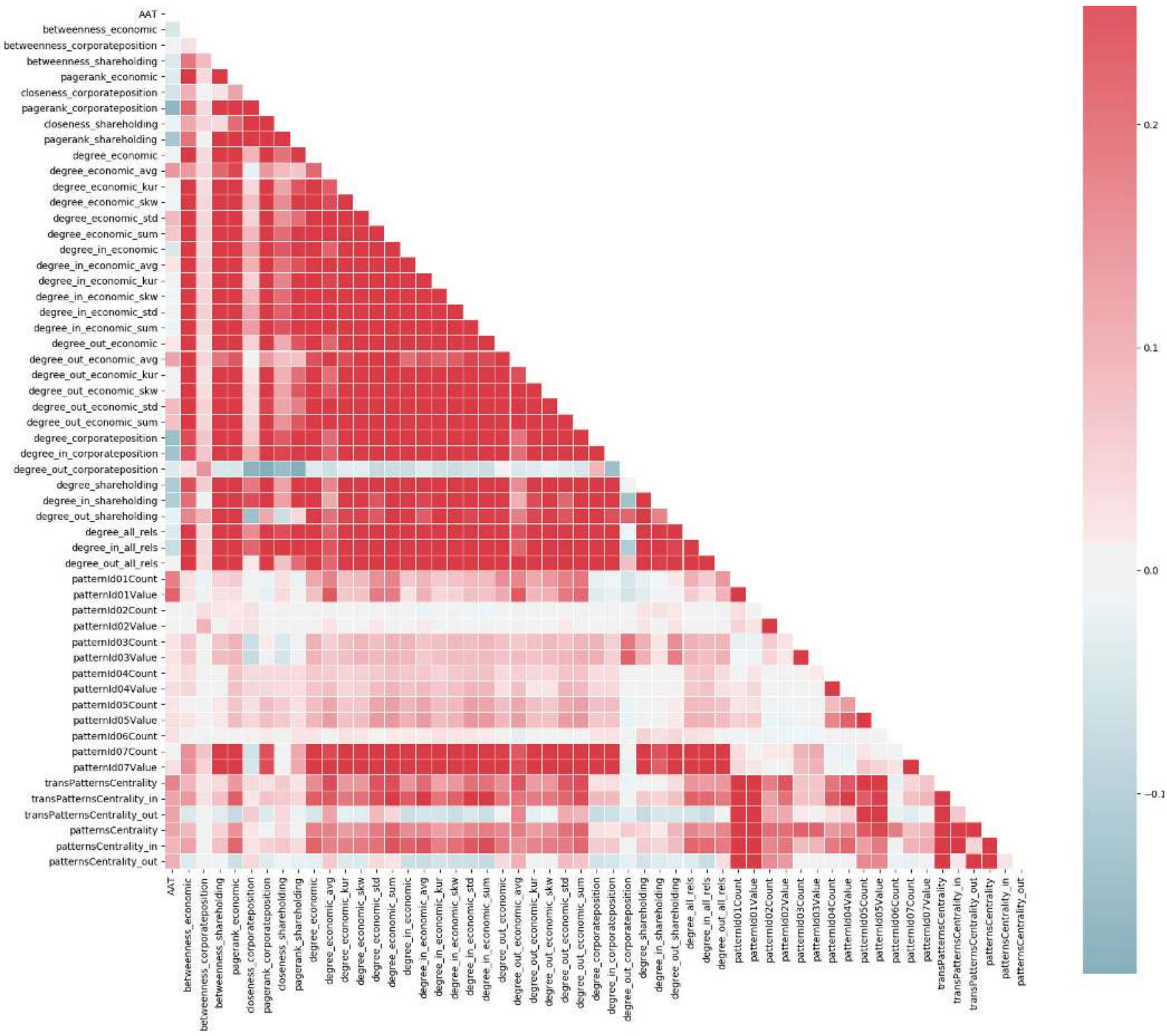


FIGURE 4. Heat map of the Pearson correlation coefficient computed for the structural features. Positive correlations correspond to the red scale; negative correlations correspond to the blue scale.

Algorithm 1. The diffusion process is broken down into a set of iterations. In the first iteration, the risk can be propagated by a set of seeds that consists of the taxpayers classified as *high-risk*. The transmission of the risk between a seed and its linked subjects takes place on the basis of a stochastic process. The greater the weight of the relationship between two taxpayers, the higher the probability of transmitting the risk along that link. The subjects to whom the risk is transmitted become the new seeds for the next iteration, and the process halts when no more seeds are created. The transmitted risk value is reduced during the propagation phases until reaching the conclusion of the process.

In the pseudo-code of Algorithm 1, the *normalize-weights* function maps each edge weight to a real number in an interval $[r_L, r_H] \subseteq [0, 1]$, which can be viewed as a probability. The function is defined as follows. Let $E(u)$ be the set of edges incident to u. For each edge $e = (u, v)$ in $E(u)$, the function normalizes its weight $w(e)$ based on its type. For edges representing economic transactions, the weight corresponds to the amount of the transaction and it is normalized using the MinMax normalization [59], as shown in (1).

$$\begin{aligned} r_L &= 0 \\ r_H &= 0.8 \\ w_{min} &= \min_{e \in E(u)} w(e) \\ w_{max} &= \max_{e \in E(u)} w(e) \\ w_{norm}(e) &= \frac{w(e) - w_{min}}{w_{max} - w_{min}} \cdot (r_L - r_H) + r_L \end{aligned} \tag{1}$$

For edges representing shareholdings, the weight corresponds to the shareholding percentage and it is multiplied by the

Algorithm 1 FiscalRiskDiffusion(G)

```
Require: A weighted taxpayer network $G$
Ensure: Fiscal risk diffusion
$\mathcal{S} \leftarrow$ taxpayers classified as high-risk
for all ($u$ in $\mathcal{S}$)) do
    seed($u$) $\leftarrow$ true
    risk($u$) $\leftarrow 1$
end for
$t \leftarrow 1$
while ($\mathcal{S}$ is not empty) do
    $\mathcal{S}' \leftarrow$ empty
    for all ($u$ in $\mathcal{S}$) do
        $E(u) \leftarrow$ set of edges $e = (u, w)$ incident to $u$
        normalize-weights($E(u)$)
        $N(u) \leftarrow$ set of taxpayers $v$ linked to $u$
        for all ($v$ in $N(u)$) do
            $w_{norm}(e) \leftarrow$ normalized weight of the edge $e = (u, v)$
            $R \leftarrow$ pseudo-random number in the range [0,1]
            if (($R < $ risk($u$) $\cdot w_{norm}(e)$) AND (seed($v$) == false)) then
                risk($v$) $\leftarrow$ risk($u$) $\cdot w_{norm}(e) \cdot \frac{1}{t^2}$
                $\mathcal{S}' \leftarrow \mathcal{S}' \cup \{v\}$
                seed($v$) $\leftarrow$ true
            end if
        end for
    end for
    $\mathcal{S} \leftarrow \mathcal{S}'$
    $t \leftarrow t + 1$
end while
```

value r_H defined in (1). In the case of multiple edges, the weights are added together because the risk is higher, but the maximum value allowed is r_H.

D. THE VISUALIZE PHASE

The last phase of the pipeline is the visual exploration of the social network enriched with the fiscal risk computed by the Learn and DIffuse phases. Through the visualization the tax officer gains a more holistic and systemic view of the network, which makes it easier: (*i*) to validate the risk scores assigned by the previous phases, (*ii*) to detect false negative risky taxpayers, and (*iii*) to discover new suspicious patterns. This phase combines the following ingredients:

- *Node-Link Layout:* The analyst can start from an automatic visualization of a relatively small portion of the taxpayer network. To this aim, a classical node-link layout of the network appears to be a natural and intuitive choice, which is preferred to other types of representations that are more suitable for larger and denser graphs (see, e.g., [60]–[63]).
- *Interactive Network Exploration:* Starting from the initial node-link layout, tax officers can use various interaction features to visually explore the taxpayer network. For example, the analyst can add to the current layout the neighbors of a node or she can remove some nodes from the visualization.
- *Interactive Filtering:* The analyst can hide from the visualization nodes or edges that are less interesting, based on the values of some specific attributes.

In our implementation of the VisualizE phase, we exploit and customize the module provided in [48] to our application domain. Namely, we visually encode the risk of the taxpayers by the background color of the corresponding nodes, as shown in Fig. 5. Also, nodes classified as *high-risk* by the Learn phase are represented with a red border, while nodes classified in the *low-risk* class have a green border color. Alternative implementations of this phase could rely on embeddable tools with built-in Neo4j connections (https://neo4j.com/developer/tools-graph-visualization/).

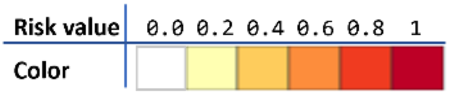


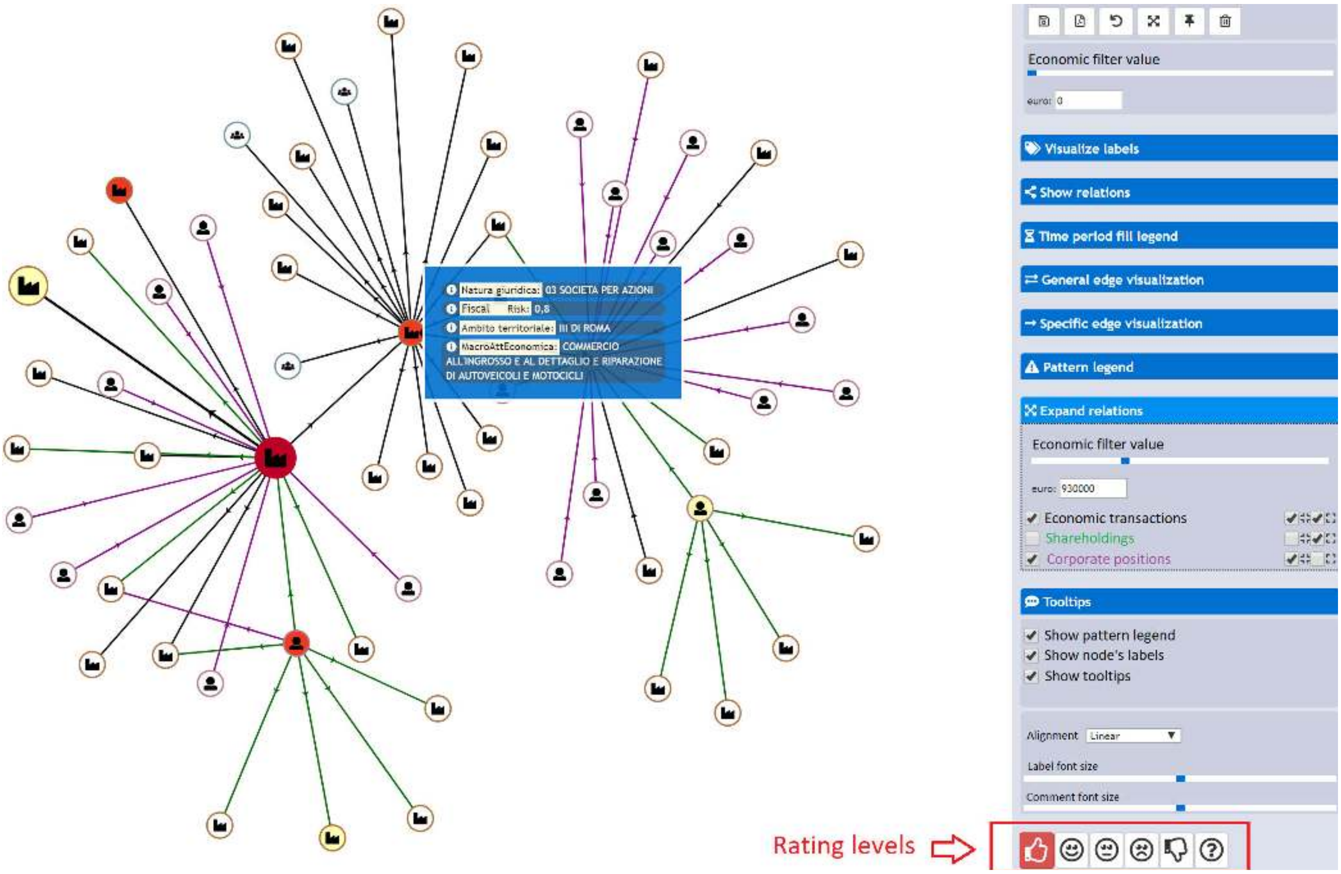
FIGURE 5. Background colors associated with the taxpayer fiscal risks. Risks values are normalized from 0 to 1.

An example of visual exploration is shown in Fig. 6. Starting from a *high-risk* taxpayer, the analyst explored a larger network by repeatedly expanding the relationships of some taxpayers. Also, by using the interactive filtering the analyst hid the shareholding relationships and the economic transactions that are below a threshold value. At the end of the exploration, the user added a feedback to the record of one of the analyzed taxpayers. This is done by selecting the icon corresponding to the rating level in the bottom-right of the interface. In this way the analyst may or may not confirm the risk level originally assigned to a taxpayer by the previous phases.

IV. CASE STUDIES

In this section we illustrate two case studies about the use of the MALDIVE approach. The case studies are based on real data and activities carried out by AdE tax officers.

We built a taxpayer network common to both case studies, based on the AdE dataset for the fiscal year 2014. First, we carried out the MAtch phase based on the pattern library described in Table 1. We then extracted the features from the taxpayer network to perform the Learn phase. We classified the *test set* of the fiscal audits sample labeling its nodes as either *high-risk* or *low-risk*. In particular, for the Learn phase we applied the Random Forest model, that according to our experimental analysis (see Section III-B) is the one that achieves the best score for the AUCROC metric. Subsequently, we applied the DIffuse phase starting from the *high-risk* taxpayers in order to propagate the fiscal risk in the taxpayer network. The resulting enriched network was the input of the VisualizE phase performed by an AdE tax officer. During the case studies, the tax officer exploited the network visual exploration phase to assess the tax risk of two taxpayers


Natura giuridica: 03 SOCIETA PER AZIONI
Fiscal Risk: 0,8
Ambito territoriale: III DI ROMA
MacroAttEconomica: COMMERCIO ALL'INGROSSO E AL DETTAGLIO E RIPARAZIONE DI AUTOVEICOLI E MOTOCICLI

Economic filter value
euro: 0

Economic filter value
euro: 930000
☑ Economic transactions
☐ Shareholdings
☑ Corporate positions

☑ Show pattern legend
☑ Show node's labels
☑ Show tooltips

Alignment: Linear

FIGURE 6. **Visualization of a portion of the taxpayer network. Starting from a *high-risk* subject (dark red background), the analyst expands the network by exploding the relationships of some taxpayers. Black edges represent economic transactions, green edges are shareholdings, and purple edges are corporate positions. At the end of the analysis the analyst may add a personal feedback to the record of the investigated taxpayers.**

one labeled as *high-risk* (Case 1) and the other labeled as *low-risk* (Case 2).

1) CASE 1

For the description of this case study refer to Fig. 7. The tax officer interacts with the Visual Exploration interface to analyze the *high-risk* taxpayer, which represents a company denoted as `v1`. Starting from `v1`, the tax officer visualizes its neighbors. After interacting with the available filters and excluding nodes and relations that are considered not relevant, the tax officer obtains the portion of network shown in Fig. 7a. The attention is captured by the presence of a double relationship between two taxpayers. Namely, node `v1` plays a dual role: It is both a supplier of a company, node `v2` in the figure, and owner of the same company for more than 90% of the share capital. The value of the economic transaction between `v1` and `v2` is high and it recommends further investigation. By enlarging the network, the officer identifies node `v3`, that owns both company `v1` and the remaining 10% of company `v2`. The officer concludes that the economic transaction between `v1` and `v2` is risky due to its value and the presence of correlated interests between the economic parties. To better analyze the tax risk of the subjects involved in the described scheme, the officer decides to visualize the spread of the tax risk obtained with the procedure described in Section III-C and shown in Fig. 7b. The officer realizes that `v2` has been initially classified as *low-risk* by the forecasting model (as represented by the green border), while the diffusion process has increased its risk level (represented by a dark orange background). Thus, after going through the whole pipeline of MALDIVE, the final decision of the officer is to consider `v2` as a *high-risk* node.

2) CASE 2

For the description of this case study refer to Fig. 8. Differently from the previous scenario, in this case study the tax officer starts from a company `n1` classified as *low-risk* by the forecasting model but whose risk has been increased by the diffusion process; in Fig. 8, the border of `n1` is indeed green while its background is dark orange. The officer decides to further investigate the reasons behind the risk diffusion towards `n1`. By considering the economic transactions of `n1`, the officer realizes that three customers of `n1` have been classified as *high-risk* (nodes `n2`, `n7`, and `n9`). The economic transactions between `n1` and these three nodes are all of relevant amount (hundreds of thousands of Euros each). The officer then inspects the corporate structure of company `n1` and finds out that its shareholders (nodes `n4` and `n3`) are

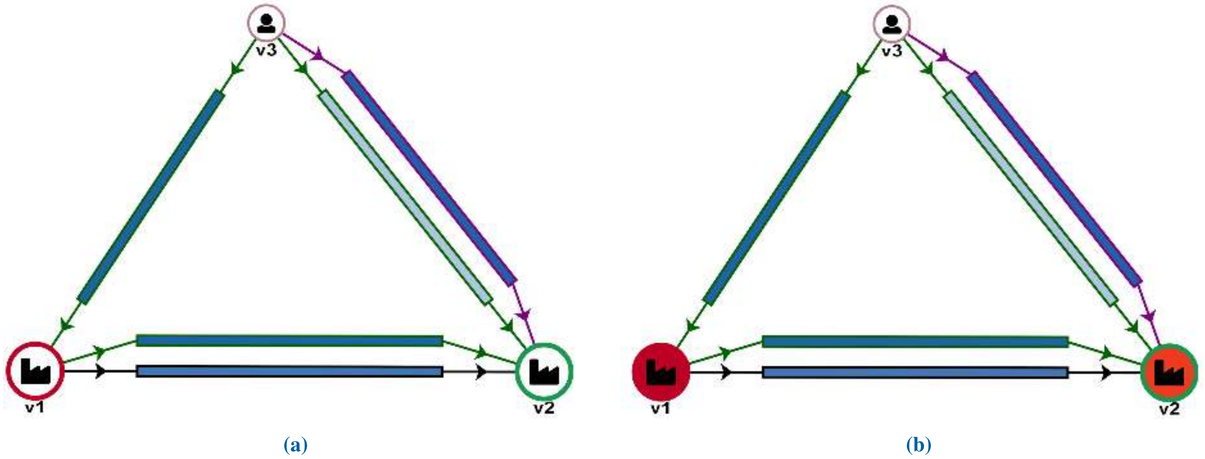


FIGURE 7. Illustration for Case 1. Portion of the taxpayer network before (a) and after (b) the risk diffusion process. Black edges represent economic transactions, green edges are shareholdings and purple edges are corporate positions. The color intensity inside each edge encodes the weight of the relationship.

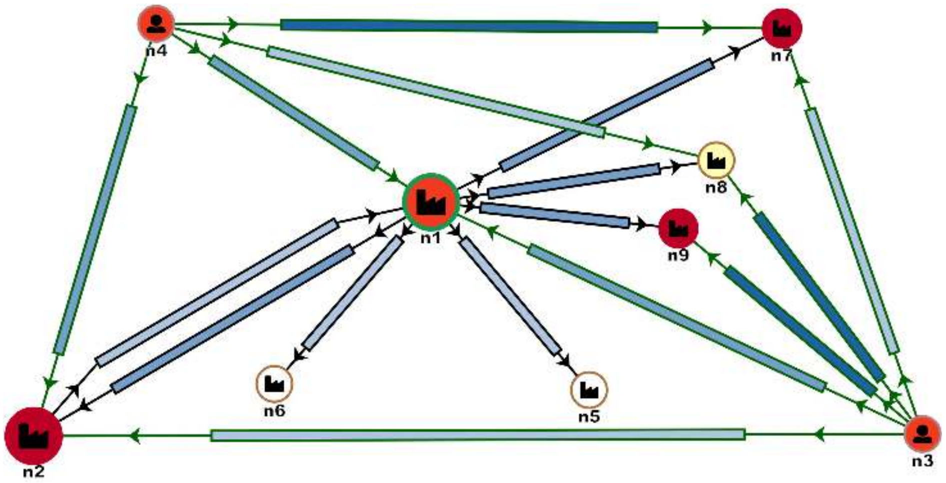


FIGURE 8. Illustration for Case 2. Portion of the taxpayer network after the risk diffusion process.

also shareholders of n2 and n7. This occurrence reveals a common tax evasion strategy that the shareholders n4 and n3 may adopt not only for the *high-risk* companies n2 and n7, but also for company n1. Thus, the officer decides to consider n1 as *high-risk*, confirming the effect of the diffusion process.

V. CONCLUSION AND FUTURE WORK

We presented MALDIVE, a novel approach for fiscal risk analysis combining different data mining and data analytics methods. MALDIVE has been designed and implemented in collaboration with the Italian Revenue Agency (Agenzia delle Entrate, AdE) and has been evaluated on real data by expert tax officers. It combines graph pattern matching, social network analysis, and machine learning to detect risky taxpayers, it exploits information diffusion processes to spread the risk in a suitably constructed taxpayer network, and it offers network visualization features to support the analyst in the exploration of the proposed results.

A key contribution of our approach is the application of combined data analysis techniques in the context of public administration, where it is crucial to always keep the decision-making process under the control of public officers. In fact, our combined approach supports final and critical decisions with a user-centric visualization environment that complements an automated classification pipeline.

In the near future, we plan to use additional data in order to improve the classification performance of our approach. Particular attention will be given to further data available in the fiscal audits, among them: (a) the profitability of a fiscal audit (expressed in terms of the amount of additional taxes paid by the taxpayer), and (b) the relevance of a fiscal audit (expressed in terms of the size of the business activity).

Finally, we are currently collecting the feedback of the analysts on the risk assigned by the system to the taxpayers. Indeed, the analyst can agree or disagree with the risk level of a taxpayer. This feedback, together with new data coming from subsequent fiscal audits, will be used to improve the training of the forecasting models.

ACKNOWLEDGMENT

This work was accomplished in cooperation with the Italian Revenue Agency (AdE). The authors would like to thank in particular Carlo Palumbo, Mario Landolfi, and Giuseppe De Luca for their support.

REFERENCES

[1] Internal Revenue Service. (2016). *Tax Gap Estimates for Tax Years 2008-2010*. [Online]. Available: https://www.irs.gov/pub/newsroom/tax%20gap%20estimates%20for%202008%20through%202010.pdf

[2] C. F. S. CASE and E. Research. (2018). *Study and Reports on the Vat Gap in the eu-28 Member States: 2018 Finalreport*. [Online]. Available: https://ec.europa.eu/taxation_customs/sites/taxation/files/2018_vat_gap_report_en.pdf

[3] Italian Government, Ministry of Economy and Finance. (2018). *Relazione Sull'Economia non Osservata e Sull'Evasione Fiscale e Contributiva Anno 2018*. [Online]. Available: http://www.mef.gov.it/documenti-allegati/2018/A6_-_Relazione_evasione_fiscale_e_contributiva.pdf

[4] *Advanced Analytics for Better Tax Administration*, OECD Publishing, Paris, France, 2016. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/9789264256453-en

[5] *The Changing Tax Compliance Environment and the Role of Audit*. OECD Publishing, Paris, France, 2017. [Online]. Available: https://www.oecd-ilibrary.org/content/publication/9789264282186-en

[6] M. Pijnenburg, W. Kowalczyk, and E. van der Hel-van Dijk, "A roadmap for analytics in taxpayer supervision," *Electron. J. e-Government*, vol. 15, p. 14, Feb. 2017.

[7] L. Šubelj, Š. Furlan, and M. Bajec, "An expert system for detecting automobile insurance fraud using social network analysis," *Expert Syst. Appl.*, vol. 38, no. 1, pp. 1039–1052, Jan. 2011, doi: 10.1016/j.eswa.2010.07.143.

[8] P. C. González and J. D. Velásquez, "Characterization and detection of taxpayers with false invoices using data mining techniques," *Expert Syst. Appl.*, vol. 40, no. 5, pp. 1427–1436, Apr. 2013.

[9] W. Didimo, L. Giamminonni, G. Liotta, F. Montecchiani, and D. Pagliuca, "A visual analytics system to support tax evasion discovery," *Decis. Support Syst.*, vol. 110, pp. 71–83, Jun. 2018.

[10] J. Ruan, Z. Yan, B. Dong, Q. Zheng, and B. Qian, "Identifying suspicious groups of affiliated-transaction-based tax evasion in big data," *Inf. Sci.*, vol. 477, pp. 508–532, Mar. 2019, doi: 10.1016/j.ins.2018.11.008.

[11] W. Didimo, G. Liotta, and F. Montecchiani, "Network visualization for financial crime detection," *J. Vis. Lang. Comput.*, vol. 25, no. 4, pp. 433–451, Aug. 2014.

[12] R. Dre ewski, J. Sepielak, and W. Filipkowski, "The application of social network analysis algorithms in a system supporting money laundering detection," *Inf. Sci.*, vol. 295, pp. 18–32, Feb. 2015, doi: 10.1016/j.ins.2014.10.015.

[13] A. Fronzetti Colladon and E. Remondi, "Using social network analysis to prevent money laundering," *Expert Syst. Appl.*, vol. 67, pp. 49–58, Jan. 2017.

[14] J. Lismont, E. Cardinaels, L. Bruynseels, S. De Groote, B. Baesens, W. Lemahieu, and J. Vanthienen, "Predicting tax avoidance by means of social network analytics," *Decis. Support Syst.*, vol. 108, pp. 13–24, Apr. 2018, doi: 10.1016/j.dss.2018.02.001.

[15] J. Zhang, L. Cheng, and H. Wang, "A multi-agent-based decision support system for bankruptcy contagion effects," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 5920–5934, Apr. 2012, doi: 10.1016/j.eswa.2011.11.112.

[16] A. L. Andrei, K. Comer, and M. Koehler, "An agent-based model of network effects on tax compliance and evasion," *J. Econ. Psychol.*, vol. 40, pp. 119–133, Feb. 2014.

[17] N. Hashimzade, G. D. Myles, F. Page, and M. D. Rablen, "Social networks and occupational choice: The endogenous formation of attitudes and beliefs about tax compliance," *J. Econ. Psychol.*, vol. 40, pp. 134–146, Feb. 2014.

[18] N. Hashimzade, G. D. Myles, and M. D. Rablen, "Predictive analytics and the targeting of audits," *J. Econ. Behav. Org.*, vol. 124, pp. 130–145, Apr. 2016.

[19] V. Van Vlasselaer, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, "GOTCHA! Network-based fraud detection for social security fraud," *Manage. Sci.*, vol. 63, no. 9, pp. 3090–3110, Sep. 2017.

[20] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *ACM SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, 2013.

[21] Council of the European Union. (May 2016). *9046/16. VAT Action Plan 'Towards a Single EU VAT Area*. [Online]. Available: http://data.consilium.europa.eu/doc/document/ST-9046-2016-INIT/en/pdf

[22] A. Beutel, L. Akoglu, and C. Faloutsos, "Fraud detection through graph-based user behavior modeling," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, I. Ray, N. Li, and C. Kruegel, Eds. Denver, CO, USA: ACM, Oct. 2015, pp. 1696–1697, doi: 10.1145/2810103.2812702.

[23] E. J. de Fortuny, M. Stankova, J. Moeyersoms, B. Minnaert, F. J. Provost, and D. Martens, "Corporate residence fraud detection," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, S. A. Macskassy, C. Perlich, J. Leskovec, W. Wang, and R. Ghani, Eds. New York, NY, USA: ACM, Aug. 2014, pp. 1650–1659, doi: 10.1145/2623330.2623333.

[24] K. Michalak and J. J. Korczak, "Graph mining approach to suspicious transaction detection," in *Proc. Federated Conf. Comput. Sci. Inf. Syst.*, M. Ganzha, L. A. Maciaszek, and M. Paprzycki, Eds. Szczecin, Poland: IEEE, Sep. 2011, pp. 69–75. [Online]. Available: http://ieeexplore.ieee.org/document/6078254/

[25] F. Tian, T. Lan, K.-M. Chao, N. Godwin, Q. Zheng, N. Shah, and F. Zhang, "Mining suspicious tax evasion groups in big data," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 10, pp. 2651–2664, Oct. 2016.

[26] A. Tselykh, M. Knyazeva, E. Popkova, A. Durfee, and A. Tselykh, "An attributed graph mining approach to detect transfer pricing fraud," in *Proc. 9th Int. Conf. Secur. Inf. Netw.*, Newark, NJ, USA, Jul. 2016, pp. 72–75, doi: 10.1145/2947626.2947655.

[27] D. Huang, D. Mu, L. Yang, and X. Cai, "CoDetect: Financial fraud detection with anomaly feature detection," *IEEE Access*, vol. 6, pp. 19161–19174, 2018, doi: 10.1109/access.2018.2816564.

[28] P. Das, A. K. Das, J. Nayak, D. Pelusi, and W. Ding, "A graph based clustering approach for relation extraction from crime data," *IEEE Access*, vol. 7, pp. 101269–101282, 2019, doi: 10.1109/access.2019.2929597.

[29] C. Sun, Z. Yan, Q. Li, Y. Zheng, X. Lu, and L. Cui, "Abnormal group-based joint medical fraud detection," *IEEE Access*, vol. 7, pp. 13589–13596, 2019, doi: 10.1109/access.2018.2887119.

[30] *Economic and Monetary Developments*, Eur. Central Bank, Frankfurt, Germany Jul. 2013.

[31] D. DeBarr and Z. Eyler-Walker, "Closing the gap: Automated screening of tax returns to identify egregious tax shelters," *SIGKDD Explor.*, vol. 8, no. 1, pp. 11–16, 2006, doi: 10.1145/1147234.1147237.

[32] K.-W. Hsu, N. Pathak, J. Srivastava, G. Tschida, and E. Bjorklund, "Data mining based tax audit selection: A case study of a pilot project at the minnesota department of revenue," in *Real World Data Mining Applications*. Cham, Switzerland: Springer, 2015, pp. 221–245.

[33] R.-S. Wu, C. Ou, H.-Y. Lin, S.-I. Chang, and D. C. Yen, "Using data mining technique to enhance tax evasion detection performance," *Expert Syst. Appl.*, vol. 39, no. 10, pp. 8769–8777, Aug. 2012.

[34] S. Basta, F. Fassetti, M. Guarascio, G. Manco, F. Giannotti, D. Pedreschi, L. Spinsanti, G. Papi, and S. Pisani, "High quality true-positive prediction for fiscal fraud detection," in *Proc. IEEE ICDM*, Miami, FL, USA, Y. Saygin, J. X. Yu, H. Kargupta, W. Wang, S. Ranka, P. S. Yu, and X. Wu, Eds. IEEE, Dec. 2009, pp. 7–12. [Online]. Available: https://ieeexplore.ieee.org/document/5360533

[35] T. Matos, J. A. F. de Macedo, and J. M. Monteiro, "An empirical method for discovering tax fraudsters: A real case study of Brazilian fiscal evasion," in *Proc. IDEAS*, B. C. Desai and M. Toyama, Eds. New York, NY, USA: ACM, 2015, pp. 41–48.

[36] Z. Assylbekov, I. Melnykov, R. Bekishev, A. Baltabayeva, D. Bissengaliyeva, and E. Mamlin, "Detecting value-added tax evasion by business entities of kazakhstan," in *Intelligent Decision Technologies*, I. Czarnowski, A. M. Caballero, R. J. Howlett, and L. C. Jain, Eds. Cham, Switzerland: Springer, 2016, pp. 37–49.

[37] X. Liu, D. Pan, and S. Chen, "Application of hierarchical clustering in tax inspection case-selecting," in *Proc. Int. Conf. Comput. Intell. Softw. Eng.*, Dec. 2010, pp. 1–4.

[38] E. Rahimikia, S. Mohammadi, T. Rahmani, and M. Ghazanfari, "Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran," *Int. J. Accounting Inf. Syst.*, vol. 25, pp. 1–17, May 2017, doi: 10.1016/j.accinf.2016.12.002.

[39] Y. J. Kim, B. Baik, and S. Cho, ''Detecting financial misstatements with fraud intention using multi-class cost-sensitive learning,'' *Expert Syst. Appl.*, vol. 62, pp. 32–43, Nov. 2016, doi: 10.1016/j.eswa.2016.06.016.

[40] H. Höglund, ''Tax payment default prediction using genetic algorithm-based variable selection,'' *Expert Syst. Appl.*, vol. 88, pp. 368–375, Dec. 2017, doi: 10.1016/j.eswa.2017.07.027.

[41] W. N. Dilla and R. L. Raschke, ''Data visualization for fraud detection: Practice implications and a call for future research,'' *Int. J. Accounting Inf. Syst.*, vol. 16, pp. 1–22, Mar. 2015.

[42] J. Kielman, J. Thomas, and R. May, ''Foundations and frontiers in visual analytics,'' *Inf. Vis.*, vol. 8, no. 4, pp. 239–246, Jan. 2009, doi: 10.1057/ivs.2009.25.

[43] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein, and J. Kuntner, ''Visual analytics for event detection: Focusing on fraud,'' *Vis. Inform.*, vol. 2, no. 4, pp. 198–212, Dec. 2018, doi: 10.1016/j.visinf.2018.11.001.

[44] E. Di Giacomo, W. Didimo, G. Liotta, and P. Palladino, ''Visual analysis of financial crimes: [System paper],'' in *Advanced Visual Interfaces*, Rome, Italy. New York, NY, USA: ACM, May 2010, pp. 393–394.

[45] R. A. Leite, T. Gschwandtner, S. Miksch, E. Gstrein, and J. Kuntner, ''Network analysis for financial fraud detection,'' in *EuroVis (Posters)*. Zürich, Switzerland: Eurographics Association, 2018, pp. 21–23.

[46] M. D. Flood, V. L. Lemieux, M. Varga, and B. L. W. Wong, ''The application of visual analytics to financial stability monitoring,'' *J. Financial Stability*, vol. 27, pp. 180–197, Dec. 2016, doi: 10.1016/j.jfs.2016.01.006.

[47] M. L. Huang, J. Liang, and Q. V. Nguyen, ''A visualization approach for frauds detection in financial market,'' in *Proc. 13th Int. Conf. Inf. Vis.*, Barcelona, Spain, E. Banissi, L. J. Stuart, T. G. Wyeld, M. Jern, G. L. Andrienko, N. Memon, R. Alhajj, R. A. Burkhard, G. G. Grinstein, D. P. Groth, A. Ursyn, J. Johansson, C. Forsell, U. Cvek, M. Trutschl, F. T. Marchese, C. Maple, A. J. Cowell, and A. V. Moere, Eds. IEEE, 2009, pp. 197–202.

[48] W. Didimo, L. Grilli, G. Liotta, F. Montecchiani, and D. Pagliuca, ''Visual querying and analysis of temporal fiscal networks,'' *Inf. Sci.*, vol. 505, pp. 406–421, Dec. 2019, doi: 10.1016/j.ins.2019.07.097.

[49] J. D. Kirkland, T. E. Senator, J. J. Hayden, T. Dybala, H. G. Goldberg, and P. Shyr, ''The nasd regulation advanced-detection system (ADS),'' *AI Mag.*, vol. 20, no. 1, pp. 55–67, 1999.

[50] R. A. Leite, T. Gschwandtner, S. Miksch, S. Kriglstein, M. Pohl, E. Gstrein, and J. Kuntner, ''EVA: Visual analytics to identify fraudulent events,'' *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 330–339, Jan. 2018, doi: 10.1109/tvcg.2017.2744758.

[51] M. Feng, J. Zheng, J. Ren, A. Hussain, X. Li, Y. Xi, and Q. Liu, ''Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data,'' *IEEE Access*, vol. 7, pp. 106111–106123, 2019, doi: 10.1109/access.2019.2930410.

[52] X. Zhu, J. Zhang, H. Li, P. Fournier-Viger, J. C.-W. Lin, and L. Chang, ''FRIOD: A deeply integrated feature-rich interactive system for effective and efficient outlier detection,'' *IEEE Access*, vol. 5, pp. 25682–25695, 2017, doi: 10.1109/access.2017.2771237.

[53] W. Didimo, F. Giacchè, and F. Montecchiani, ''Kojaph: Visual definition and exploration of patterns in graph databases,'' in *Graph Drawing and Network Visualization* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2015, pp. 272–278. [Online]. Available: https://link.springer.com/book/10.1007/978-3-319-27261-0#about

[54] R. Pienta, F. Hohman, A. Endert, A. Tamersoy, K. Roundy, C. Gates, S. Navathe, and D. H. Chau, ''VIGOR: Interactive visual exploration of graph query results,'' *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 215–225, Jan. 2018.

[55] J. Scott, ''Social network analysis,'' *Sociology*, vol. 22, no. 1, pp. 109–127, 1988.

[56] L. Page, S. Brin, R. Motwani, and T. Winograd, ''The pagerank citation ranking: Bringing order to the Web,'' Stanford InfoLab, Stanford, CA, USA, Tech. Rep. 1999-66, 1999.

[57] J. West and M. Bhattacharya, ''Intelligent financial fraud detection: A comprehensive review,'' *Comput. Secur.*, vol. 57, pp. 47–66, Mar. 2016, doi: 10.1016/j.cose.2015.09.005.

[58] J. H. Friedman, ''Greedy function approximation: A gradient boosting machine,'' *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.

[59] Y. K. Jain and S. K. Bhandare, ''Min max normalization based data perturbation method for privacy protection,'' *Int. J. Comput. Commun. Technol.*, vol. 2, no. 8, pp. 45–50, 2011.

[60] M. Ghoniem, J.-D. Fekete, and P. Castagliola, ''A comparison of the readability of graphs using node-link and matrix-based representations,'' in *Proc. IEEE Symp. Inf. Vis.*, Apr. 2005, pp. 17–24.

[61] N. Henry, J.-D. Fekete, and M. J. Mcguffin, ''NodeTrix: A hybrid visualization of social networks,'' *IEEE Trans. Vis. Comput. Graph.*, vol. 13, no. 6, pp. 1302–1309, Nov. 2007.

[62] V. Batagelj, F. J. Brandenburg, W. Didimo, G. Liotta, P. Palladino, and M. Patrignani, ''Visual analysis of large graphs using (X,Y)-clustering and hybrid visualizations,'' *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 11, pp. 1587–1598, Nov. 2011.

[63] L. Angori, W. Didimo, F. Montecchiani, D. Pagliuca, and A. Tappini, ''ChordLink: A new hybrid visualization model,'' in *Graph Drawing and Network Visualization* (Lecture Notes in Computer Science). Cham, Switzerland: Springer, 2019, pp. 276–290. [Online]. Available: https://link.springer.com/book/10.1007/978-3-030-35802-0#about

WALTER DIDIMO received the Ph.D. degree in computer engineering from the University of Rome ''La Sapienza,'' in 2000. He is currently an Associate Professor of computer engineering with the University of Perugia. His research interests include graph drawing, network visualization, algorithm engineering, as well as software design and experiments. He collected more than 150 international publications in the above areas. He is a member of the Steering Committee of the International Symposium on Graph Drawing and Network Visualization. He is an Associate Editor of IEEE Access journal.

LUCA GRILLI received the Ph.D. degree in information engineering from the University of Perugia, in 2007. He is currently an Assistant Professor of computer engineering with the University of Perugia. His research interests include graph drawing, information visualization, visual analytics, computational geometry, and very recently blockchain and DLTs technologies. In these research areas, he has collected about 40 publications and served as an external referee of international conferences and journals.

GIUSEPPE LIOTTA received the Ph.D. degree in computer science from the University of Rome ''La Sapienza,'' in 1995. He is currently a Professor with the Department of Engineering, University of Perugia. His research interests include information visualization, graph drawing, and computational geometry. On these topics, he has published more than 250 articles and gave invited lectures worldwide. He has served and chaired program committees of international symposiums and has served in the editorial board of international journals. His research has been founded by the Italian National Research Council, the Italian Ministry of Research and Education, the EU, and industrial sponsors.

LORENZO MENCONI received the degree in computer engineering from the University of Siena, in 2002, and the Ph.D. degree in cognitive science from the University of Siena, in 2012. He worked as a Software Designer and IT Project Manager at the Italian Court of Auditors and at the Italian Revenue Agency (Agenzia delle Entrate). He is currently working with the Istituto per la Vigilanza sulle Assicurazioni (IVASS). His research interests include machine learning, artificial intelligence, and data mining.

FABRIZIO MONTECCHIANI received the Ph.D. degree in information engineering from the University of Perugia, in 2014. He is currently an Assistant Professor of computer engineering with the University of Perugia. His research interests include graph drawing and graph algorithms, computational geometry, information visualization and visual analytics, as well as algorithm engineering for Big Data. He has collected more than 80 international publications in the above areas. He has been the Guest Editor of the *Journal of Graph Algorithms and Applications* and a member of the Program Committee of the International Symposium on Graph Drawing and Network Visualization.

DANIELE PAGLIUCA is currently pursuing the Ph.D. degree in information engineering with the University of Perugia. He is also a Public Officer of the Italian Revenue Agency, where he has gained experience as a Data Analyst, managing various data analysis projects for tax risk assessment. His research interests include decision support systems, visual analytics systems, graph databases, data mining, machine learning, and information visualization. He is an author of various national and international publications ranging from statistics to computer science.