Mafia Network: a Machine-Learning Approach

Adriano Amati[1], Monica Billio[1], Marco Di Cataldo[1,2]

[1] *Ca' Foscari University of Venice*
[2] *London School of Economics*

Abstract

This paper develops a novel early-warning system to predict firm confiscation due to mafia infiltration in Italy. To do this, we construct a unique, dynamic corporate ownership network by tracing the first- and second-degree connections of all Italian firms confiscated over the last two decades. Using this dataset, we train a series of machine learning models that integrate firm-level financial data with a comprehensive set of network topology measures, successfully forecasting confiscation up to three years in advance. Our out-of-sample results demonstrate exceptionally high predictive power, with the integrated model significantly outperforming those relying on a single data source. We find that network features are crucial for comprehensive detection (high recall), while firm financials provide accuracy (high precision); our framework effectively balances this critical trade-off. A feature-contribution analysis using SHAP values reveals that firm size is the single most important predictor, showing that smaller and financially distressed firms face the highest confiscation risk. Network topology features collectively emerge as the second most important group of indicators. Taken together, our findings underscore the value of integrating ownership network intelligence with firm fundamentals to create a powerful and interpretable tool for risk-based supervision and enforcement against organized crime.

JEL Classification: L1, K14, L25, C25, C38

Keywords: mafia, organized crime, legal economy, network, machine-learning, forecasting, firms, Italy.

Email addresses: adriano.amati@unive.it (Adriano Amati), billio@unive.it (Monica Billio), marco.dicataldo@unive.it (Marco Di Cataldo)

1. Introduction

Over the past few decades, organized crime groups (OCGs) have pushed well beyond their traditional spheres of illicit activity, making inroads into legitimate businesses and economies across the globe. A recent Europol report underscores this escalating trend, revealing that around 86% of the EU's most dangerous criminal networks infiltrate lawful enterprises to bolster, conceal, and enable a broad range of illegal operations, as well as to launder their proceeds (Europol, 2024). The scope and complexity of this phenomenon have increasingly drawn the attention of economic scholars. Researchers in the field are examining how mafia *infiltration*, or the presence of OCGs in legitimate markets, can alter economic incentives, undermine fair competition and produce distortions in the broader economic environment (Pinotti, 2015; Di Cataldo and Mastrorocco, 2022; Le Moglie and Sorrenti, 2022). In light of this, it is of vital importance to understand the strategic behaviour of criminals in forming such distortions (Mastrobuoni, 2020; Daniele and Dipoppa, 2023), as well as to investigate the extent to which anti-mafia legislation can be effective in tackling this issue (Calamunci and Drago, 2020; Chircop et al., 2023; Boeri et al., 2024).

While the infiltration of legal economies by OCGs is drawing increasing scholarly attention (Mirenda et al., 2022; Arellano-Bover et al., 2024; Le Moglie and Sorrenti, 2022), research has largely focused on ex-post analysis of the effects of OCG infiltration. This paper breaks new ground by developing an early-warning system that is the first to integrate dynamic corporate ownership networks for the purpose of predicting firm confiscation due to OCG infiltration. By applying machine learning to a unique dataset combining network topology with firm-level data, we create a framework that successfully forecasts confiscation up to three years in advance. Our analysis reveals the distinct predictive roles of network versus financial data and identifies the key firm characteristics associated with infiltration risk. Most importantly, we show the critical importance of integrating both feature sets to build a robust predictive model.

It is important to clarify our predictive target: confiscation, an unambiguous and judicially confirmed signal of OCG infiltration. The broader phenomenon of infiltration itself is latent, as authorities may not detect every compromised firm; thus, not all infiltrated firms are ultimately confiscated. By definition, however, a confiscation event serves as definitive confirmation that a firm was, in fact, infiltrated. Although this focus on judicially-verified outcomes restricts our analysis to a subset of all infiltrated firms,

it is a necessary trade-off. This approach provides the analytical certainty required to train a robust early-warning system on a ground truth of confirmed cases. To this end, a promising avenue for future research is the use of a Graph Neural Network (GNN) to learn latent representations for each firm in the ownership graph. This deep learning approach can automatically generate powerful node embeddings, i.e. latent representations, that capture complex, non-linear dynamic patterns of connectivity and economic performance, which could help uncover the subtle structural signatures that distinguish different types of firms at risk of confiscation. Building on this novel predictive framework, our paper makes three primary contributions.

Our first contribution is the creation of a novel, dynamic ownership network for Italy. To ensure our analysis is anchored on a ground truth free of false positives, we build our network starting from judicially-confirmed confiscated firms rather than those merely suspected of infiltration. From these seed firms, we reconstruct their corporate environment by linking them to their direct and second-order owners and subsidiaries –encompassing both firms (*Persone Giuridiche*) and individuals (*Persone Fisiche*). The result is a directed, weighted, and dynamic graph of corporate control, with edges representing ownership stakes timestamped at the daily level. After integrating this network information with comprehensive firm-level financials, we obtain a rich, attributed graph that serves as the foundation for our predictive modelling. The unique granularity of this dataset enables our core methodological shift: instead of an ex-post analysis of consequences, we leverage the network's evolution to test if the changing topology of corporate ownership is a powerful leading indicator of confiscation and, consequently, infiltration risk.

Our second key contribution stems from our predictive modelling, which sheds light on crime infiltration with practical implications for law enforcement authorities. Using a supervised machine learning framework, we develop an early-warning system designed to predict the likelihood of a firm's future confiscation due to mafia infiltration. Specifically, our analysis reveals a stark precision-recall trade-off between the different feature sets. Models trained exclusively on firm-level financial data tend to be highly precise; that is, when they flag a firm as high-risk, they are very likely to be correct, i.e. up to ≈ 0.99. However, their critical weakness is low recall, ≈ 0.67, meaning they fail to identify a large portion of the firms that are actually infiltrated, allowing many potential targets to go undetected. On the other hand, models relying solely on network topology data are comprehensive, achieving high recall by flagging a broad set of suspicious companies,

i.e. up to ≈ 0.94. The significant drawback is lower precision, i.e. ≈ 0.72, generating a substantial number of false positives and consequently making it less accurate. Crucially, our integrated model, which combines both data sources, successfully resolves this operational trade-off. It achieves a powerful and effective balance, maintaining the high precision needed to avoid false alarms while retaining the high recall necessary to ensure comprehensive detection of potential targets, i.e. an F1-score of up to ≈ 0.91 with precision ≈ 0.97 and recall ≈ 0.86. From a law enforcement standpoint, this balanced outcome is vital, as it provides a practical tool that allows for the efficient allocation of scarce investigative resources to a list of the highest-risk firms that is both accurate and complete. This adds a significant contribution to the literature on the *efficiency-security* trade-off, in the context of criminal networks and anti-mafia enforcement (Morselli, 2008; Lu et al., 2025).

Third, we run a feature-contribution analysis, where we complement the model's performance results by identifying the specific characteristics linked to a higher risk of confiscation. To achieve this, we employ SHapley Additive exPlanations, or SHAP, values, a method that allows us to precisely quantify the impact of each feature on the model's output for any given firm (Lundberg et al., 2020). This granular analysis reveals a clear hierarchy of predictive factors. The findings show that firm size, measured by the number of employees, is the single most influential predictor. The relationship is unambiguous: smaller and financially distressed firms consistently face the highest confiscation risk. Following size, financial indicators of distress –such as low profitability and poor liquidity –emerge as another crucial set of predictors. This empirical finding aligns with previous research, which posits that smaller, financially vulnerable firms are not only prime targets for criminal infiltration but are also more likely to be discovered by authorities, as they are often directly involved in illicit activities (Le Moglie and Sorrenti, 2022; Arellano-Bover et al., 2024). We also find that the models learn complex, non-linear patterns from network topology that defy simple interpretation, suggesting a promising area for future research. Taken together, our findings highlight the value of combining ownership-network intelligence with firm fundamentals to produce timely and interpretable early-warning signals, offering a practical basis for risk-based supervision and prioritized audits.

Related Literature A growing body of research examines how OCGs embed themselves within the legal economy and reshape incentives, firm dynamics, and public resource allocation. Many recent studies attempt to trace the firm-level footprints of

infiltration. Using administrative and balance-sheet data, Mirenda et al. (2022) show that targeted firms exhibit anomalous revenue growth unaccompanied by proportional input expansion, followed by deteriorating fundamentals—patterns consistent with rent extraction and money laundering rather than productivity gains. In a broader perspective, Pinotti (2015) document sizeable aggregate costs of organized crime in Southern Italy, while Le Moglie and Sorrenti (2022) suggest that OCG entry tends to occur where firms are smaller and financially weaker, pointing to systematic selection into vulnerability. Other research focuses on the strategic motives behind infiltration, modelling criminal infiltration as an investment decision of OCGs (Arellano-Bover et al., 2024). The study posits that OCGs may establish small-sized firms for criminal activities, infiltrate medium-sized firms post creation, while they infiltrate large firms for clean financial returns and political connections (Arellano-Bover et al., 2024).

A complementary literature uses networks to understand how criminal organizations penetrate and exploit corporate structures. Research on enforcement and asset recovery documents how interventions propagate through economic networks: Calamunci and Drago (2020) study real-time firm outcomes under court-appointed management of confiscated firms. De Luca et al. (2025) looks at changes in firms' ownership structure during the Covid-19 pandemic, finding that prolonged state-mandated business closures restrict changes in ownership, but the effect is weaker in areas with strong mafia presence. The authors interpret this as OCGs infiltrating firms in financial distress (De Luca et al., 2025). Although our aim is fundamentally different from these studies, we expand on their approach by building a time-dynamic network of infiltrated and non-infiltrated firms, in order to systematically test whether the evolving structure of corporate ownership can serve as a leading indicator of infiltration risk.

Since crime is fundamentally a social phenomenon, economic theory has increasingly turned to social network analysis as a powerful framework for understanding and combating it (Lindquist and Zenou, 2019; Ballester et al., 2004). The very architecture of a criminal network is therefore theoretically important, as it can either amplify or dampen criminal activity. This approach allows us to understand the structural roles of nodes –criminals in this case, and help to further characterize and understand criminal activity (Lindquist and Zenou, 2019; Ballester et al., 2004). Moreover, from a statistical modeling perspective, other research frames criminal network structure as the outcome of a fundamental efficiency-security trade-off. This view posits that organizations must create enough connections to operate efficiently while remaining sparse enough to ensure

security and avoid detection (Lu et al., 2025). This tension shapes their architecture, leading to unique features such as groups of redundant criminals and a mix of overt and deliberately covert ties (Lu et al., 2025). Our study builds on this foundation by shifting the focus from the internal social dynamics of criminal-only networks to the economic interactions of OCGs within the legal economy. We consider a mixed environment where criminal actors and legitimate firms interact, allowing us to identify the patterns that signal a firm's risk of confiscation—whether it was originally created by an OCG or was a legitimate enterprise later preyed upon.

Recent advances in predictive modelling underscore the feasibility of anticipating illicit behaviour with administrative data. For example, Ash et al. (2025) develop a machine-learning framework to forecast corruption in Brazilian municipalities, while Mastrobuoni (2020) show how digital information systems enhance police productivity. Applying these predictive techniques to OCG infiltration, Cariello et al. (2024) use a machine learning approach on a unique dataset of Italian firms to detect potentially infiltrated companies. These results collectively motivate our approach, which also focuses on ex-ante risk assessment by forecasting confiscation risk. Our study also expands on this prior work by integrating not only firm fundamentals but also a rich set of network topology measures. While our firm-level only model achieves performance comparable to that reported by Cariello et al. (2024), our integrated model with network features demonstrates substantially superior predictive power, highlighting the significant value added by incorporating network topology.

The existing literature provides a strong foundation by identifying the economic footprints of infiltrated firms, exploring the network effects of anti-mafia interventions, and demonstrating the power of machine learning for forecasting illicit activities. However, these streams of research have largely evolved in parallel. A significant gap remains in leveraging the dynamic evolution of corporate ownership networks as a primary tool for ex-ante risk assessment. Our study bridges this divide by uniting these three approaches. Our network approach systematically tests whether changing topological features can serve as a leading indicator of confiscation. In doing so, we shift the analytical focus from identifying the consequences of infiltration to building a predictive framework that can anticipate it, providing a new lens through which to view and combat organized crime's encroachment into the legal economy.

2. Data

This section will briefly describe all the data sources and how they have been used for the study.

Confiscated Firms Data Data on confiscated firms has been provided by the National Agency for Seized and Confiscated Mafia Assets *Agenzia Nazionale per i beni sequestrati e confiscati alla criminalità organizzata* (ANBSC) which reports detailed data on all firms ever confiscated for illicit activities in relation to OCG. This dataset comprises over 6,200 firms confiscated between 2000 to 2024 and includes key variables such as the percentage of confiscated assets, the date of confiscation, the economic sector in which the firm operates, and precise geo-referenced data, including the firm's address. Furthermore, the data documents the complete judicial history of the confiscation process, from the initial seizure of assets by authorities –*sequestro*, to the final legal outcome, be it the confirmation of the confiscation through multiple judicial instances or its eventual revocation. Crucially, in contrast to prior research (e.g., Chircop et al., 2023), our dataset provides an unprecedentedly complete picture of this phenomenon at a national scale, with the number of confiscated firms being an order of magnitude larger.

For the remainder of this paper, the confiscation date refers to the initial judicial seizure of proprietary assets by authorities, which marks the beginning of legal proceedings against a firm suspected of mafia involvement. Figure A.1 in the Appendix displays the number of confiscations over time. We can observe an increase in confiscations in the years 2009 to 2017 and a sharp decline in later years. This is probably due to the lack of processing of more recent events by ANBSC. Figure A.2 displays the spatial distribution of confiscated firms at the municipality-level activity. We can immediately see that there is a high concentration of confiscated entities in Rome and Milan, followed by Naples and Palermo. Overall, there seems to be a higher concentration of confiscated firms in the South of Italy, which is consistent with past literature on mafia presence in Italy (Arellano-Bover et al., 2024; Pinotti, 2015).

Ownership Network We also collected firm-level data from *InfoCamere*, the Italian company responsible for managing the official business registry. To analyze the economic environment of the confiscated firms, we reconstructed their ownership networks. The network nodes represent either firms (*Persone Giuridiche*, PG) or individuals (*Persone*

Fisiche, PF), and the directed edges represent ownership stakes, timestamped at the day level. We mapped the network surrounding each confiscated firm up to two degrees of separation considering *all* ownership stakes[1] to define a connection. The reconstruction followed a two-step process:

1. **First-Degree Connections**: For each confiscated firm, we first identified its immediate connections. We compiled a list of all its direct shareholders (those who owned it) and all of its direct subsidiaries (those it owned) over a 20-year period, covering the 10 years before and after the confiscation.
2. **Second-Degree Connections**: For each upstream shareholder, we included all other firms in their portfolio, ensuring these connections fell within the 20-year time window of the original confiscated firm. Similarly, for each downstream subsidiary, we included all of its owners and any other firms it controlled, again respecting the same temporal window. In cases where a second-degree entity was connected to multiple confiscated firms, we retained only the link associated with the earliest confiscation date to provide a unique temporal anchor for our analysis.

This methodology captures a detailed view of the local ownership structure, revealing both the investors and assets linked to the confiscated entities. The resulting network is a weighted–i.e., ownership percentage–, directed–i.e., ownership direction–heterogeneous[2]–dynamic graph, spanning from 1993 to 2024.[3]

This network-based approach represents a fundamentally different way to measure the degree of OCG involvement in the legal economy. Rather than attempting to identify potentially infiltrated firms from the entire population (Mirenda et al., 2022; Arellano-Bover et al., 2024), our methodology begins with a solid core of firms whose involvement in criminal activity is judicially confirmed through confiscation. From this reliable anchor of definitively infiltrated entities, we then expand outward, building a targeted network of all connected firms and individuals up to two degrees of separation. The resulting dataset provides a unique view into the immediate economic environment surrounding confirmed OCG activity, enabling us to analyze the firms within this high-risk ecosystem

[1] i.e. $> 0\%$

[2] as it includes both people and firms. In this context people can only be *source* nodes, i.e. they cannot be owned by any other node

[3] Ideally, this network would be benchmarked against a random sample of the Italian firm population ownership network. However, access to the comprehensive ownership data required to construct such a baseline is presently restricted.

to find predictive signals of infiltration.

Firm-Level Data From *InfoCamere*, we also collect comprehensive details on company registrations, firm characteristics –i.e. number of employees, municipality, industry, and so forth –economic and financial performance as well as balance-sheet data at the yearly level. The two main sources are the business registry (*registro*), with quarterly data covering years from 2001 to 2024, and firm-level company financials (*bilancio*), with yearly data covering the period 2005 to 2024. By integrating the ANBSC confiscation records with the Infocamere ownership and firm-level datasets, we construct a rich, attributed network of yearly snapshots. Figure A.3 displays the distribution across sectors of our sample of firms. We can observe that most confiscated entities operate in construction, real estate, wholesale and manufacturing, with fewer confiscations occurring in the remaining sectors. Table A.2 in the Appendix provides a detailed list of all firm-level variables used in our analysis, categorized into firm characteristics, financial indicators, balance sheet assets, balance sheet liabilities and equity, and income statement items.

The Network By integrating the confiscation records, ownership data, and firm-level financials, we construct the final, rich, attributed network used in our analysis. To explore this complex dataset, we developed an interactive 3D rendering of the graph.[4] A key feature of this tool is its dynamic capability; a time slider allows for navigation through yearly snapshots, making it possible to visualize how ownership structures evolve over time. In addition to this dynamic view, the rendering can display the complete static network, which aggregates all connections across the entire period. The static version of the graph counts 40,425 edges and 36,323 nodes, of which 11,202 PFs and 25,121 PGs. Figure A.4a displays the static version of the graph, Figure A.4b displays the static graph's main component and Table 1 displays the static and yearly networks' descriptive statistics. In both Figures A.4a and A.4b, as well as in Table 1, we can observe that there is a giant component accounting for 60% of the nodes and 70% of the edges in the static version of the network. Interestingly, this feature disappears when we look at the yearly snapshots, where the largest component accounts for only 10-15% of the nodes and edges. This stark difference between the long-term aggregate and short-term snapshots may hint at strategic infiltration patterns, such as the use of temporary

[4]More on this on Section B in the Appendix

or rapidly changing ownership structures to evade detection. While a full analysis of this network dynamic is beyond the scope of this paper, it represents a compelling area for future research.

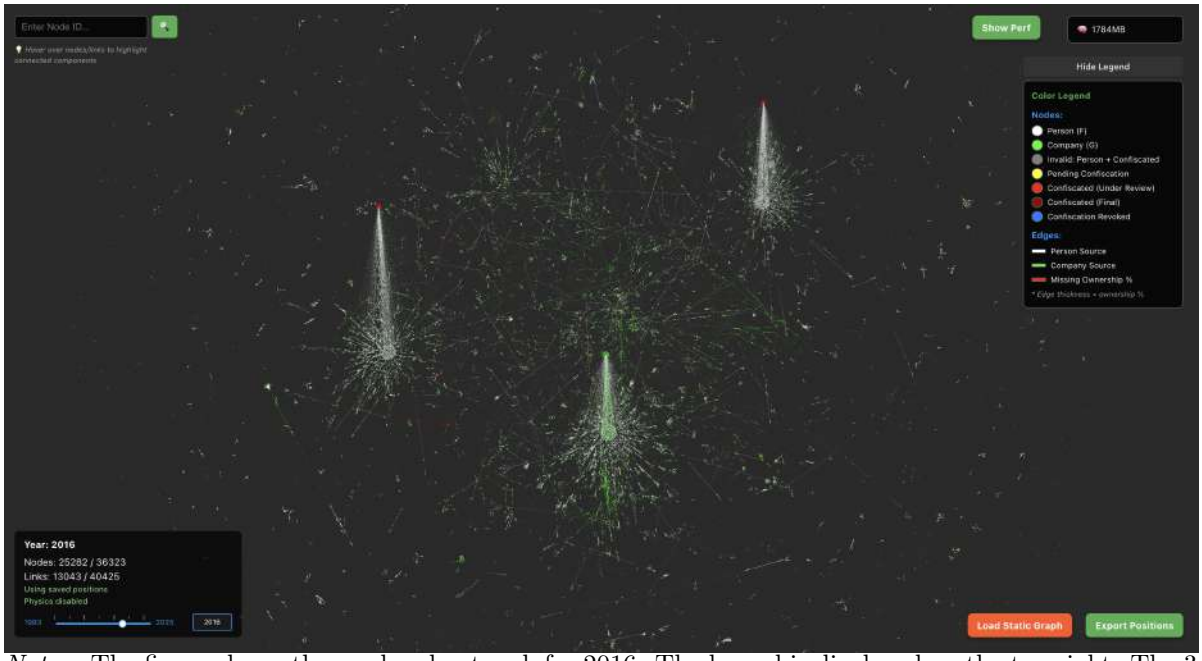
Figures 1 and A.5 provide some visualizations of the network's yearly snapshots. Figure 1 illustrates the network's structure in 2016, which is dominated by three prominent hubs. These hubs are all part of a single, large, weakly connected component that, as detailed in the Appendix (Figure A.5d), encompasses nearly 6,000 nodes, just under 7,000 edges, and over 150 confiscated firms. A closer look at the hubs reveals a striking contrast among their most central nodes: two are confiscated micro-firms–i.e. less than 10 employees (indicated by red nodes), while the third is anchored by a large, non-confiscated financial institution (a green node). Panels A.5a to A.5c display other recurrent network structures. Panel A.5a shows a star-like structure centered on a confiscated firm, with many direct connections to both individuals and other firms. Here edges are red to indicate that information on the ownership percentage is missing. Panel A.5b shows a more complex structure, with multiple interconnected firms and individuals, suggesting a web of relationships that could facilitate illicit activities. Panel A.5c displays a dense cluster of non-confiscated firms with a person (white node) at the center, indicating a potentially influential individual within this network.

It is also important to note that this final dataset yields a non-random network centered on confiscated firms, which presents a distinct challenge for our classification task. One might expect mafia-related firms to have a unique network topology that simplifies identification, although a descriptive analysis suggests this is not the case. As shown in Figures A.6, A.7, A.8 and Table A.1 the distributions of node degrees and centralities show no clear separation across firm types and PFs. It could also be that this sample might actually cluster together similar firms, making the classification task much harder. Still, a definitive assessment would require a more comprehensive graph that also maps the ownership structures of a random sample of firms with no known links to criminal activity.

3. Methodology

Our goal is to build a model that estimates the probability, p_i, that a firm i will be confiscated because of OCG infiltration. This prediction is based on the firm's historical data, including its financial fundamentals and network signals, which we collect into a

Figure 1: Graph Snapshot for 2016



Notes: The figure shows the rendered network for 2016. The legend is displayed on the top right. The 3 most central nodes have been highlighted. Note that each of these hub nodes belong to the same weakly connected component, see A.5d

feature set, $\mathbf{X}_i$. The machine learning task is to train a function, f, that takes the firm's data as input and outputs an estimated confiscation risk:

$$p_i = f(\mathbf{X_i}).$$

We then test this model's early-warning capabilities by evaluating its performance at different forecast horizons, from the year of the event ($h = 0$) up to three years in advance ($h = 3$). This forecasting window is particularly relevant from a policy perspective, as the large majority of judicial inquiries begin one to two years before the initial asset seizure, i.e. *sequestro*, making our model a potentially valuable tool for early-stage investigations.[5]

3.1. Data pre-processing

We assemble a firm-year panel from the dynamic ownership graphs' node attributes and we expand it with network information. Centrality measures and neighbourhood

[5]This typical one-to-two-year timeline for inquiries preceding a *sequestro* was confirmed in personal communication with ANBSC officials and data.

Table 1: Descriptive statistics - Directed Mafia Network

Metric	1993	1998	2003	2008	2013	2015	2017	2019	2021	2023	Whole
Network characteristics:											
Nodes	7,957	11,384	17,819	25,411	25,577	25,366	24,726	23,097	21,340	19,218	36,323
Links	326	3,860	9,271	18,341	13,485	12,686	11,967	9,159	6,988	4,742	40,425
Average Degree	0.041	0.339	0.520	0.722	0.527	0.500	0.484	0.397	0.327	0.247	1.113
Companies (G nodes)	7,757	9,978	14,057	18,189	19,370	19,218	18,881	18,262	17,394	16,197	25,121
Persons (F nodes)	200	1,406	3,762	7,222	6,207	6,148	5,845	4,835	3,946	3,021	11,202
Prop. Companies	0.975	0.876	0.789	0.716	0.757	0.758	0.764	0.791	0.815	0.843	0.692
Confiscated Companies	3,353	3,712	4,380	5,346	5,914	5,957	5,867	5,791	5,676	5,416	6,284
Prop. Confiscated	0.432	0.372	0.312	0.294	0.305	0.310	0.311	0.317	0.326	0.334	0.250
Avg. Confiscation Quota	92.167	91.823	90.909	90.093	90.173	90.246	90.316	90.473	90.566	90.677	90.357
N. of components	7,672	8,502	10,254	10,000	13,696	14,045	14,047	14,749	14,903	14,780	4,513
	(7631)	(7517)	(8544)	(7779)	(12074)	(12694)	(12762)	(13930)	(14355)	(14472)	(4576)
Average Clustering Coefficient	0.000	0.002	0.002	0.004	0.003	0.002	0.002	0.002	0.001	0.001	0.009
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Overall Clustering Coefficient	0.014	0.003	0.003	0.003	0.005	0.004	0.005	0.005	0.004	0.005	0.004
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Largest Component characteristics:											
Nodes	18	123	825	4,472	2,615	4,100	5,610	3,340	2,732	2,267	22,071
Links	26	385	1,388	5,728	3,070	4,708	6,442	3,759	3,018	2,449	28,948
Average Degree	1.444	3.130	1.682	1.281	1.174	1.148	1.148	1.125	1.105	1.080	1.312
Companies (G nodes)	13	106	654	2,811	1,486	1,946	2,772	1,334	958	609	15,222
Persons (F nodes)	5	17	171	1,661	1,129	2,154	2,838	2,006	1,774	1,658	6,849
Prop. Companies	0.722	0.862	0.793	0.629	0.568	0.475	0.494	0.399	0.351	0.269	0.690
Confiscated Companies	10	1	42	166	77	87	146	113	124	110	1,013
Prop. Confiscated	0.769	0.009	0.064	0.059	0.052	0.045	0.053	0.085	0.129	0.181	0.067
Fraction of Nodes	0.002	0.011	0.046	0.176	0.102	0.162	0.227	0.145	0.128	0.118	0.608
	(0.001)	(0.003)	(0.068)	(0.544)	(0.092)	(0.030)	(0.019)	(0.003)	(0.002)	(0.001)	(0.849)
Average Clustering Coefficient	0.046	0.000	0.011	0.008	0.007	0.006	0.006	0.004	0.002	0.002	0.011
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Overall Clustering Coefficient	0.100	0.000	0.001	0.003	0.006	0.006	0.005	0.005	0.004	0.004	0.004
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

Notes: The table reports network characteristics for the mafia network for selected years. Number in parenthesis are from simulated Erdős-Renyi random networks with the same number of nodes, links and density as the observed network. Time intervals have been chosen to be eavery 5 years for older snapshots and every two years for more recent ones. The last column refers to the whole network, i.e. the one including all links and nodes from 1993 to 2023.

counts [6] are computed on yearly directed and undirected graphs—both unweighted and weighted by the ownership stake–and merged back to the nodes and construct a firm-level count of adjacent individuals (PFs). Also, we remove PF nodes, as cannot be confiscated entities.

Because our goal is to forecast confiscation either in the year it occurs (t^c) or in a

[6] i.e. number of direct PFs connections

specified number of years prior ($t^c - h$), every observation must be aligned to a *reference year* representing the forecast horizon. For confiscated firms, this reference year is observed directly. For non-confiscated firms, no such date exists, so we assign a synthetic reference year via propensity-score matching. The reason for this is to make the comparison as honest as possible and prevent the model from learning a spurious temporal rule. If the reference years for all non-confiscated firms were clustered at the end of the sample period, the model could make artificially easy predictions based on the time-period alone, rather than by identifying the underlying financial and network signals present in the years prior.

Specifically, we estimate a scalar proximity score from observable firm characteristics and select the nearest confiscated "twin" for each non-confiscated firm. The matching variables include industry sector (NACE 2-digit level), firm size and geographic location. The choice of these covariates is motivated by their relative stability over time. Unlike financial performance metrics which can be highly volatile, a firm's industry, location, and general size category tend to change infrequently, making them robust variables for identifying a structurally comparable "twin". The reference year for the non-confiscated firm is then set equal to that of its matched confiscated twin. It is important to stress that this matching procedure is in no way related to recovering treatment effects of being confiscated, it is just needed to obtain a synthetic confiscation year for non-confiscated firms. We also consider an alternative methodology, where we sample confiscation dates for non-confiscated firms from the same empirical distribution function of observed confiscation years. Figure A.9 displays the result of both methodologies.

Then, for a given forecast horizon h and any firm i, we set

$$t^c_{i,h} = t^c_i - h,$$

drop observations without sufficient history, and keep only years $t \leq t^c_{i,h}$. This produces a longitudinal history for every firm that ends at its (shifted) reference year. This enables us to build a dataset that includes both firm-level data[7] and network measures[8] for each firm, and their lags. We linearly impute missing values and standardize all numerical features. For the categorical variables we consider two lags (i.e. $K = 2$) and

[7] All variables listed on Table A.2.

[8] All network measures listed on Table A.1 and the count of PFs direct neighbours.

for continuous variables, five lags (i.e. $L = 5$). So for each firm i we obtain the following row-vector of features to feed to the classifier:

$$\mathbf{X}_{i,h} = \Big[\underbrace{\mathbf{x}_{i,t_h^c-1}, \ldots, \mathbf{x}_{i,t_h^c-L}, \mathbf{c}_{i,t_h^c-1}, \ldots, \mathbf{c}_{i,t_h^c-K}}_{\text{Firm level data}}, \underbrace{\text{PN}_{i,t_h^c-1}, \ldots, \text{PN}_{i,t_h^c-L},\ \text{NC}_{i,t_h^c-1}, \ldots, \text{NC}_{i,t_h^c-L}}_{\text{Network data}} \Big],$$

where $\mathbf{x}_{i,t}$ is the vector of firm-level financial variables at time t, $\mathbf{c}_{i,t}$ is the vector of categorical firm-level variables at time t, $\text{PN}_{i,t}$ is the number of PFs connected to firm i at time t and $\text{NC}_{i,t}$ are the network centrality measures of firm i at time t.

Finally, we label each observation with a binary target variable, $y_{i,h}$, which equals 1 if the firm is confiscated in year $t_{i,h}^c$ and 0 otherwise. This results in a balanced panel dataset with one firm per row, where each row contains the historical features up to the reference year and the corresponding confiscation label. In order to understand the importance of network features, we run the same classification task with three different feature sets: *Firm-level* data only, *Network* data only, *Firm-level and Network* data. The resulting matrices, with number of observations and average confiscation rate is displayed on Table 2. As expected, there is a natural decline in the number of observations as the forecast horizon increases, as well as a mild class imbalance –i.e. confiscated firms are a minority of the total sample –that becomes more pronounced at longer horizons. The next subsection will describe how we address this imbalance in our classification task.

Table 2: Descriptive Statistics by Forecast Horizon

	$h = 0$	$h = 1$	$h = 2$	$h = 3$
Observations	17,967	17,484	16,767	16,204
Mean Confiscation	0.2739	0.2697	0.2562	0.2426

Notes: The table reports descriptive statistics for the classification dataset across different forecast horizons. $h = 0$ corresponds to the contemporaneous dataset, i.e. $t_h^c = t^c$, while $h > 0$ indicates the number of years ahead for the prediction task. The number of observations decreases for higher forecast horizons due to data availability constraints.

3.2. Classification

To build our predictive model, we employ a supervised machine learning framework designed to distinguish between confiscated and non-confiscated firms. Similarly to Ash

et al. (2025), we test four distinct classification algorithms: Logistic Regression, Random Forest (Breiman, 2001) and two gradient boosting algorithms, LightGBM (Ke et al., 2017) and XGBoost (Chen and Guestrin, 2016). These models were chosen to span a range of complexity and flexibility. We start with Logistic Regression, which assumes a linear relationship between the predictors and the outcome, providing a strong baseline.[9] We then evaluate more powerful tree-based ensemble models—Random Forest, LightGBM, and XGBoost—which relax this linearity assumption and are adept at capturing complex, non-linear relationships and interactions within the data[10]. The performance of these four models is systematically compared across the three distinct feature sets for each different forecast horizon.

Our experimental design follows a rigorous pipeline for data preparation, training, and evaluation. Since confiscations occur in different points in time, arranging the data in a random train/test split could bias the results, i.e. temporal *data leakage*.[11] To address this, we employ a chronological split, training our models on all firm-year observations up to a specific cut-off date. We chose 2018 as cutoff date, to provide sufficient training examples to the algorithms –i.e. $t_i^c \leq 2018 - h$, but we also consider 2016 as a cutoff date as a robustness check. We then evaluate the models' performance exclusively on observations from subsequent years. The train-test split is displayed on Table A.9. Moreover, to address the inherent class imbalance—where non-confiscated firms outnumber confiscated ones—we incorporate the Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al., 2002). Applied only to the training data, SMOTE generates new synthetic examples of the minority (confiscated) class, creating a balanced dataset that prevents the classifiers from developing a bias towards the majority class.

The final stage involves training the models and evaluating their out-of-sample performance. For each combination of model and feature set, we perform hyperparameter tuning using a randomized search with 5-fold cross-validation and 10 random draws of the parameter space. This process systematically explores different model configura-

[9] To be precise, while the sigmoid function means the model's probability output is a non-linear function of the features, Logistic Regression is considered a linear classifier because its decision boundary is linear. That is, it separates the classes in the feature space using a hyperplane (e.g., a line in two dimensions).

[10] see Section C for a more in-depth discussion.

[11] Although there might not be data leakage of the confiscation of firm i on the data concerning firm j, the aim of this classification exercise is to make the problem as close as possible as the one that law enforcement authorities have to deal with.

tions to identify the one that maximizes the ROC-AUC, a key metric that evaluates a model's ability to discriminate between positive and negative classes across all possible classification thresholds. Once the best-performing model is identified for each scenario, its predictive power is conclusively assessed on the unseen test set. We report a comprehensive suite of metrics: Accuracy (the overall correct predictions), Precision (the proportion of correctly identified confiscated firms among all firms flagged as such), Recall (the proportion of actual confiscated firms that were correctly identified), the F1-score (the harmonic mean of Precision and Recall), and the ROC-AUC score. This entire procedure is independently repeated for each forecast horizon, from $h = 0$ to $h = 3$, allowing us to evaluate the model's early-warning capabilities at predicting confiscation up to three years in advance.

4. Results

This section will evaluate the model performance across the three feature sets and then will provide a more detailed breakdown of the feature contributions for predictions.

4.1. Main Results

Table 3 reports the results for forecasting confiscation in the year it occurs ($h = 0$), while Table 4 evaluates the models' early-warning capability one year in advance ($h = 1$). The findings are presented across three distinct feature sets—Firm-Level Only, Network Only, and the combined Firm + Network—to isolate the predictive contribution of each data source. As we evaluate these results, it is crucial to remember that our model predicts *confiscation*, not *infiltration* directly. While a confiscation event provides a certain label for an infiltrated firm, the absence of confiscation does not guarantee the absence of infiltration. This implies that some of the model's false positives—firms predicted to be at high risk that are not in our confiscated sample—could represent successful identifications of firms that are truly infiltrated but remain undetected by law enforcement. Consequently, the measured accuracy of a model displaying high recall may be biased downwards.

A primary finding, consistent across all models and time horizons, is the substantial performance gain achieved by combining firm-level and network features (Panel C in both tables). This combined effect is evident across all metrics. For instance, in the contemporaneous prediction task (Table 3), the XGBoost model's AUC score increases

Table 3: Model Performance Comparison Across Feature Sets, $h = 0$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.830	0.809	0.799	0.793
Precision	0.950	**0.991**	**0.985**	**0.979**
Recall	0.742	0.674	0.660	0.652
F1	0.833	0.802	0.791	0.783
AUC	0.893	0.902	0.908	0.904
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.539	0.765	0.758	0.761
Precision	0.795	0.725	0.723	0.723
Recall	0.266	**0.949**	**0.938**	**0.947**
F1	0.398	0.822	0.816	0.820
AUC	0.632	0.841	0.858	0.796
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.871**	**0.905**	**0.888**	**0.904**
Precision	**0.958**	0.976	0.935	0.971
Recall	**0.811**	0.855	0.865	0.857
F1	**0.878**	**0.912**	**0.899**	**0.911**
AUC	**0.958**	**0.962**	**0.960**	**0.975**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table 4: Model Performance Comparison Across Feature Sets, $h = 1$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.796	0.806	0.796	0.795
Precision	0.924	**0.988**	**0.968**	**0.982**
Recall	0.695	0.663	0.659	0.648
F1	0.793	0.794	0.785	0.781
AUC	0.890	0.900	0.896	0.890
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.543	0.759	0.765	0.769
Precision	0.786	0.723	0.715	0.721
Recall	0.258	**0.928**	**0.971**	**0.961**
F1	0.389	0.812	0.823	0.824
AUC	0.628	0.835	0.721	0.718
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.873**	**0.889**	**0.893**	**0.901**
Precision	**0.967**	0.977	0.898	0.943
Recall	**0.802**	0.822	0.914	0.877
F1	**0.877**	**0.893**	**0.906**	**0.909**
AUC	**0.966**	**0.961**	**0.963**	**0.963**

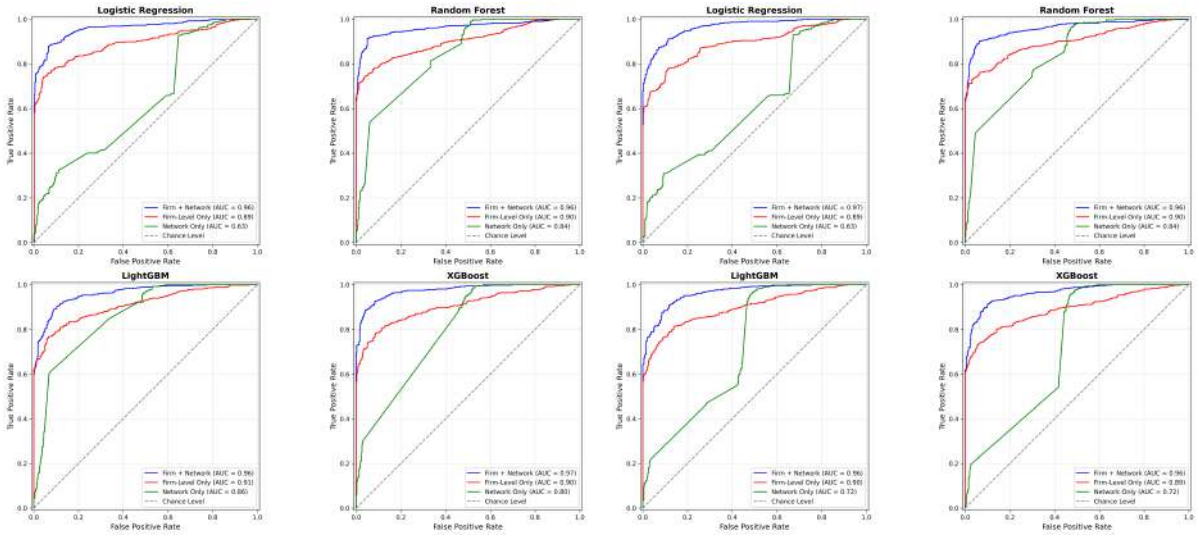
Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

from 0.904 (firm-level only) and 0.796 (network only) to an exceptional 0.975 when both feature sets are used. Similarly, the F1-score, which balances precision and recall, jumps from 0.783 to 0.911, indicating a far more reliable classification. In the appendix Tables A.3 to A.4 show the performance results on the test-set for longer forecast horizons, where we can observe a similar pattern as in Table 3 and 4. Most importantly, model performance is similar for different forecast horizons, signalling that the model is robust over time and effective as an early-warning system, capable of identifying at-risk firms several years in advance. Except for a very low recall score, the Firm-Level Only model achieves performance levels that are comparable to, or worse than, those reported in Cariello et al. (2024), suggesting that our non-random sample of Italian firms is suitable for meaningful comparison with earlier work. This is not the case for the Firm and Network model, which consistently outperforms the results in Cariello et al. (2024), highlighting the significant value added by incorporating network topology.

While the combined feature set is clearly superior, a comparison of the models using isolated feature sets reveals their complementary predictive roles. Models trained exclusively on firm-level data (Panel A) consistently achieve higher precision, suggesting that financial indicators, when triggered, are strong signals of risk. Conversely, models relying solely on network data (Panel B) demonstrate markedly higher recall. For example, at $h = 1$, the Random Forest, LightGBM, and XGBoost models achieve recall scores of 0.928, 0.971, and 0.961 respectively with network data, compared to scores of 0.663, 0.659, and 0.648 with firm-level data. This suggests that ownership network topology is exceptionally effective at flagging the broader set of potentially infiltrated firms. This pattern is also evident in Tables A.3 to A.4, where recall is highest using only network data.

These divergent outcomes highlight the trade-off with significant strategic implications for law enforcement. A model relying solely on financial indicators, while precise, operates with lower recall, risking that many infiltrated firms go undetected. Conversely, a network-only approach, despite its excellent recall, would generate a high volume of false positives, hampering the efficient allocation of investigative resources. The superior performance of the combined model demonstrates its ability to navigate this trade-off effectively. By integrating both data sources, our framework achieves an efficient balance, maintaining high recall without unduly sacrificing precision. This approach is crucial for operational effectiveness, as it enables authorities to cast a wide yet targeted net, ensuring that potential threats are not overlooked while simultaneously optimizing the

Figure 2: Model Performance ROC Curves by Forecast Horizon



(a) Contemporaneous Prediction ($h = 0$)

(b) One-Year-Ahead Prediction ($h = 1$)

Notes: ROC-AUC Curves for each classification model across three feature sets. The left panel (a) shows performance for contemporaneous predictions ($h = 0$), while the right panel (b) shows performance for one-year-ahead predictions ($h = 1$). The blue line represents the combined Network and Firm-level feature set, the green line represents only network data, and the red line represents only firm-level data. All the plots refer to the best performing hyperparameter configuration.

deployment of limited enforcement efforts. This is also clearly shown in the confusion matrices displayed in Figures A.10 and A.11: the off-diagonal elements –i.e. Type I and Type II errors, are clearly minimized with the combined model.

Figure 2a and Figure 2b visualize these findings by plotting the ROC curves for the contemporaneous ($h = 0$) and one-year-ahead ($h = 1$) predictions, respectively. Across all four models, the ROC curve for the combined Firm + Network feature set (blue line) consistently dominates, positioned closest to the ideal top-left corner. In statistical terms, the curve for the combined model stochastically dominates the others, meaning it is unequivocally superior at distinguishing between classes regardless of the chosen classification threshold. This visually confirms its superior discriminative power, achieving a high True Positive Rate (recall) for any given False Positive Rate.

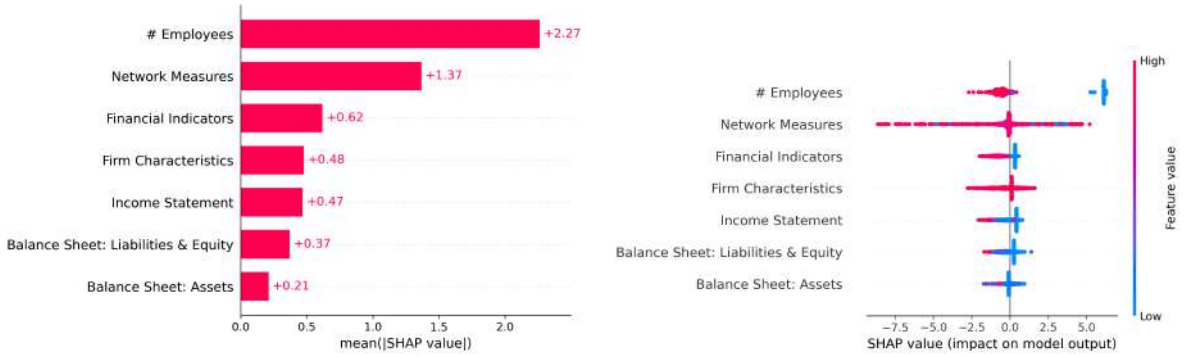
4.2. SHAP Values

To move beyond aggregate performance metrics and understand which characteristics drive confiscation risk, we conduct a feature contribution analysis using SHAP (SHapley Additive exPlanations) values. Originating from cooperative game theory, SHAP provides a unified framework for interpreting model predictions by assigning each fea-

ture an importance value for a particular prediction (Lundberg et al., 2020). For a given observation, each feature's SHAP value quantifies its contribution to pushing the model's output from the baseline (the average prediction over the training set) to its final value. Crucially, this analysis does not explain the real-world causes of confiscation; rather, it reveals the internal logic the model uses to make its predictions. This approach allows for both local interpretability—explaining why an individual firm received its risk score—and global interpretability, as we can aggregate the values across the entire dataset to assess the overall impact and directionality of each feature. For the remainder of this subsection, we will be focussing on the best performing models, i.e. XGBoost and LightGBM. Figures A.16 and A.17 in the Appendix report the results for the latter. For more details on SHAP, see Section D in the Appendix.

The magnitude of a SHAP value has a direct quantitative interpretation based on its additive property on the model's output scale, which for these classifiers is log-odds, i.e. $\log\left(\frac{p}{1-p}\right)$. For instance, a SHAP value of 0.5 for a given feature indicates that this feature increases the log-odds of confiscation by 0.5, holding all other features constant. This translates to an increase in the odds of confiscation by a factor of $e^{0.5} \approx 1.65$, or a 65% increase in odds, assuming all other factors remain unchanged. Conversely, a negative SHAP value would indicate a decrease in the log-odds, thus reducing the odds of confiscation.

Figure 3: XGBoost SHAP Values by Feature, $h = 0$



(a) Feature Importance

(b) Features' impact on model's output

Notes: On panel (a): SHAP Value for each aggregated category of feature. The x-axis the mean absolute SHAP Value, i.e. average impact of each feature group on the magnitude of the model's prediction. Panel (b): each point on the plot corresponds to a single observation for a given feature. The position on the x-axis is the SHAP value, representing the impact of that feature on the model's prediction for that observation, if $x > 0$ observation is more likely to be confiscated. The color indicates feature value, from low (blue) to high (red).

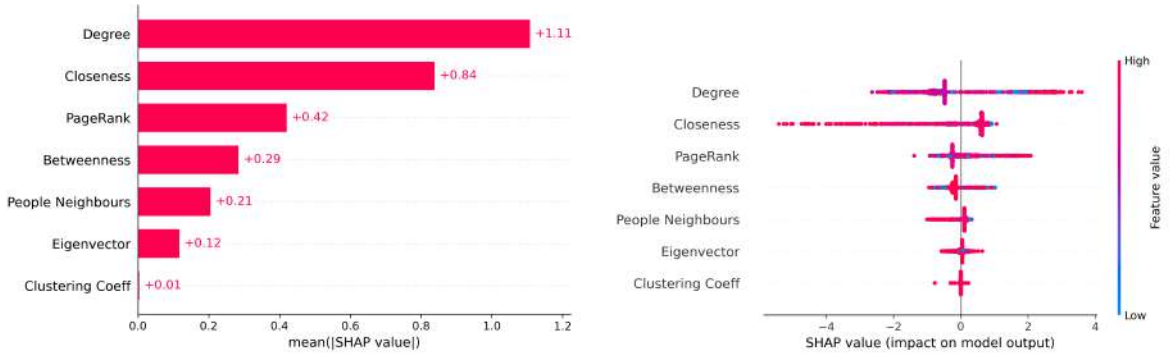
Figure 3 presents the mean absolute SHAP values for each group of features, revealing a clear and consistent hierarchy of feature importance. The quantitative impact of the top predictors is substantial; the number of employees is the single most influential feature set, with an average absolute SHAP value of 2.27, which translates to a change in the odds of confiscation by a factor of approximately 9.7 ($e^{2.27}$). The network measures follow as the second most important group, with a value of 1.37, corresponding to a change in the odds by a factor of nearly 4 ($e^{1.37}$).

On the right panel, Figure 3b provides a more granular view of how individual features within these groups influence the model's predictions. Each point represents a single firm's observation for a specific feature, with its position on the x-axis indicating the SHAP value (the feature's contribution to the log-odds of confiscation) and its color representing the actual feature value (from low in blue to high in red). The figure shows a clear separation: smaller firms consistently contribute positively to a confiscation prediction, whereas larger firms do the opposite. This empirical finding resonates with the work of Arellano-Bover et al. (2024), where they argue that smaller firms are more likely to be directly involved in criminal activities, thus making their discovery by authorities more probable.

The emergence of network measures as the second most important feature set is also a key finding. This result strengthens our conclusions, as it suggests the model's high performance is not merely an artifact of our sampling strategy; if the network's construction around confiscated firms were creating an artificially easy task, network features would likely have been the single dominant predictor, which is not what we observe. Following firm size and network metrics, financial indicators emerge as another crucial set of predictors. The beeswarm plot again provides insight, demonstrating that firms exhibiting signs of financial distress –such as low profitability (e.g., a low ROE) and poor liquidity (e.g., cash and cash equivalents)—have a higher probability of being confiscated. This observation is consistent with the mechanism described by Le Moglie and Sorrenti (2022), whereby firms might turn to investments from organized crime groups (OCGs) when legitimate avenues for borrowing are exhausted.

From Figures 3 and 4 we can see that network topology measures are also important predictors. We see that Degree (both node's in and out degree and degree centrality) changes the odds by a factor of 3, followed by Closeness which prompts a change in the odds of a factor of 2, and all the other measures. However, both from 3b and 4b there is no clear pattern pushing the probability on either side given feature values. The mix

Figure 4: XGBoost SHAP Values by Network Feature, $h = 0$



(a) Feature Importance

(b) Features' impact on model's output

Notes: On panel (a): SHAP Value for each aggregated category of feature. The x-axis shows the mean absolute SHAP Value, i.e. average impact of each feature group on the model's prediction. Panel (b): each point on the plot corresponds to a single observation for a given feature. The position on the x-axis is the SHAP value, representing the impact of that feature on the model's prediction for that observation, if $x > 0$ observation is more likely to be confiscated. The color indicates feature value, from low (blue) to high (red).

of blue and red dots in both Figures suggests that the model has learned nuanced and non-linear relationships, without a clear separation. We can only observe the distinct hierarchy emerging from Figure 4a. The relevant SHAP plots for the remaining forecast horizons, i.e. $h > 0$, are reported in Figures A.14 and A.15. For every event horizon, the results are largely left unchanged, with similar patterns in both feature importance and values, although minor differences can be observed for the LightGBM model (see A.16 and A.17). In sum, although a large impact on the odds from a given feature does not, in itself, guarantee better performance, the outstanding accuracy of our model on the test set validates its decision-making process. Therefore, we can conclude that the features identified by SHAP as most influential are indeed reliable and powerful signals of confiscation risk.

5. Robustness

Alternative Matching To ensure that our results are not driven by the specific matching strategy employed to assign synthetic confiscation years to non-confiscated firms, we re-evaluate our models using an alternative approach. Instead of propensity-score matching, we randomly sample confiscation years for non-confiscated firms from the empirical distribution of observed confiscation years, as discussed in Section 3. This

method maintains the temporal distribution of confiscation events without relying on firm characteristics for matching. The results, presented in Tables A.5 to A.8, confirm the robustness of our main findings. The combined feature set continues to outperform individual sets, and the complementary roles of network and firm-level data remain evident.

Removing Lags from the Features Another question concerning our modeling approach is the role of historical data. To assess this, we compare the performance of our models with lagged features (Table 3) against models trained only on contemporaneous data, i.e. Table A.9. We mostly consider the F1 and AUC score as they provide a comprehensive picture of a model's classification ability. This is because they collectively assess two critical aspects of performance: the model's underlying power to distinguish between positive and negative cases (AUC) and its practical success at balancing the trade-off between finding all true cases and avoiding false alarms (F1-score).

The inclusion of lags consistently improves performance, particularly for the more complex tree-based models. For instance, the XGBoost model's AUC score in the combined feature set (Panel C) increases from 0.945 to 0.975 with the addition of lags, and its F1-score rises from 0.880 to 0.911. While the largest performance gains are observed in the Network Only feature set (Panel B), the difference is smallest in the Firm-Level Only models (Panel A), suggesting that current-year financial distress is already a very powerful stand-alone predictor. This pattern can be observed also for longer prediction horizons, i.e. Tables A.10 to A.12

Interestingly, the simpler Logistic Regression model does not benefit from the added complexity of lagged features. This is likely because its rigid linear decision boundary struggles to effectively map the high-dimensional space created by the historical variables to an accurate confiscation probability. Without the complexity of lags, the model can find a clearer linear separation between the classes, resulting in a more stable performance. All in all, these results suggest that more complex models, particularly gradient boosting algorithms, are better suited to capturing the subtle temporal dynamics and non-linear patterns present in the data, leveraging historical information to achieve a higher degree of predictive accuracy.

Alternative Cutoff Date To further validate the robustness of our findings, we re-evaluate our models using an earlier cutoff date of 2016 for the train-test split, as

opposed to the original 2018. This adjustment allows us to test the models' performance with a longer out-of-sample period, thereby assessing their stability and generalizability over time. Figure A.18 illustrates the distribution of actual and synthetic confiscation dates under this new cutoff. The results, presented in Tables A.13 to A.16, reaffirm the robustness of our main conclusions: the combined feature set continues to deliver superior performance across all metrics. Notably, while the integrated model remains superior, the distinct precision-recall trade-off observed between the firm-level and network-only feature sets in our main analysis is less evident with this earlier cutoff date. The network data feature set still displays higher recall, but now the combined model also achieves the highest precision.

6. Conclusion

The infiltration of legitimate economies by organized crime groups represents a critical and evolving threat. This paper introduces a novel early-warning system designed to predict firm confiscation due to mafia involvement in Italy. By constructing a unique, time-resolved corporate ownership network and integrating it with firm-level financial data, we develop a machine-learning framework capable of identifying at-risk firms up to three years in advance with a high degree of accuracy.

Our findings yield three key contributions. First, we demonstrate that a hybrid approach, combining network topology with firm fundamentals, is substantially more powerful than relying on either data source alone, achieving an out-of-sample AUC score of up to 0.975. Second, we uncover the complementary roles of these data sources: network features are paramount for ensuring high recall (capturing a wide net of potential targets), while financial indicators provide the precision needed to pinpoint high-risk cases. Third, our feature-contribution analysis confirms existing theories that smaller, financially vulnerable firms are prime targets for infiltration. It also establishes the pivotal importance of network structure as an important indicator of risk, providing an empirical basis for a new dimension of risk-based supervision.

From a policy perspective, our framework provides a potential tool to strengthen early-warning systems for OCG infiltration, offering an additional way to signal firms at higher risk and support supervisory efforts. We acknowledge, however, that our dataset is centered on confiscated firms and their connected ownership networks, which limits external validity. Our primary contribution should therefore be understood as highlight-

ing the importance of ownership-network information for improving prediction, while recognizing that applying such models in real-world classification tasks would require broader data access and validation.

A promising avenue for future research lies in leveraging the rich, dynamic graph structure of our data through more advanced deep learning models, particularly Graph Neural Networks (GNNs). By training a GNN on the entire ownership graph, it would be possible to learn complex, non-linear patterns of connectivity that are characteristic of infiltrated networks (Rossi et al., 2020; Xu et al., 2020). This approach would allow the model to automatically generate powerful feature representations—or node embeddings—for each firm and individual, capturing nuanced information about their role and position within the economic ecosystem. These learned embeddings could not only serve as powerful inputs to improve classification performance but could also be analyzed directly to uncover subtle structural signatures of infiltration, offering deeper insights into the evolving strategies OCGs use to penetrate and exploit the legal economy.

In conclusion, this study demonstrates that the architecture of corporate control contains clear, early signals of criminal infiltration. By systematically integrating network analysis and machine learning, we offer a new and powerful framework for understanding and combating one of the most persistent threats to modern economies, paving the way for more effective and data-driven enforcement.

References

Arellano-Bover, J., De Simoni, M., Guiso, L., Macchiavello, R., Marchetti, D. J., and Prem, M. (2024). Mafias and firms.

Ash, E., Galletta, S., and Giommoni, T. (2025). A machine learning approach to analyze and support anticorruption policy. *American Economic Journal: Economic Policy*, 17(2):162–193.

Ballester, C., Calvó-Armengol, A., and Zenou, Y. (2004). Who's who in crime network. wanted the key player. Technical report, IUI Working Paper.

Boeri, F., Di Cataldo, M., and Pietrostefani, E. (2024). Localized effects of confiscated and re-allocated real estate mafia assets. *Journal of Economic Geography*, 24(2):219–240.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Calamunci, F. and Drago, F. (2020). The economic impact of organized crime infiltration in the legal economy: Evidence from the judicial administration of organized crime firms. *Italian Economic Journal*, 6(2):275–297.

Cariello, P., De Simoni, M., and Iezzi, S. (2024). A machine learning approach for the detection of firms infiltrated by organised crime in italy. *Quaderni dell'antiriciclaggio d'Italia, Unità di Informazione Finanziaria per l'Italia, Roma, Banca*.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Chircop, J., Fabrizi, M., Malaspina, P., and Parbonetti, A. (2023). Anti-mafia police actions, criminal firms, and peer firm tax avoidance. *Journal of Accounting Research*, 61(1):243–277.

Daniele, G. and Dipoppa, G. (2023). Fighting organized crime by targeting their revenue: Screening, mafias, and public funds. *The Journal of Law, Economics, and Organization*, 39(3):722–746.

De Luca, R., Greco, R., Immordino, G., and Oliviero, T. (2025). Mafia infiltration and ownership dynamics in italian companies during covid-19. Technical report, Centre for Studies in Economics and Finance (CSEF), University of Naples, Italy.

Di Cataldo, M. and Mastrorocco, N. (2022). Organized crime, captured politicians, and the allocation of public resources. *The Journal of Law, Economics, and Organization*, 38(3):774–839.

Europol (2024). Leveraging legitimacy: How the eu's most threatening criminal networks abuse legal business structures. Technical report, Publications Office of the European Union. ISBN 978-92-95236-92-9, DOI: 10.2813/7731594, QL-01-24-014-EN-N.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Le Moglie, M. and Sorrenti, G. (2022). Revealing "mafia inc."? financial crisis, organized crime, and the birth of new enterprises. *Review of Economics and Statistics*, 104(1):142–156.

Lindquist, M. J. and Zenou, Y. (2019). Crime and networks: Ten policy lessons. *Oxford Review of Economic Policy*, 35(4):746–771.

Lu, C., Durante, D., and Friel, N. (2025). Zero-inflated stochastic block modelling of efficiency-security trade-offs in weighted criminal networks. *Journal of the Royal Statistical Society Series A: Statistics in Society*, page qnaf029.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.

Mastrobuoni, G. (2020). Crime is terribly revealing: Information technology and police productivity. *The Review of Economic Studies*, 87(6):2727–2753.

Mirenda, L., Mocetti, S., and Rizzica, L. (2022). The economic effects of mafia: Firm level evidence. *American Economic Review*, 112(8):2748–2773.

Morselli, C. (2008). The efficiency–security trade-off. In *Inside Criminal Networks*, pages 1–9. Springer.

Pinotti, P. (2015). The economic costs of organised crime: Evidence from southern italy. *The Economic Journal*, 125(586):F203–F232.

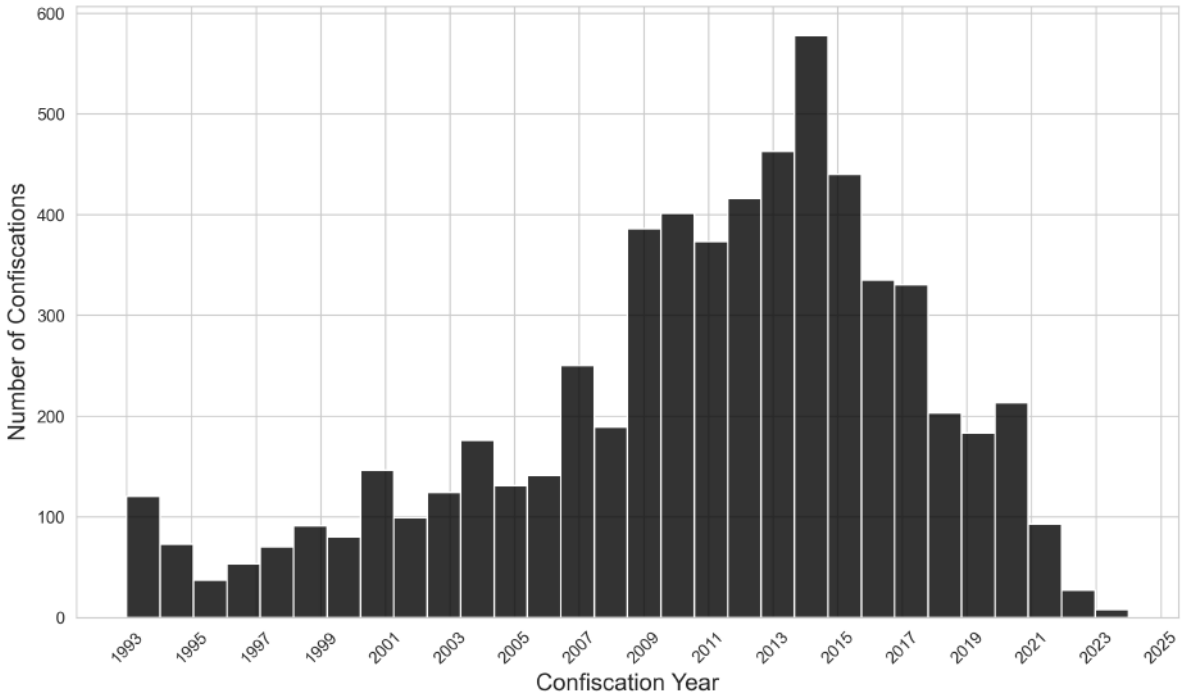
Rossi, E., Chamberlain, B., Frasca, F., Eynard, D., Monti, F., and Bronstein, M. (2020). Temporal graph networks for deep learning on dynamic graphs. *arXiv preprint arXiv:2006.10637*.

Xu, D., Ruan, C., Korpeoglu, E., Kumar, S., and Achan, K. (2020). Inductive representation learning on temporal graphs. *arXiv preprint arXiv:2002.07962*.

A. Appendix

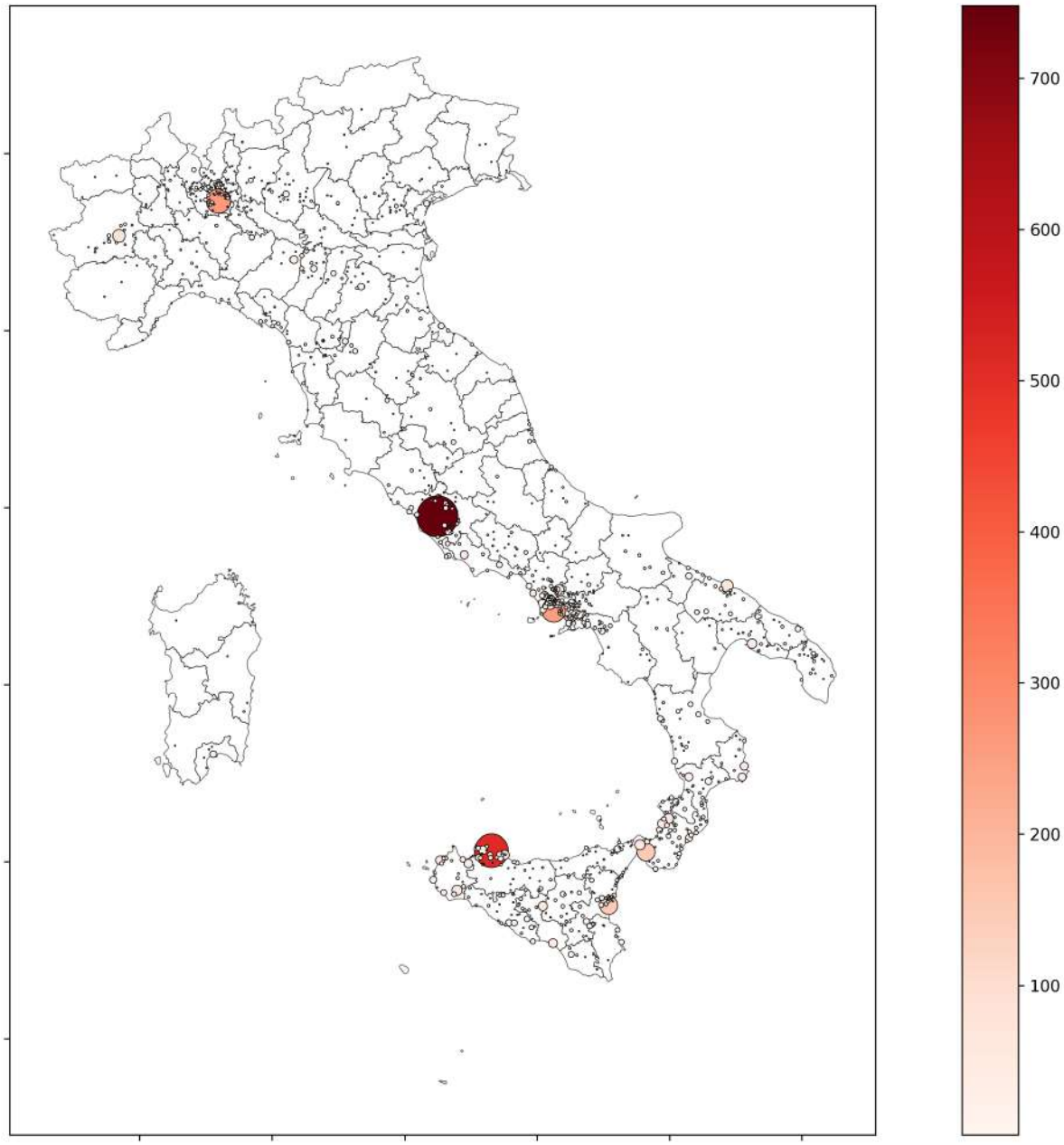
A.1. Figures

Figure A.1: Distribution of Confiscations over time



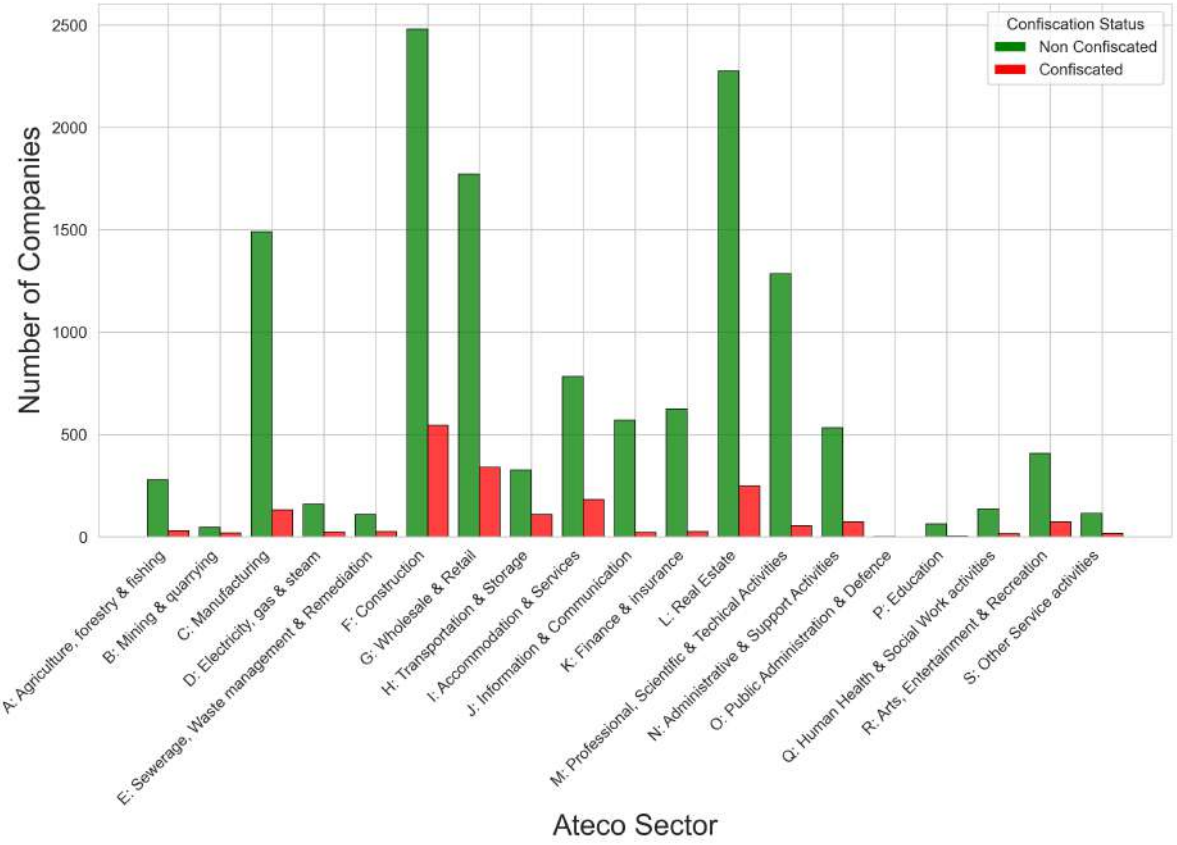
Notes: The figure shows the distribution of confiscation dates over time as reported by ANBSC. The confiscation date indicates the beginning of legal proceedings against the firm.

Figure A.2: Geographical Distribution of Confiscations



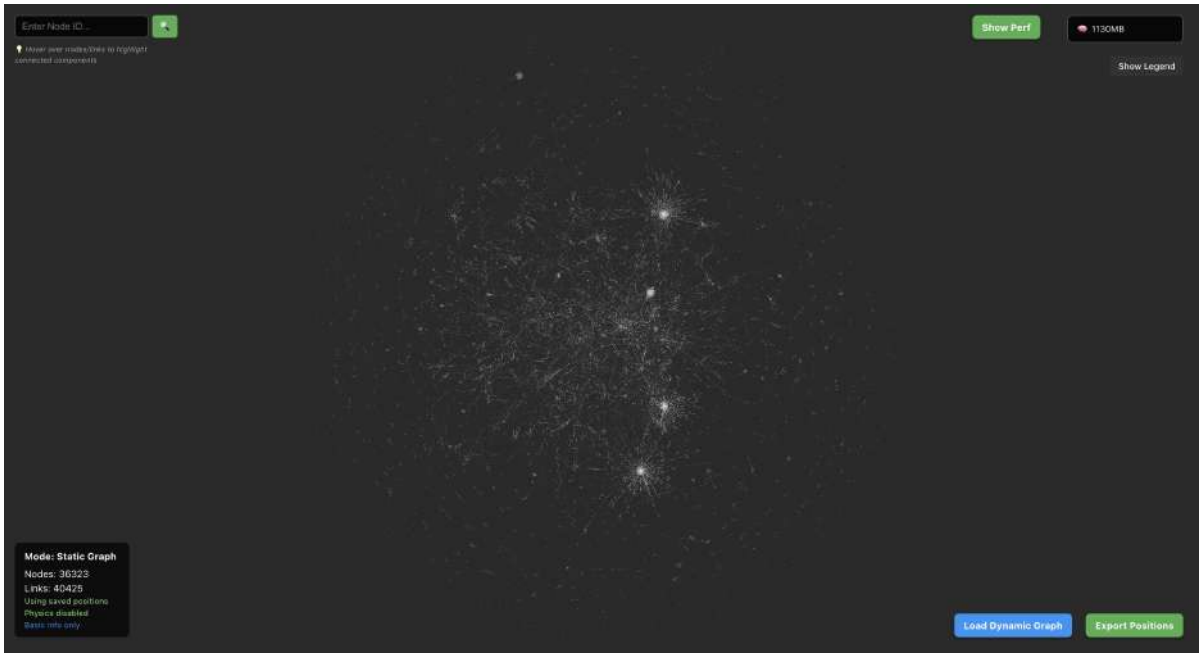
Notes: The figure shows the municipality-level geographic distribution of confiscations over Italy as reported by ANBSC. The size of the circle indicates the number of firms.

Figure A.3: Sector Distribution of Confiscated and Non-Confiscated Firms

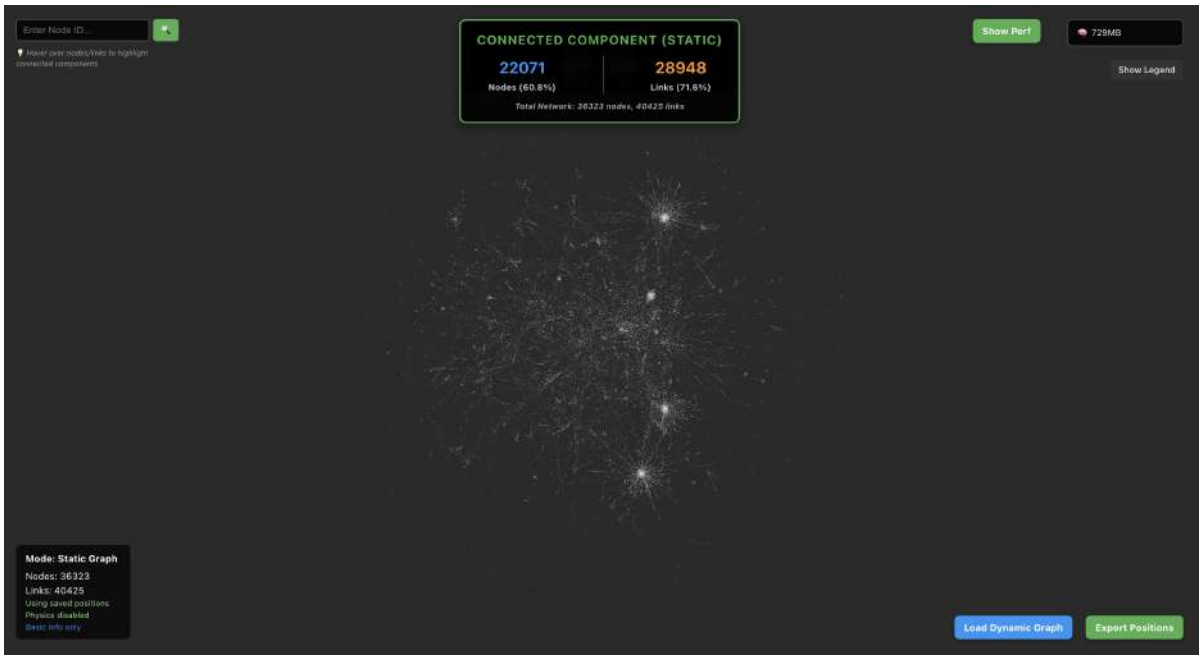


Notes: The figure shows the sectorial distribution of confiscated and non-confiscated firms, following the NACE classification.

Figure A.4: Static Graph



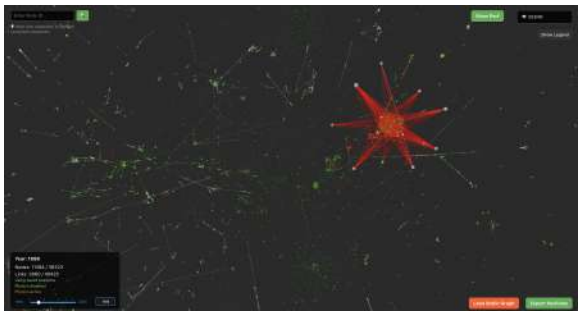
(a) All Nodes



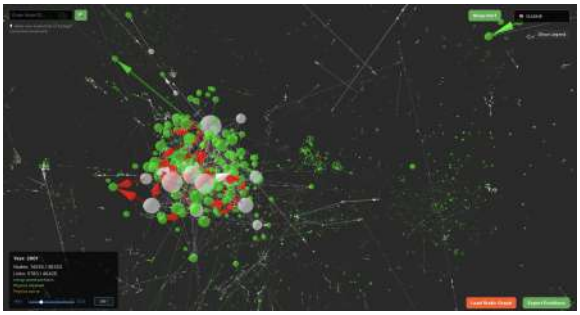
(b) Main Component

Notes: Panel (a) shows the static graph, including all nodes and edges over time. Panel (b) highlights nodes belonging to the main component. No node coloring for the static graph.

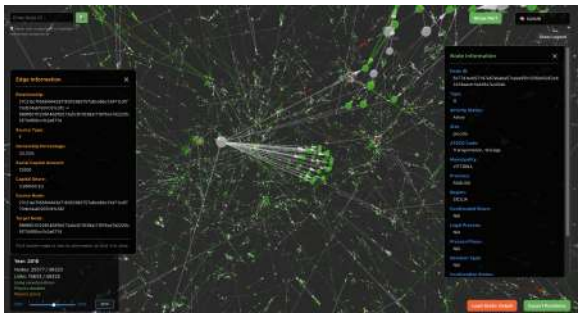
Figure A.5: Network Snapshots over Time



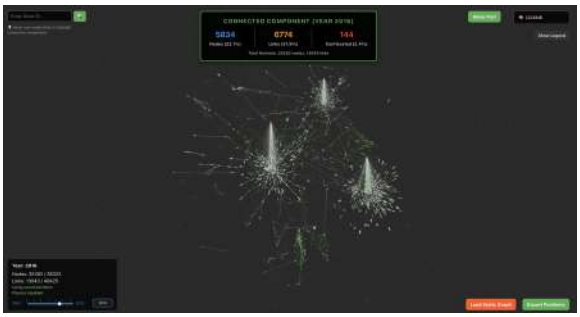
(a) Network Snapshot: 1998



(b) Network Snapshot: 2001



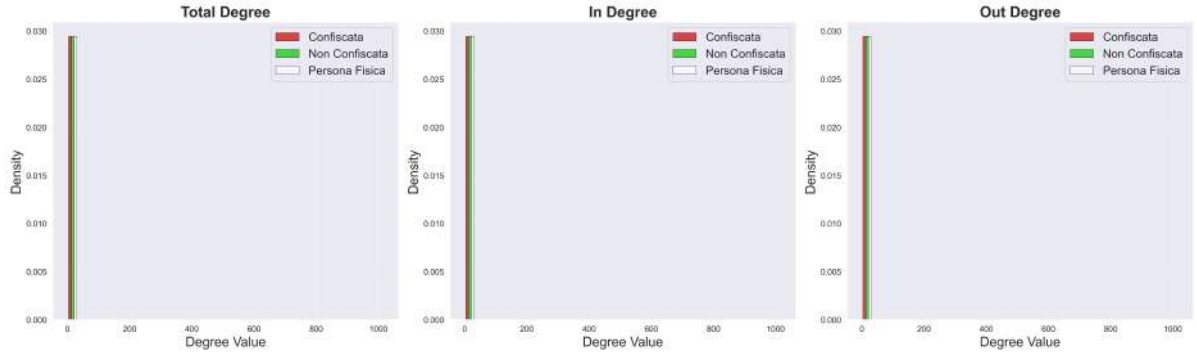
(c) Network Snapshot: 2010



(d) Network Snapshot: 2016

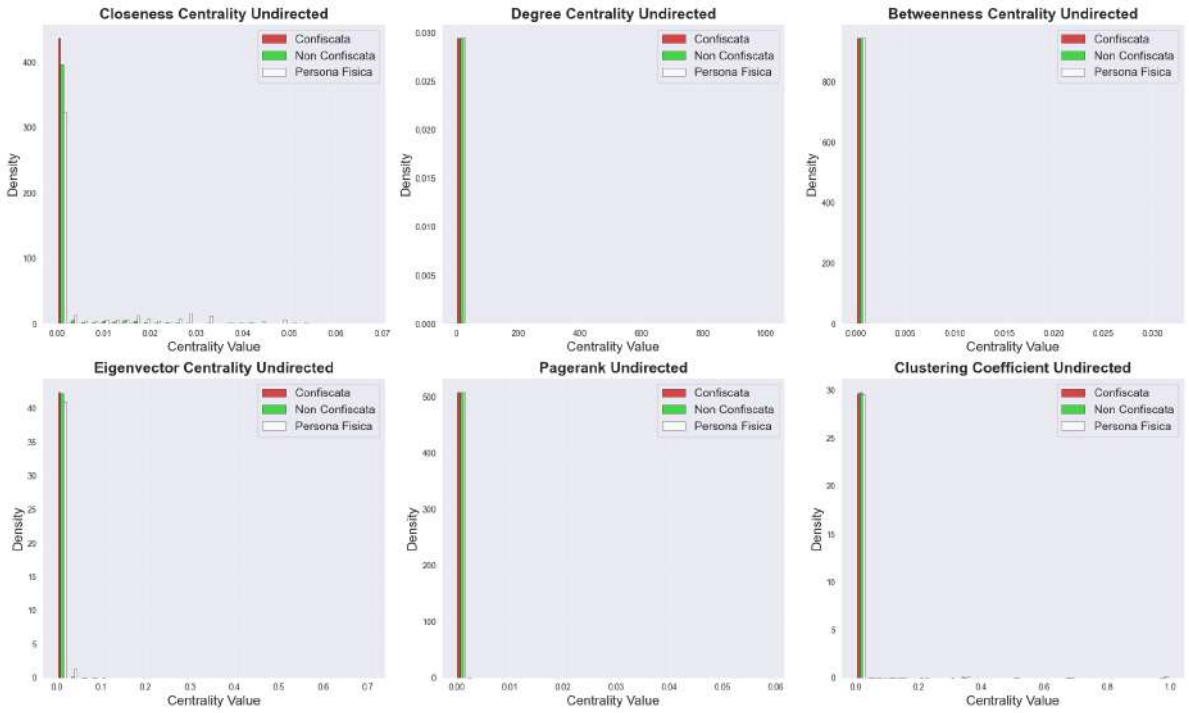
Notes: Different snapshots of the dynamic graph, showcasing visualization features. In panel A.5c the right banner displays node attributes and the left banner edge attributes. Panel A.5d displays the main connected component with the three hub nodes

Figure A.6: Degree Distributions



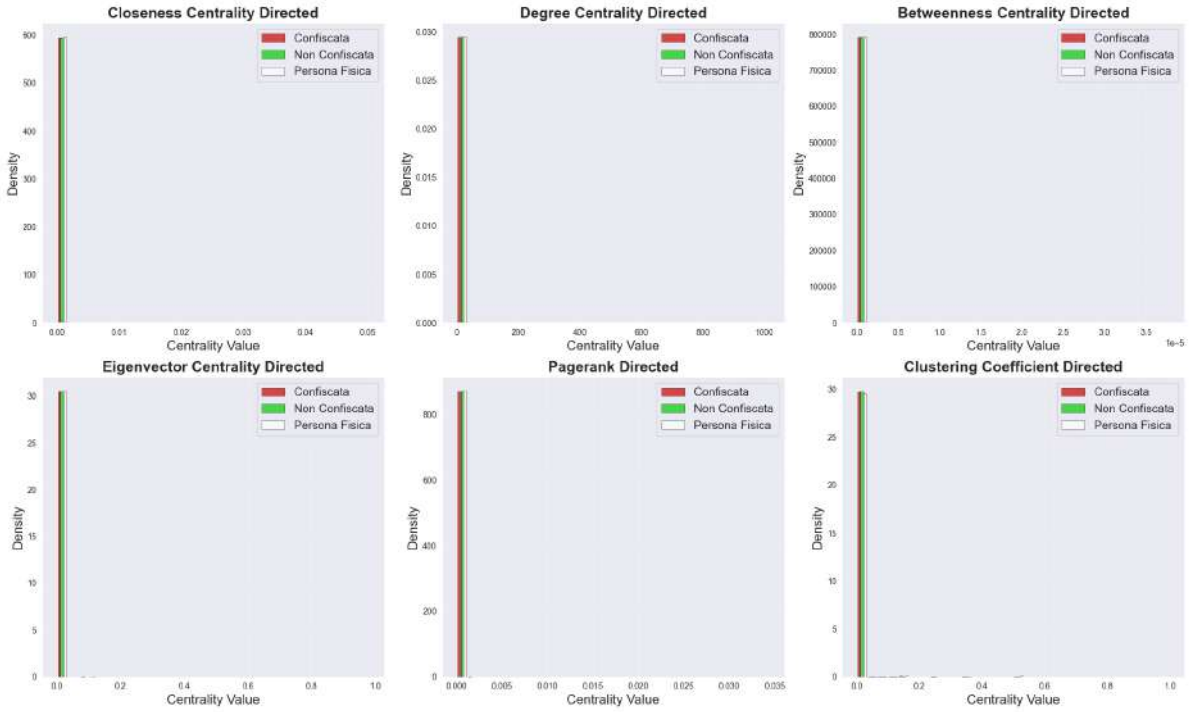
Notes: Yearly degree distributions by node type.

Figure A.7: Undirected Centralities



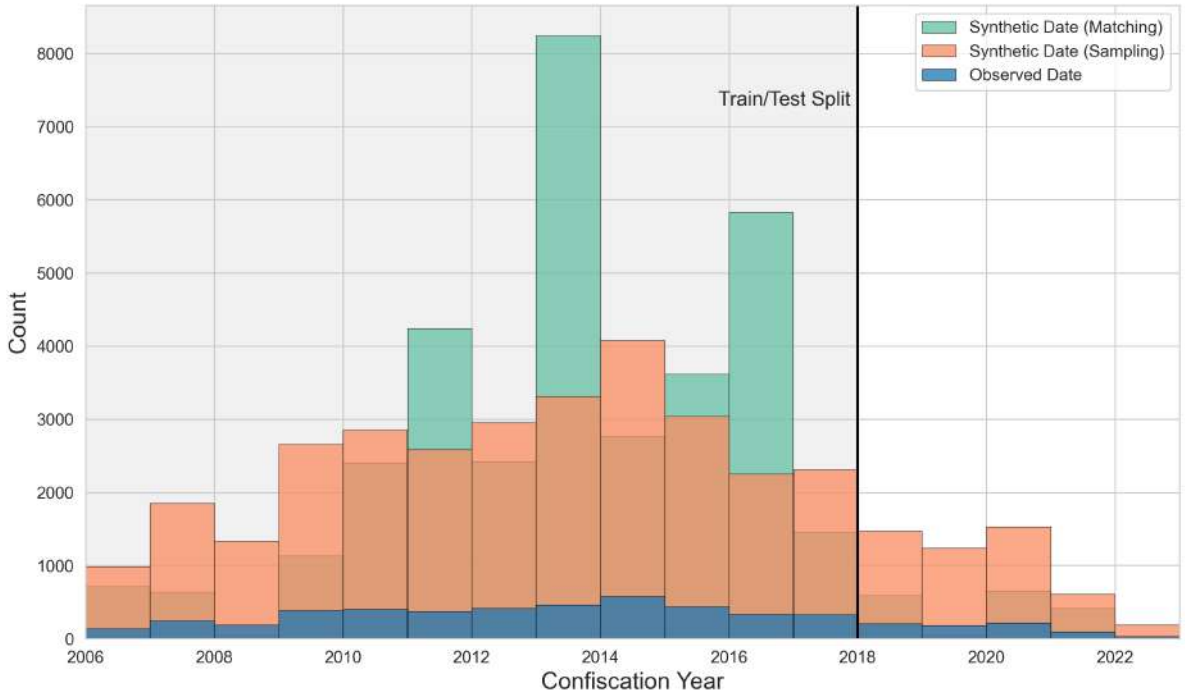
Notes: Yearly Undirected centralities by node type.

Figure A.8: Directed Centralities



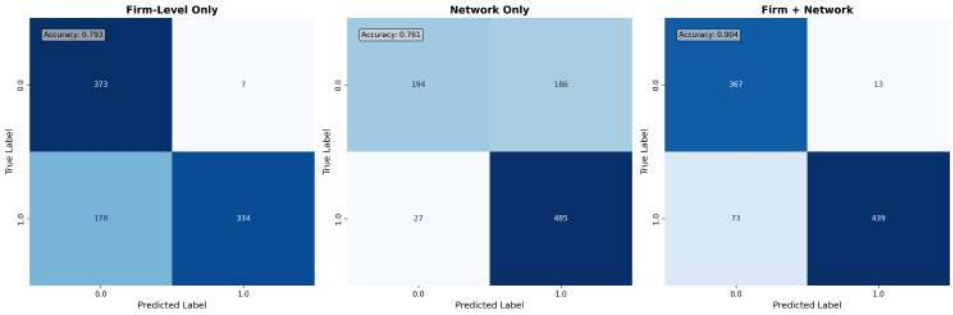
Notes: Yearly directed centralities by node type.

Figure A.9: Actual and Synthetic Confiscation Dates

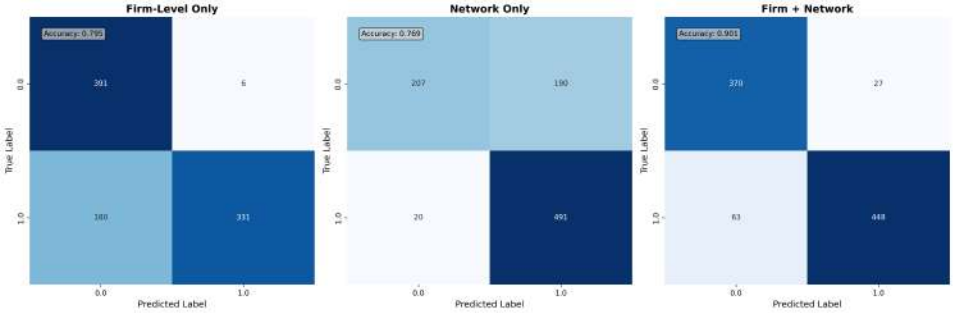


Notes: Output of the matching procedure for inferring "confiscation" dates for non-confiscated firms. Both matching and sampling methodologies for creating synthetic confiscation dates are displayed.

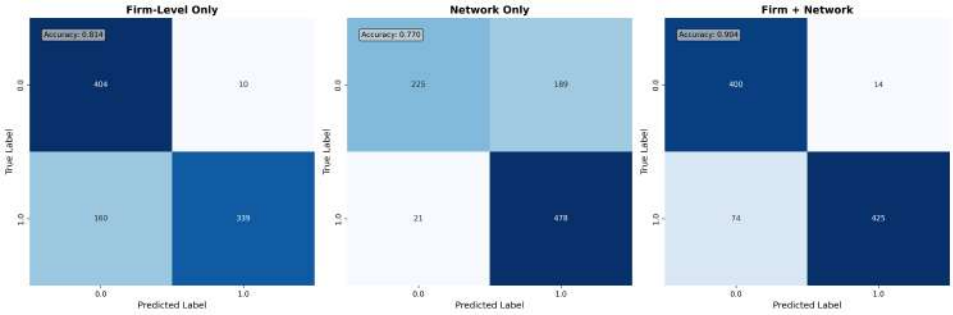
Figure A.10: XGBoost, Confusion Matrices



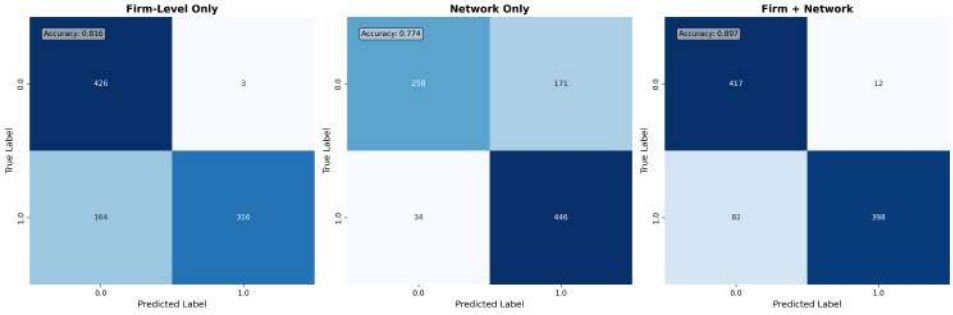
(a) $h = 0$



(b) $h = 1$



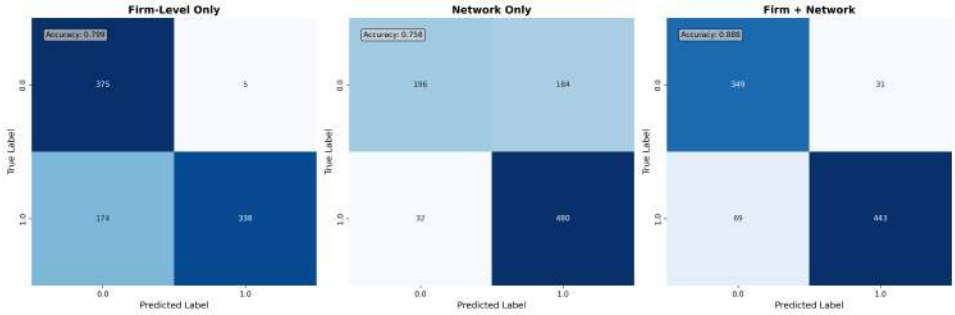
(c) $h = 2$



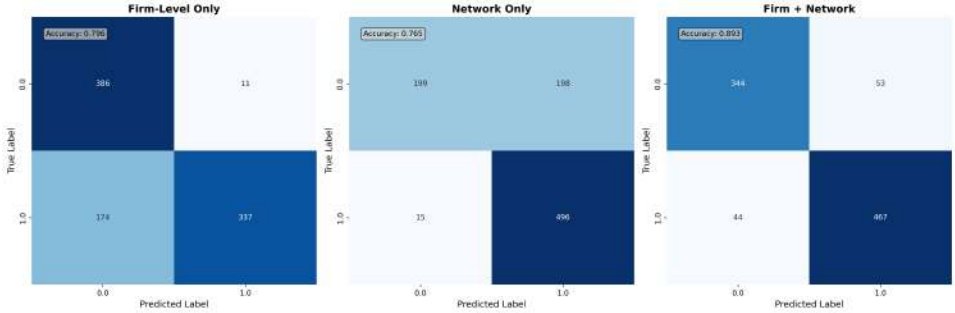
(d) $h = 3$

Notes: Distribution of correct predictions, along with Type I and Type II Errors for XGBoost model across different feature sets.

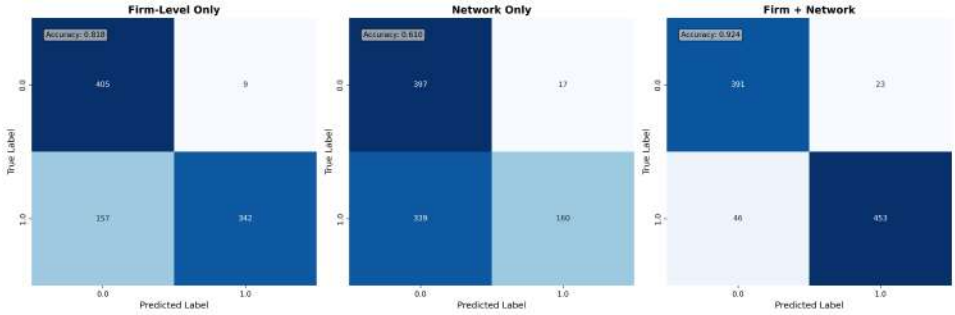
Figure A.11: LightGBM, Confusion Matrices



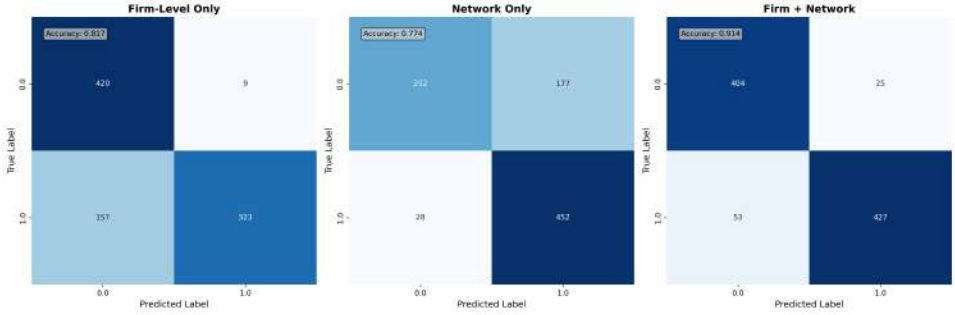
(a) $h = 0$



(b) $h = 1$



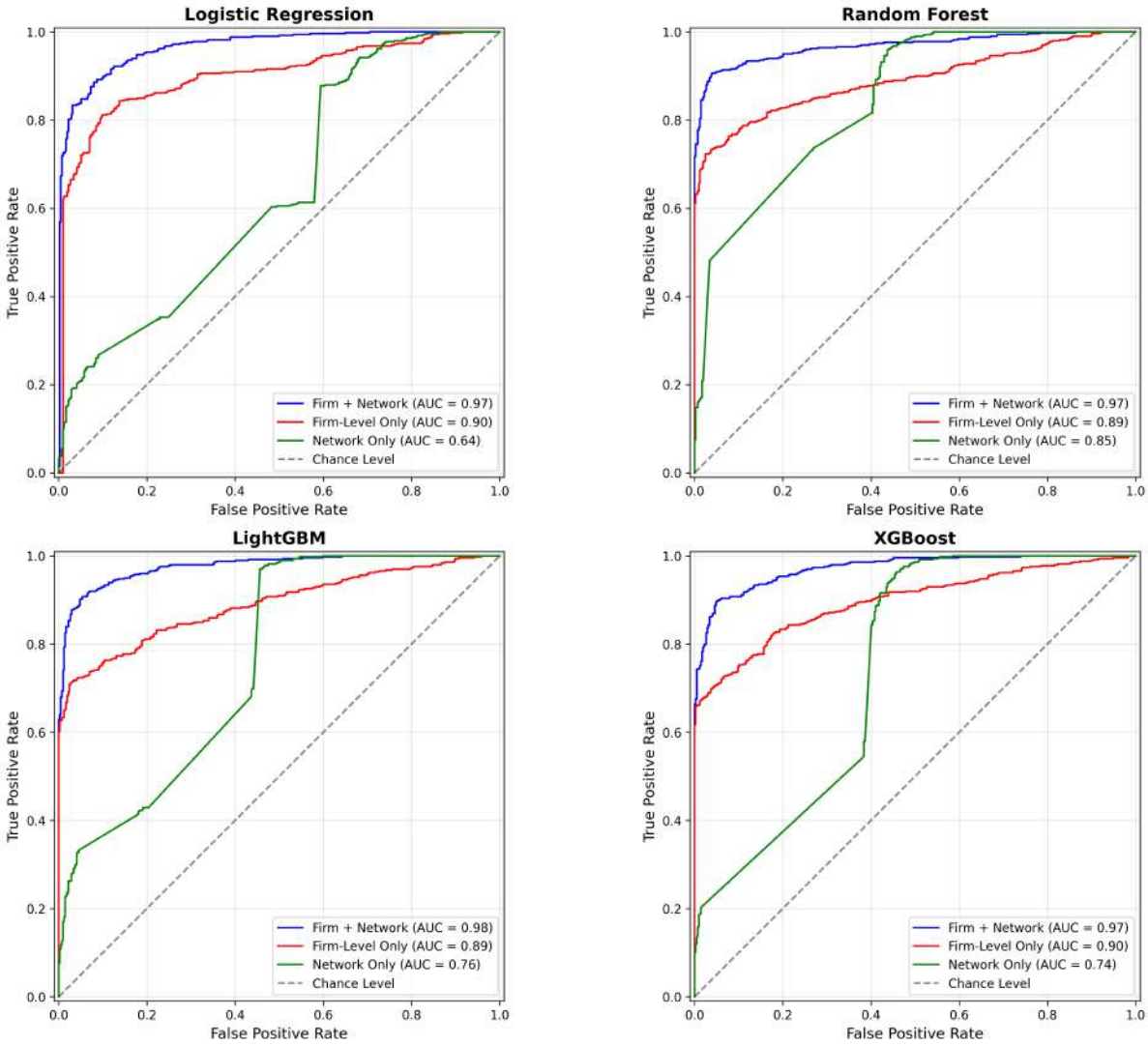
(c) $h = 2$



(d) $h = 3$

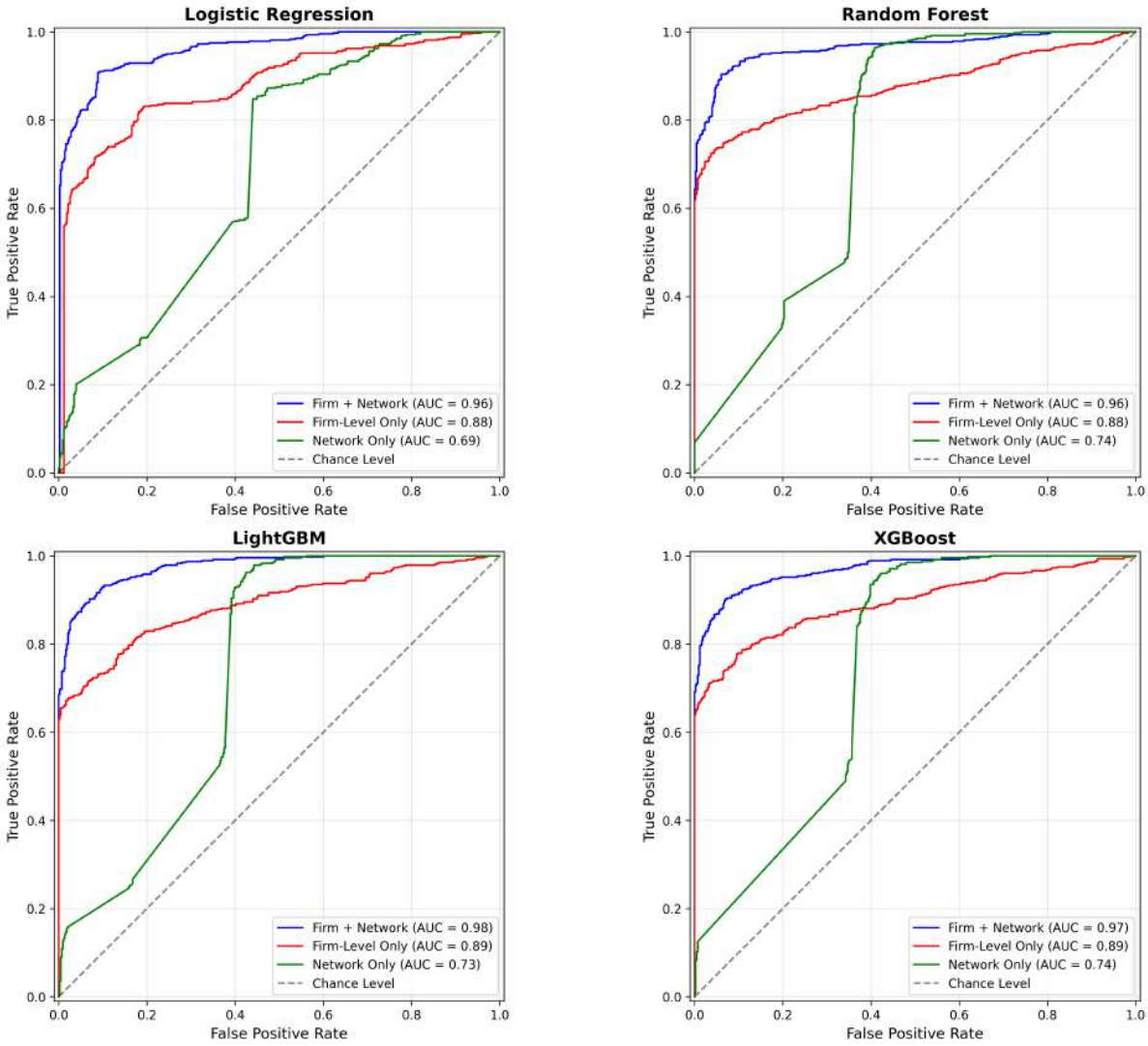
Notes: Distribution of correct predictions, along with Type I and Type II Errors for XGBoost model across different feature sets.

Figure A.12: Model Performance, $h = 2$



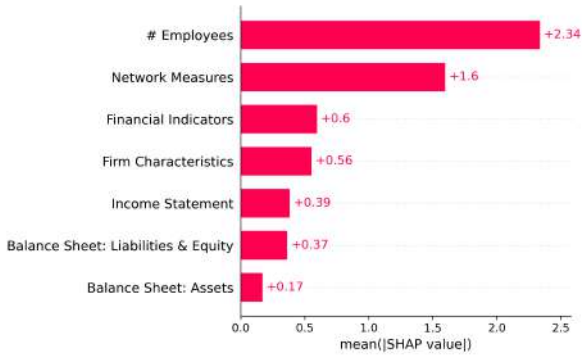
Notes: ROC-AUC Curves across the three different feature sets for each classification model. The blue line is for the Network and Firm-level data feature set, the green line for only network data and the red for firm-level data only.

Figure A.13: Model Performance, $h = 3$

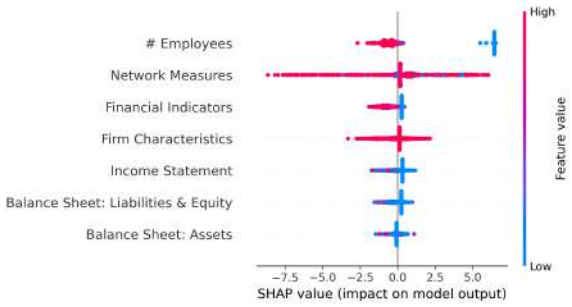


Notes: ROC-AUC Curves across the three different feature sets for each classification model. The blue line is for the Network and Firm-level data feature set, the green line for only network data and the red for firm-level data only.

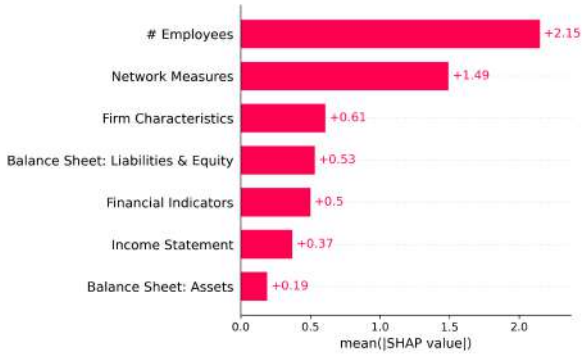
Figure A.14: XGBoost SHAP Values by Feature, $h \in \{1, 2, 3\}$



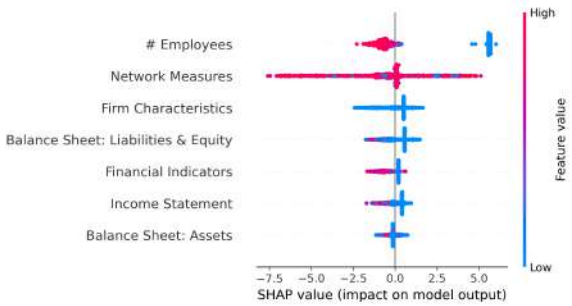
(a) Feature Importance, $h = 1$



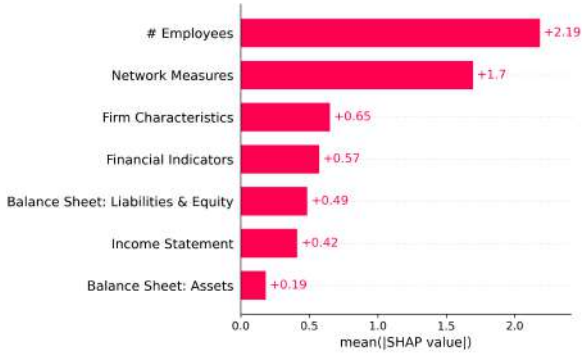
(b) Features' impact on model's output, $h = 1$



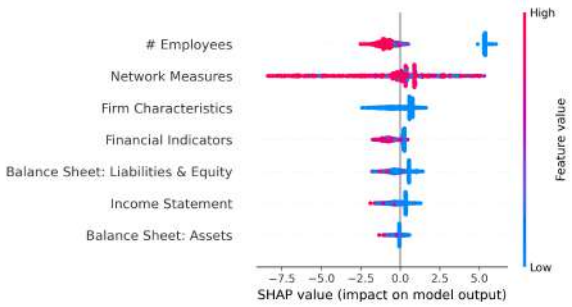
(c) Feature Importance, $h = 2$



(d) Features' impact on model's output, $h = 2$



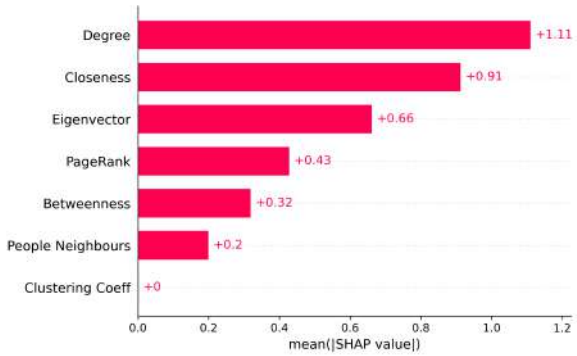
(e) Feature Importance, $h = 3$



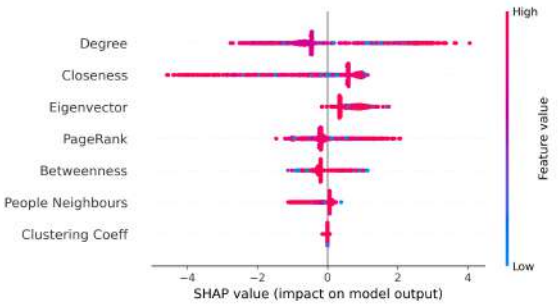
(f) Features' impact on model's output, $h = 3$

Notes: On the left column: SHAP Value for each aggregated category of feature. The x-axis the mean absolute SHAP Value, i.e. average impact of each feature group on the magnitude of the model's prediction. On the right column: each point on the plot corresponds to a single observation for a given feature. The position on the x-axis is the SHAP value, representing the impact of that feature on the model's prediction for that observation, if $x > 0$, observation is more likely to be confiscated. The color indicates feature value, from low (blue) to high (red).

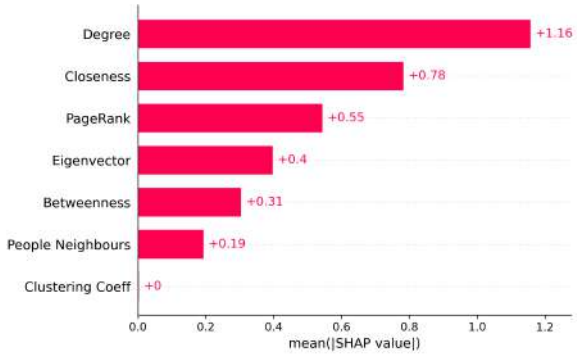
Figure A.15: XGBoost SHAP Values by Network Feature, $h \in \{1, 2, 3\}$



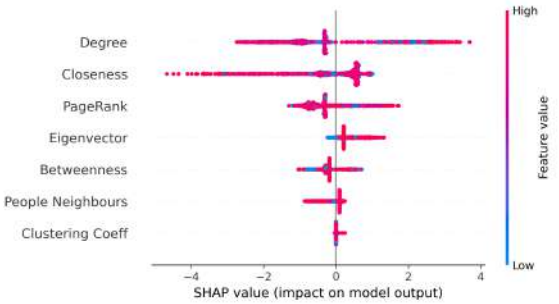
(a) Feature Importance, $h = 1$



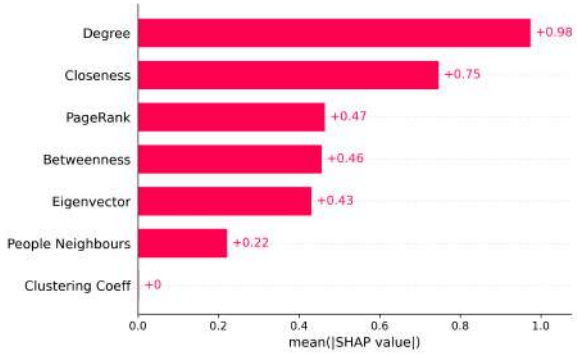
(b) Features' impact on model's output, $h = 1$



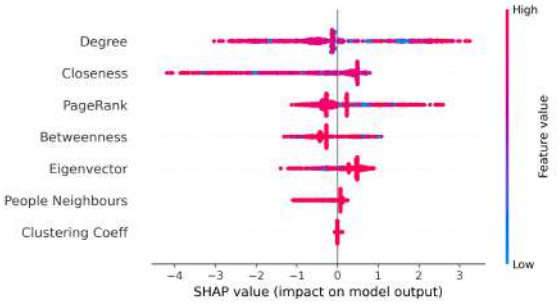
(c) Feature Importance, $h = 2$



(d) Features' impact on model's output, $h = 2$



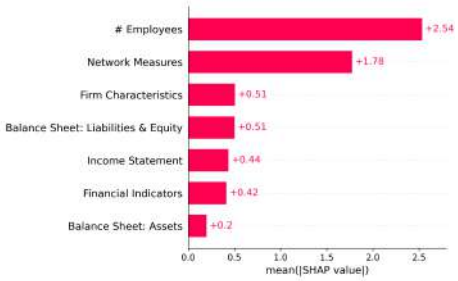
(e) Feature Importance, $h = 3$



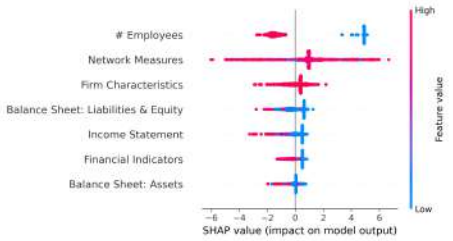
(f) Features' impact on model's output, $h = 3$

Notes: On the left column: SHAP Value for each aggregated category of feature. The x-axis the mean absolute SHAP Value, i.e. average impact of each feature group on the magnitude of the model's prediction. On the right column: each point on the plot corresponds to a single observation for a given feature. The position on the x-axis is the SHAP value, representing the impact of that feature on the model's prediction for that observation, if $x > 0$ observation is more likely to be confiscated. The color indicates feature value, from low (blue) to high (red).

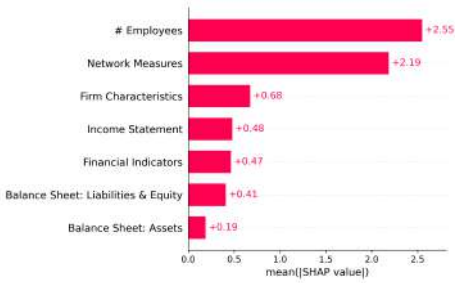
Figure A.16: LightGBM SHAP Values by Feature, $h \in \{0, 1, 2, 3\}$



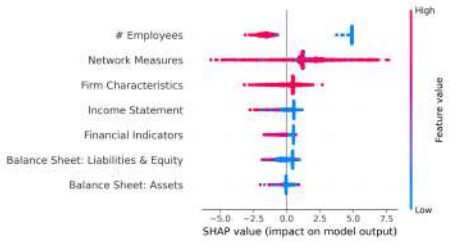
(a) Feature Importance, $h = 0$



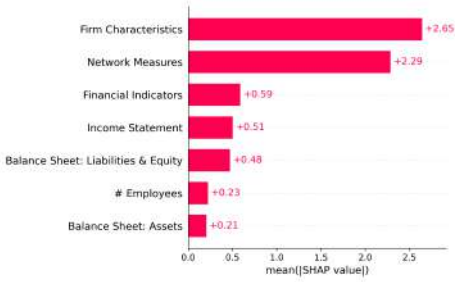
(b) Features' impact on model's output, $h = 0$



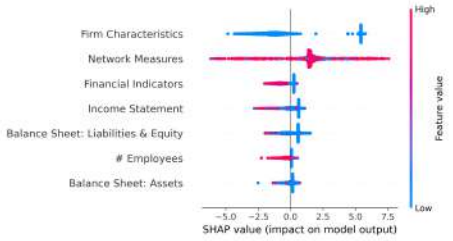
(c) Feature Importance, $h = 1$



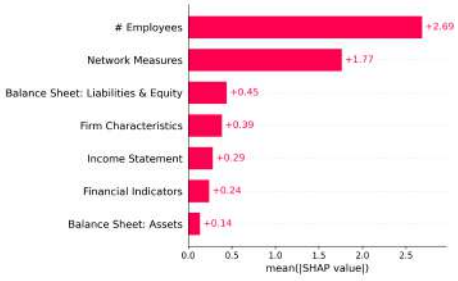
(d) Features' impact on model's output, $h = 1$



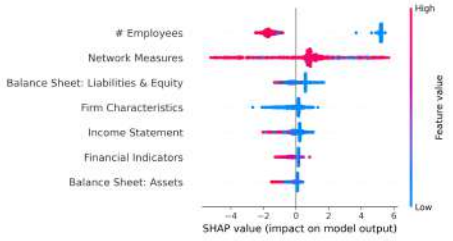
(e) Feature Importance, $h = 2$



(f) Features' impact on model's output, $h = 2$



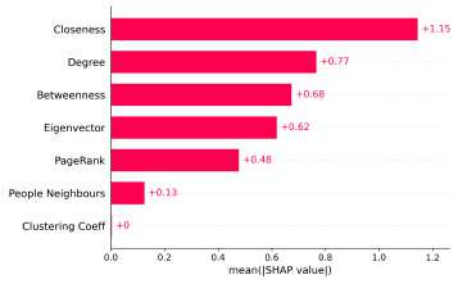
(g) Feature Importance, $h = 3$



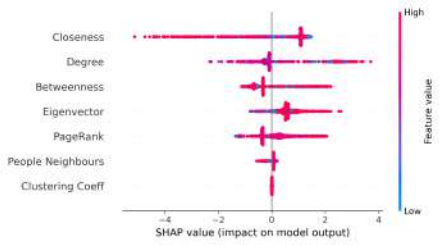
(h) Features' impact on model's output, $h = 3$

Notes: On the left column: SHAP Value for each aggregated category of feature. The x-axis the mean absolute SHAP Value, i.e. average impact of each feature group on the magnitude of the model's prediction. On the right column: each point on the plot corresponds to a single observation for a given feature. The position on the x-axis is the SHAP value, representing the impact of that feature on the model's prediction for that observation, if $x > 0$ observation is more likely to be confiscated. The color indicates feature value, from low (blue) to high (red).

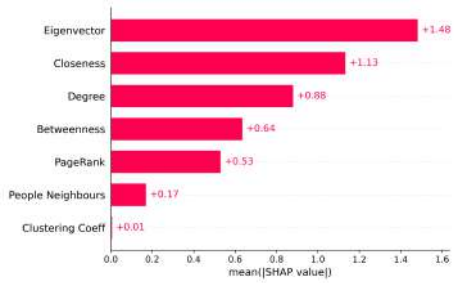
Figure A.17: LightGBM SHAP Values by Network Feature, $h \in \{0, 1, 2, 3\}$



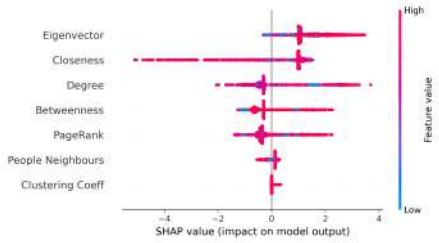
(a) Feature Importance, $h = 0$



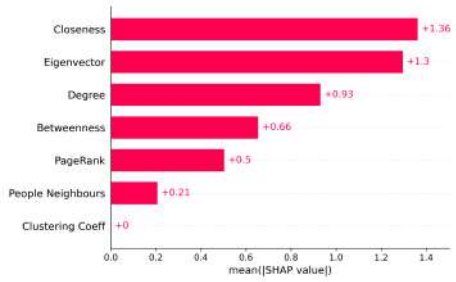
(b) Features' impact on model's output, $h = 0$



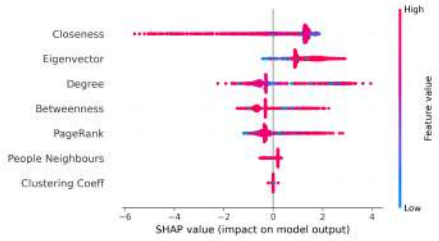
(c) Feature Importance, $h = 1$



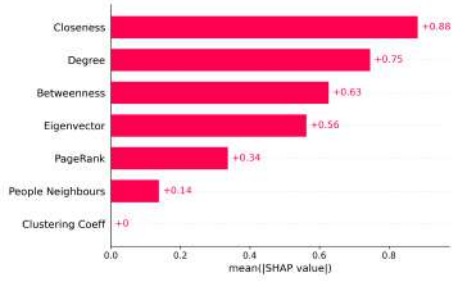
(d) Features' impact on model's output, $h = 1$



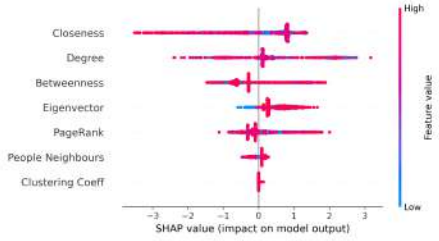
(e) Feature Importance, $h = 2$



(f) Features' impact on model's output, $h = 2$



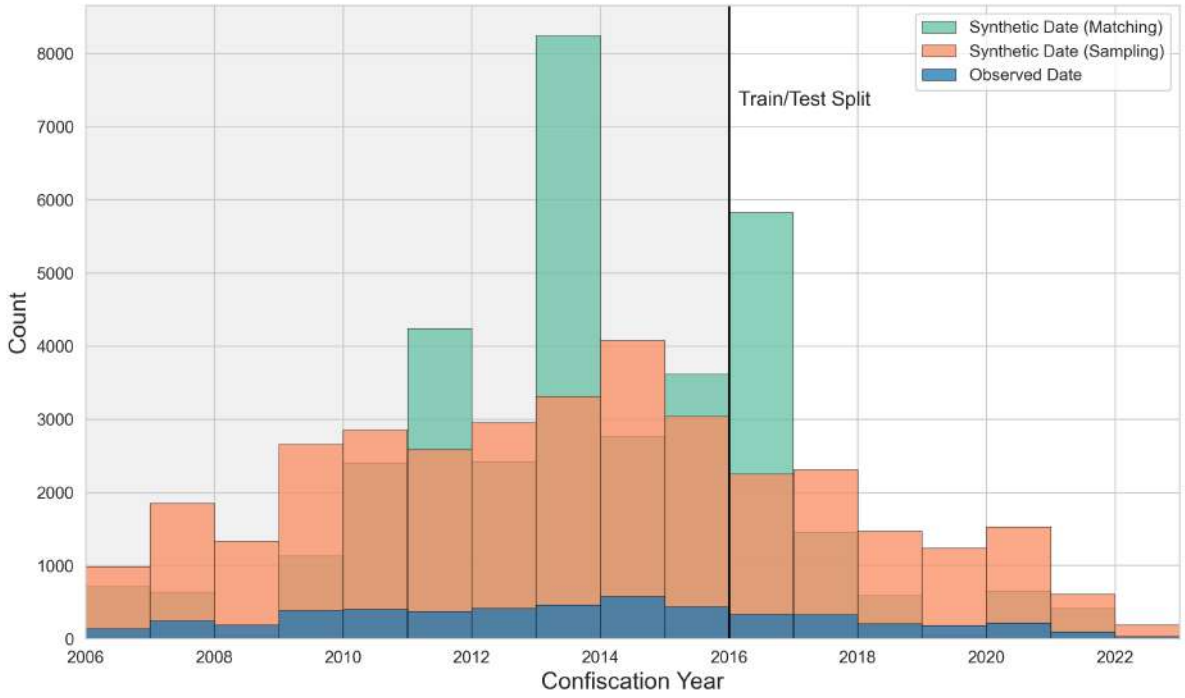
(g) Feature Importance, $h = 3$



(h) Features' impact on model's output, $h = 3$

Notes: On the left column: SHAP Value for each aggregated category of feature. The x-axis the mean absolute SHAP Value, i.e. average impact of each feature group on the magnitude of the model's prediction. On the right column: each point on the plot corresponds to a single observation for a given feature. The position on the x-axis is the SHAP value, representing the impact of that feature on the model's prediction for that observation, if $x > 0$ observation is more likely to be confiscated. The color indicates feature value, from low (blue) to high (red).

Figure A.18: Earlier Cut-off Date: Actual and Synthetic Confiscation Dates



Notes: Output of the matching procedure for inferring "confiscation" dates for non-confiscated firms. Both matching and sampling methodologies for creating synthetic confiscation dates are displayed. Here the train/test cut-off date is 2016.

A.2. Tables

Table A.1: Centrality Measures Descriptive Statistics by Network Type and Node Category

	Yearly Networks			Static Network		
Centrality Measure	People	Non-Confiscated	Confiscated	People	Non-Confiscated	Confiscated
Degree Centrality (Directed)	1.9903	0.6577	0.6511	3.1355	1.6868	2.2088
	(2.4802)	(2.9199)	(9.9793)	(4.5168)	(7.7052)	(19.5877)
Degree Centrality (Undirected)	1.9903	0.6574	0.6508	3.1355	1.6861	2.2059
	(2.4802)	(2.9186)	(9.9791)	(4.5168)	(7.7032)	(19.5851)
PageRank (Directed)	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000
	(0.0000)	(0.0001)	(0.0003)	(0.0000)	(0.0001)	(0.0003)
PageRank (Undirected)	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.0001)	(0.0001)	(0.0004)	(0.0000)	(0.0001)	(0.0002)
Eigenvector Centrality (Directed)	0.0000	0.0003	0.0002	0.0000	0.0005	0.0004
	(0.0000)	(0.0084)	(0.0068)	(0.0000)	(0.0062)	(0.0060)
Eigenvector Centrality (Undirected)	0.0027	0.0003	0.0001	0.0021	0.0001	0.0001
	(0.0104)	(0.0049)	(0.0074)	(0.0061)	(0.0013)	(0.0089)
Closeness Centrality (Directed)	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001
	(0.0000)	(0.0002)	(0.0005)	(0.0000)	(0.0006)	(0.0009)
Closeness Centrality (Undirected)	0.0064	0.0020	0.0003	0.0428	0.0490	0.0094
	(0.0123)	(0.0066)	(0.0019)	(0.0369)	(0.0303)	(0.0221)
Clustering Coefficient (Directed)	0.0032	0.0019	0.0023	0.0070	0.0103	0.0075
	(0.0348)	(0.0287)	(0.0288)	(0.0469)	(0.0656)	(0.0444)
Clustering Coefficient (Undirected)	0.0065	0.0038	0.0046	0.0140	0.0198	0.0145
	(0.0690)	(0.0543)	(0.0558)	(0.0935)	(0.1230)	(0.0835)
Betweenness Centrality (Directed)	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
Betweenness Centrality (Undirected)	0.0000	0.0000	0.0000	0.0001	0.0001	0.0002
	(0.0001)	(0.0001)	(0.0002)	(0.0007)	(0.0012)	(0.0025)

Notes: Mean values of centrality measures with standard deviations in parentheses. For yearly networks, statistics are computed by first averaging across years for each node, then by node category. Static networks show direct statistics by node category.

Table A.2: List of Firm-level Variables from *Infocamere* by Category

Firm Characteristics	**Financial Indicators**	**Balance Sheet: Assets**	**Balance Sheet: Liabilities & Equity**	**Income Statement**
size	i_valpro	aa_crediti_verso_soci	ba_totale_patrimonio_netto	ea_totale_valore_della_produzione
status	i_valagg	ab_totale_immobilizzazioni	ba1_capitale_sociale	ea01_ricavi_vendite_e_prestazioni
comune	i_marope	ab1_totale_immobilizzazioni_immateriali	ba2_riserva_da_sovrapprezzo	ea02_variazioni_rimanenze_prodotti
provincia	i_risuope	ab2_totale_immobilizzazioni_materiali	ba3_riserva_di_rivalutazione	ea03_variazione_lavori_in_corso_su_ordinazione
regione	i_risugca	ab3_totale_immobilizzazioni_finanziarie	ba4_riserva_legale	ea04_incrementi_di_immobilizz_per_lav_interni
n_addetti_tot	i_risuaim	ac_attivo_circolante	ba5_riserva_statutaria	ea05_altri_ricavi
status_attivita	i_risunet	ac1_totale_rimanenze	ba6_riserva_azioni_proprie	eb_totale_costi_della_produzione
c_s_attivita	i_immob	ac2_totale_crediti	ba7_altre_riserve	eb06_acquisti_di_materie
c_attivita	i_rimane	ac3_totale_attivita_finanziarie	ba8_utile_perdita_a_nuovo	eb07_servizi
c_ateco_benc	i_crediti	ac4_disponibilita_liquide	ba9_utile_perdita_di_esercizio	eb08_godimento_di_beni_di_terzi
gruppo_impresa	i_displiq	ad_ratei_e_risconti	bax_capitale_proprio	eb09_costi_per_il_personale
ateco_letter	i_totatt	at_totale_attivo	bb_totale_fondi_rischi	eb10_ammortamenti_e_svalutazioni
industry	i_patrnet		bb1_fondo_di_quiescenza	eb11_variazione_rimanenze_materie_prime
	i_passcons		bb2_fondo_imposte	eb12_accantonamenti_per_rischi
	i_passcorr		bb3_altri_fondi	eb13_altri_accantonamenti
	i_totpass		bc_trattamento_di_fine_rapporto	eb14_oneri_diversi_di_gestione
	i_roe		bd_totale_debiti	ec_totale_proventi_e_oneri_finanziari
	i_roi		bd1_obbligazioni	ec15_proventi_da_partecipazioni
	i_indfin		bd10_debiti_verso_imprese_collegate	ec16_altri_proventi_finanziari
	i_copimm		bd11_debiti_verso_controllanti	ec17_interessi_e_altri_oneri_finanziari
	i_liqimm		bd2_obbligazioni_convertibili	ed18_rivalutazioni
			bd4_totale_debiti_vs_banche	ed19_svalutazioni
			bd5_totale_debiti_vs_altri_finanziatori	ee_totale_proventi_e_oneri_straordinari
			bd6_acconti_anticipi_da_clienti	ef00_risultato_prima_delle_imposte
			bd7_totale_debiti_vs_fornitori	ef22_imposte_sul_reddito
			bd9_debiti_verso_imprese_controllate	ef23_utile_perdita_di_esercizio
			bdx_totale_debiti_commerciali	
			bdy_totale_debiti_residui	
			bdz_a_totale_debiti_entro_esercizio	
			bdz_b_totale_debiti_oltre_esercizio	
			be_ratei_e_risconti	
			bt_totale_passivo	

Notes: The table reports the firm-level features used for the predictive modeling, categorized into five groups: Firm Characteristics, Financial Indicators, Balance Sheet Assets, Balance Sheet Liabilities & Equity, and Income Statement. Each variable is represented by its code as per the *Infocamere* database.

Table A.3: Model Performance Comparison Across Feature Sets, $h = 2$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.807	0.805	0.818	0.814
Precision	0.952	**0.988**	**0.974**	**0.971**
Recall	0.681	0.651	0.685	0.679
F1	0.794	0.785	0.805	0.800
AUC	0.901	0.891	0.889	0.897
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.553	0.774	0.610	0.770
Precision	0.786	0.724	0.904	0.717
Recall	0.251	**0.948**	0.321	**0.958**
F1	0.380	0.821	0.473	0.820
AUC	0.640	0.847	0.758	0.744
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.883**	**0.907**	**0.924**	**0.904**
Precision	**0.971**	0.981	0.952	0.968
Recall	**0.810**	0.846	**0.908**	0.852
F1	**0.883**	**0.909**	**0.929**	**0.906**
AUC	**0.966**	**0.968**	**0.976**	**0.972**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.4: Model Performance Comparison Across Feature Sets, $h = 3$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.806	0.807	0.817	0.816
Precision	0.906	**0.990**	**0.973**	**0.991**
Recall	0.706	0.642	0.673	0.658
F1	0.794	0.779	0.796	0.791
AUC	0.878	0.877	0.891	0.894
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.673	0.777	0.774	0.774
Precision	0.635	0.727	0.719	0.723
Recall	**0.898**	**0.925**	**0.942**	**0.929**
F1	0.744	0.814	0.815	0.813
AUC	0.688	0.743	0.732	0.743
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.872**	**0.882**	**0.914**	**0.897**
Precision	**0.957**	0.963	0.945	0.971
Recall	0.794	0.808	0.890	0.829
F1	**0.868**	**0.879**	**0.916**	**0.894**
AUC	**0.960**	**0.965**	**0.976**	**0.970**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.5: Model Performance Comparison Across Feature Sets (Sampled Years), $h = 0$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.816	0.871	0.864	0.869
Precision	0.646	0.824	0.777	0.812
Recall	0.742	0.681	0.715	0.687
F1	0.691	0.745	0.744	0.745
AUC	0.877	0.889	0.902	0.902
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.429	0.617	0.616	0.613
Precision	0.322	0.417	0.415	0.415
Recall	**0.956**	**0.953**	**0.941**	**0.964**
F1	0.482	0.580	0.576	0.580
AUC	0.582	0.808	0.766	0.767
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.880**	**0.925**	**0.930**	**0.930**
Precision	**0.765**	**0.902**	**0.895**	**0.912**
Recall	0.818	0.820	0.848	0.829
F1	**0.791**	**0.859**	**0.871**	**0.868**
AUC	**0.937**	**0.957**	**0.972**	**0.972**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. The synthetic reference year for non-confiscated firms has been obtained sampling from the EDF of confiscation years, see A.9. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.6: Model Performance Comparison Across Feature Sets (Sampled Years), $h = 1$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.830	0.868	0.868	0.856
Precision	0.668	0.813	0.805	0.755
Recall	0.744	0.670	0.681	0.696
F1	0.704	0.735	0.738	0.724
AUC	0.879	0.879	0.893	0.896
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.458	0.647	0.646	0.646
Precision	0.320	0.433	0.431	0.431
Recall	**0.879**	**0.966**	**0.951**	**0.958**
F1	0.469	0.598	0.593	0.595
AUC	0.597	0.807	0.775	0.785
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.881**	**0.928**	**0.928**	**0.926**
Precision	**0.765**	**0.914**	**0.912**	**0.902**
Recall	0.812	0.812	0.814	0.818
F1	**0.788**	**0.860**	**0.860**	**0.858**
AUC	**0.937**	**0.958**	**0.969**	**0.969**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. The synthetic reference year for non-confiscated firms has been obtained sampling from the EDF of confiscation years, see A.9. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.7: Model Performance Comparison Across Feature Sets (Sampled Years), $h = 2$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.835	0.873	0.869	0.873
Precision	0.669	0.814	0.798	0.798
Recall	0.748	0.673	0.677	0.699
F1	0.706	0.737	0.733	0.745
AUC	0.875	0.876	0.892	0.892
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.493	0.671	0.671	0.665
Precision	0.331	0.444	0.443	0.439
Recall	**0.893**	**0.959**	**0.947**	**0.951**
F1	0.483	0.607	0.604	0.601
AUC	0.623	0.817	0.790	0.790
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.892**	**0.931**	**0.929**	**0.928**
Precision	**0.783**	**0.924**	**0.901**	**0.907**
Recall	0.818	0.808	0.821	0.812
F1	**0.800**	**0.862**	**0.859**	**0.857**
AUC	**0.938**	**0.959**	**0.968**	**0.970**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. The synthetic reference year for non-confiscated firms has been obtained sampling from the EDF of confiscation years, see A.9. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.8: Model Performance Comparison Across Feature Sets (Sampled Years), $h = 3$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.863	0.872	0.875	0.874
Precision	0.748	0.811	0.810	0.825
Recall	0.712	0.662	0.678	0.656
F1	0.730	0.729	0.738	0.731
AUC	0.886	0.878	0.891	0.885
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.493	0.697	0.696	0.697
Precision	0.333	0.460	0.459	0.460
Recall	**0.950**	**0.965**	**0.961**	**0.959**
F1	0.493	0.623	0.622	0.621
AUC	0.670	0.831	0.833	0.815
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.892**	**0.929**	**0.933**	**0.930**
Precision	**0.768**	**0.913**	**0.894**	**0.906**
Recall	0.837	0.802	0.843	0.817
F1	**0.801**	**0.854**	**0.868**	**0.859**
AUC	**0.946**	**0.959**	**0.970**	**0.970**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. The synthetic reference year for non-confiscated firms has been obtained sampling from the EDF of confiscation years, see A.9. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.9: Model Performance Comparison Across Feature Sets (No Lag Features), $h = 0$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.795	0.794	0.811	0.815
Precision	0.898	**0.994**	**0.965**	**0.968**
Recall	0.725	0.645	0.695	0.701
F1	0.802	0.782	0.808	0.813
AUC	0.874	0.886	0.892	0.898
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.655	0.677	0.668	0.669
Precision	0.637	0.645	0.639	0.643
Recall	**0.928**	**0.971**	**0.971**	**0.951**
F1	0.755	0.775	0.771	0.768
AUC	0.587	0.735	0.693	0.657
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.871**	**0.883**	**0.854**	**0.868**
Precision	**0.954**	0.972	0.875	0.919
Recall	0.814	0.820	0.871	0.844
F1	**0.879**	**0.890**	**0.873**	**0.880**
AUC	**0.951**	**0.953**	**0.935**	**0.945**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. All lagged features have been excluded from the feature sets. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.10: Model Performance Comparison Across Feature Sets (No Lag Features), $h = 1$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.796	0.794	0.792	0.809
Precision	0.894	**0.991**	**0.965**	**0.967**
Recall	0.724	0.640	0.654	0.685
F1	0.800	0.778	0.779	0.802
AUC	0.876	0.878	0.889	0.892
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.556	0.683	0.682	0.680
Precision	0.814	0.644	0.646	0.642
Recall	0.274	**0.977**	**0.963**	**0.973**
F1	0.410	0.776	0.773	0.774
AUC	0.596	0.754	0.580	0.698
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.868**	**0.896**	**0.869**	**0.851**
Precision	**0.960**	0.969	0.860	0.884
Recall	**0.798**	0.843	0.916	0.847
F1	**0.872**	**0.902**	**0.887**	**0.865**
AUC	**0.951**	**0.949**	**0.947**	**0.938**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. All lagged features have been excluded from the feature sets. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.11: Model Performance Comparison Across Feature Sets (No Lag Features), $h = 2$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.805	0.818	0.807	0.820
Precision	0.902	**1.000**	**0.974**	0.956
Recall	0.721	0.667	0.665	0.703
F1	0.802	0.800	0.790	0.811
AUC	0.880	0.901	0.894	0.897
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.521	0.679	0.690	0.689
Precision	0.552	0.638	0.642	0.642
Recall	0.655	**0.952**	**0.980**	**0.972**
F1	0.599	0.764	0.776	0.774
AUC	0.607	0.610	0.622	0.621
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.888**	**0.897**	**0.904**	**0.887**
Precision	**0.963**	0.976	0.960	**0.965**
Recall	**0.828**	0.832	0.860	0.824
F1	**0.890**	**0.898**	**0.907**	**0.889**
AUC	**0.955**	**0.958**	**0.968**	**0.963**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. All lagged features have been excluded from the feature sets. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.12: Model Performance Comparison Across Feature Sets (No Lag Features), $h = 3$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.809	0.815	0.809	0.825
Precision	0.882	**0.997**	**0.966**	**0.979**
Recall	0.735	0.652	0.660	0.683
F1	0.802	0.788	0.785	0.805
AUC	0.877	0.882	0.890	0.892
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.672	0.693	0.690	0.697
Precision	0.624	0.641	0.639	0.643
Recall	**0.952**	**0.950**	**0.946**	**0.958**
F1	0.754	0.766	0.763	0.770
AUC	0.627	0.612	0.631	0.619
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.893**	**0.910**	**0.877**	**0.888**
Precision	**0.961**	0.974	0.911	0.950
Recall	0.831	0.852	0.850	0.831
F1	**0.892**	**0.909**	**0.879**	**0.887**
AUC	**0.953**	**0.958**	**0.949**	**0.956**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. All lagged features have been excluded from the feature sets. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.13: Robustness (2016 Train-Test Split) - Model Performance Comparison Across Feature Sets, $h = 0$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.761	0.813	0.832	0.814
Precision	0.724	0.883	0.919	0.877
Recall	0.794	0.694	0.702	0.702
F1	0.757	0.777	0.796	0.780
AUC	0.871	0.878	0.891	0.893
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.574	0.738	0.714	0.708
Precision	0.623	0.668	0.647	0.643
Recall	0.228	**0.875**	**0.859**	0.844
F1	0.334	0.758	0.738	0.730
AUC	0.620	0.805	0.792	0.788
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.850**	**0.890**	**0.901**	**0.906**
Precision	**0.850**	**0.912**	**0.945**	**0.944**
Recall	**0.826**	0.848	0.838	**0.850**
F1	**0.838**	**0.879**	**0.888**	**0.895**
AUC	**0.934**	**0.960**	**0.967**	**0.971**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.14: Robustness (2016 Train-Test Split) - Model Performance Comparison Across Feature Sets, $h = 1$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.805	0.828	0.820	0.825
Precision	0.817	0.908	0.870	0.896
Recall	0.742	0.697	0.717	0.702
F1	0.778	0.789	0.786	0.787
AUC	0.885	0.872	0.881	0.885
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.580	0.675	0.719	0.708
Precision	0.635	0.655	0.647	0.692
Recall	0.206	0.624	**0.858**	0.661
F1	0.311	0.639	0.738	0.676
AUC	0.640	0.802	0.753	0.797
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.854**	**0.907**	**0.896**	**0.890**
Precision	**0.858**	**0.965**	**0.934**	**0.945**
Recall	**0.818**	**0.829**	0.832	**0.808**
F1	**0.838**	**0.892**	**0.880**	**0.872**
AUC	**0.933**	**0.963**	**0.960**	**0.962**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.15: Robustness (2016 Train-Test Split) - Model Performance Comparison Across Feature Sets, $h = 2$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.780	0.845	0.835	0.847
Precision	0.750	0.894	0.886	0.908
Recall	0.762	0.741	0.725	0.731
F1	0.756	0.810	0.798	0.810
AUC	0.877	0.865	0.885	0.892
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.601	0.744	0.742	0.743
Precision	0.643	0.655	0.651	0.653
Recall	0.241	**0.903**	**0.911**	**0.910**
F1	0.351	0.759	0.759	0.760
AUC	0.694	0.799	0.793	0.800
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.877**	**0.904**	**0.901**	**0.906**
Precision	**0.875**	**0.948**	**0.926**	**0.955**
Recall	**0.846**	0.832	0.847	0.828
F1	**0.860**	**0.886**	**0.885**	**0.887**
AUC	**0.941**	**0.959**	**0.969**	**0.972**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC). All the metrics refer to the best performing hyperparameter configuration identified via cross-validation on the training set.

Table A.16: Robustness (2016 Train-Test Split) - Model Performance Comparison Across Feature Sets, $h = 3$

Panel A: Firm-Level Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.790	0.859	0.849	0.856
Precision	0.752	**0.942**	**0.908**	**0.930**
Recall	0.766	0.717	0.723	0.720
F1	0.759	0.814	0.805	0.812
AUC	0.872	0.863	0.885	0.889
Panel B: Network Only				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	0.628	0.772	0.771	0.762
Precision	0.540	0.678	0.673	0.670
Recall	**0.940**	**0.899**	**0.916**	**0.884**
F1	0.686	0.773	0.775	0.762
AUC	0.668	0.847	0.842	0.800
Panel C: Firm + Network				
	Logistic Regression	Random Forest	LightGBM	XGBoost
Accuracy	**0.887**	**0.887**	**0.905**	**0.907**
Precision	**0.879**	0.873	0.893	0.917
Recall	0.855	0.865	0.886	0.863
F1	**0.867**	**0.869**	**0.890**	**0.889**
AUC	**0.942**	**0.954**	**0.973**	**0.969**

Notes: The table reports key classification performance metrics for four models across three feature sets using time-series split validation. Bold values indicate the best performing feature set for a given model and metric. Panel A uses only firm-level features, Panel B uses only network features, and Panel C combines both feature sets. The evaluation metrics include Accuracy, Precision, Recall, F1 Score, and Area Under the ROC Curve (AUC).

B. Network Visualization: Software and Settings

To facilitate the exploration of the complex, high-dimensional network dataset constructed for this paper, we developed a custom interactive 3D visualization tool. The software is built in `React` and utilizes the `react-force-graph-3d` library to render the graph. The tool is designed to be accessible to a non-technical audience while providing deep analytical capabilities. Its primary features are detailed below.

B.1. Interactive Features and Data Exploration

The visualization is designed to be a fully interactive analytical tool, allowing users to explore the network's structure and attributes in detail.

Dynamic and Static Views The tool supports two primary visualization modes. The *dynamic view* allows users to navigate through time using a slider, displaying a snapshot of the network for any given year. This feature is crucial for observing the evolution of ownership structures, the formation of new clusters, and the changing roles of specific entities over the three-decade period. The *static view* aggregates all nodes and links across the entire period into a single graph, providing a comprehensive overview of the network's full scale and identifying entities that are central over the long run.

Interactive Information Panels The graph is richly attributed and fully interactive. Clicking on any node (a firm or an individual) or edge (an ownership link) opens a detailed information panel.

- **Node Information:** For a selected firm, the panel displays key attributes such as its industry sector, size, location, and its complete judicial history, including the status and date of confiscation.
- **Edge Information:** For a selected ownership link, the panel details the ownership percentage, the corresponding monetary value of the stake, and the type of entity at the source of the link.

Visual Symbology To convey information intuitively, the rendering uses a clear visual symbology. Node colors distinguish between individuals, non-confiscated firms, and firms at various stages of the confiscation process (e.g., pending, final, or revoked). The thickness of an edge is directly proportional to the percentage of the ownership stake it

represents, making it easy to identify significant control relationships at a glance.

Component Highlighting and Search The tool includes analytical features to aid exploration. A search function allows for the immediate identification and navigation to any specific node in the network. Furthermore, hovering the cursor over a node or link automatically highlights the entire weakly connected component it belongs to, providing an immediate sense of the size and scope of a particular subgroup within the larger network.

B.2. Network Layout and Physics Simulation

The spatial arrangement of the nodes in the 3D space is determined by a force-directed layout algorithm, but it is specifically designed to ensure temporal consistency for tracking changes over time. This physics-based approach places connected entities closer together while pushing apart unconnected ones, which helps to visually reveal the underlying cluster structure of the network.

A Master Layout for Temporal Consistency To make meaningful comparisons between different yearly snapshots, the visualization is anchored on a single, consistent layout. This is achieved by first generating a force-directed layout for the complete *static graph*, which aggregates all nodes and links across the entire time period. The resulting (x, y, z) coordinate for each node is then stored and used as its fixed "master" position throughout the visualization. When a user navigates to a specific year, nodes present in that snapshot are displayed at their master coordinates. This methodology is crucial as it ensures that a node's position remains constant from one year to the next, preventing the layout from "jumping" and allowing for a clear and intuitive analysis of how the network evolves. For this reason, the computationally expensive live physics simulation is disabled during normal use, though nodes can still be dragged for inspection. In the absence of pre-calculated positions (for instance, when analyzing a new dataset), the tool can run a live, aggressive physics simulation to generate a layout from scratch.

C. Methodological Details: Machine-Learning Models

This appendix provides a detailed description of the machine learning models, hyperparameter tuning strategy, and data balancing techniques employed in the classification task.

C.1. Models

Three tree-based ensemble models were selected for their high performance in classification tasks: Random Forest, LightGBM, and XGBoost. These models are implemented within a `scikit-learn` pipeline framework.

Random Forest The Random Forest model is an ensemble learning method that operates by constructing a multitude of decision trees at training time. For classification, the final output is the class selected by the most trees. It mitigates the overfitting tendencies of individual decision trees by employing two key techniques:

1. **Bagging (Bootstrap Aggregating):** Each tree is trained on a different bootstrap sample (random sample with replacement) of the training data.
2. **Feature Randomness:** When splitting a node, each tree considers only a random subset of the features, which decorrelates the trees and further reduces variance.

This combination results in a robust model that generalizes well to new data.

LightGBM LightGBM is a high-performance gradient boosting framework based on decision trees. It belongs to the family of Gradient Boosting Decision Tree (GBDT) algorithms, which build models in a sequential, stage-wise fashion. Each new tree corrects the errors of its predecessor. LightGBM's key innovations for speed and efficiency are:

1. **Gradient-based One-Side Sampling (GOSS):** To reduce the number of data instances used for computing gradients, GOSS retains all instances with large gradients (i.e., those that are poorly trained) and randomly samples from instances with small gradients.
2. **Exclusive Feature Bundling (EFB):** It bundles mutually exclusive features (i.e., features that rarely take non-zero values simultaneously) to reduce the feature dimensionality.

These techniques allow LightGBM to train significantly faster than other boosting algorithms, especially on large datasets.

XGBoost XGBoost, which stands for Extreme Gradient Boosting, is another highly efficient and scalable implementation of the GBDT algorithm. It is renowned for its performance in machine learning competitions. Key features include:

1. **Regularization:** It incorporates both L1 (Lasso) and L2 (Ridge) regularization terms in its objective function, which penalizes model complexity and helps prevent overfitting.
2. **Sparsity Awareness:** It handles missing values efficiently by learning default directions for them in each tree node.
3. **Parallelization:** It allows for parallel processing during tree construction, significantly speeding up the training process.

The metric used for evaluation is *logarithmic loss*.

C.2. Hyperparameters and Randomized Cross-Validation

To optimize model performance, we tuned key hyperparameters using `scikit-learn`'s randomized search with time-series cross-validation. Instead of exhaustively searching all possible hyperparameter combinations like a grid search, randomized search samples a fixed number of parameter settings (i.e. 10 iterations) from specified probability distributions. This is computationally more efficient and often yields comparable results.

The cross-validation strategy considers 5 splits. This is crucial for our temporal dataset, as it ensures that the training data always precedes the validation data in each fold, mimicking a real-world scenario where the model is trained on past data to predict future outcomes. The performance of each hyperparameter combination is evaluated using the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) score, which is well-suited for imbalanced classification tasks.

The hyperparameter search spaces for each model were defined as follows:

- **Random Forest**
  - `n_estimators`: Number of trees in the forest [100, 200, 300].
  - `max_depth`: Maximum depth of the trees [10, 20, 30, None].

- `min_samples_split`: Minimum number of samples required to split an internal node [2, 5, 10].

- **LightGBM**
  - `n_estimators`: Number of boosting rounds [100, 200, 400].
  - `learning_rate`: Step size shrinkage to prevent overfitting [0.01, 0.1, 0.2].
  - `num_leaves`: Maximum number of leaves in one tree [20, 31, 40].

- **XGBoost**
  - `n_estimators`: Number of boosting rounds [100, 200, 300].
  - `learning_rate`: Step size shrinkage [0.01, 0.1, 0.2].
  - `max_depth`: Maximum depth of a tree [3, 5, 7].

C.3. Synthetic Minority Over-sampling TEchnique (SMOTE)

Our dataset exhibits a significant class imbalance, where the number of non-confiscated firms (majority class) far exceeds the number of confiscated firms (minority class). To address this, we integrated the Synthetic Minority Over-sampling TEchnique (SMOTE) (Chawla et al., 2002), from the `imblearn` library, into our training pipeline. SMOTE is applied only to the training data within each cross-validation fold to prevent data leakage into the validation set.

The SMOTE algorithm works by creating synthetic samples for the minority class instead of simply duplicating existing ones (Chawla et al., 2002). This helps the classifier learn a more generalized decision boundary. The process is as follows:

1. For each instance $\vec{x}_i$ in the minority class, its k-nearest neighbors from the same class are identified.
2. One of these neighbors, $\vec{x}_{zi}$, is randomly selected.
3. A new synthetic sample, $\vec{x}_{new}$, is generated at a random point along the line segment connecting the original instance and its selected neighbor.

Mathematically, a new synthetic instance is created using the formula:

$$\vec{x}_{new} = \vec{x}_i + \lambda \times (\vec{x}_{zi} - \vec{x}_i)$$

where λ is a random number uniformly distributed in the interval $[0, 1]$. This procedure effectively populates the feature space around existing minority class samples, creating a larger and less sparse decision region for the minority class, which helps the model to better learn its characteristics.

D. Model Interpretability with SHAP

To understand the drivers behind the predictions of our best-performing models (LightGBM and XGBoost), we employed SHAP (SHapley Additive exPlanations) (Lundberg et al., 2020). This appendix details the theoretical underpinnings of SHAP and its practical implementation in our analysis. SHAP is a unified approach to explaining the output of any machine learning model, based on the principles of cooperative game theory and their classical Shapley values. The core idea is to treat features as "players" in a game, where the "payout" is the model's prediction. SHAP values represent the fairest way to distribute this payout among the features, indicating the contribution of each feature to the prediction for a specific instance.

A Shapley value is the average marginal contribution of a feature value across all possible feature combinations (coalitions). It is the only allocation method that satisfies three desirable properties:

1. **Local Accuracy (Efficiency):** The sum of the SHAP values for all features for a given instance is equal to the difference between the model's prediction for that instance and the base value (the average prediction over the entire dataset). This is expressed as:
$$f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i$$
where $f(x)$ is the model's prediction for instance x, ϕ_0 is the base value, M is the number of features, and ϕ_i is the SHAP value for feature i.
2. **Missingness:** A feature that has no impact on the model's prediction receives a SHAP value of 0.
3. **Consistency:** If a model changes so that a feature's marginal contribution increases or stays the same (regardless of other features), its SHAP value will not decrease.

The Shapley value for a feature i is computed by considering every possible subset S of features that does not include feature i. The marginal contribution of adding feature i to the subset S is calculated. This is then averaged over all possible subsets S:

$$\phi_i(f, x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)]$$

Here, F is the full set of features, and $f_x(S)$ is the model prediction for instance x given only the feature values in subset S. Our analysis was implemented using the `shap` library in Python.