ARCHIVIO ISTITUZIONALE DELLA RICERCA

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

Insurance Fraud Detection: A Statistically-Validated Network Approach

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:
Riccardo Cesari, Michele Tumminello, Andrea Consiglio, Pietro Vassallo, Fabio Farabullini (2023). Insurance Fraud Detection: A Statistically-Validated Network Approach. JOURNAL OF RISK AND INSURANCE, 90(2), 381-419 [10.1111/jori.12415].

Availability:
This version is available at: https://hdl.handle.net/11585/957367 since: 2024-02-13

Published:

DOI: http://doi.org/10.1111/jori.12415

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/).
When citing, please refer to the published version.

(Article begins on next page)

11 September 2026

This is the final peer-reviewed accepted manuscript of:

Tumminello, M., Consiglio, A., Vassallo, P., Cesari, R., Farabullini, F. Insurance fraud detection: A statistically validated network approach (2023) ***Journal of Risk and Insurance*****, 90 (2), pp. 381-419**

The final published version is available online at: ***https://doi.org/10.1111/jori.12415***

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (https://cris.unibo.it/)

When citing, please refer to the published version.

Insurance Fraud Detection:

A Statistically-Validated Network Approach

Michele Tumminello and Andrea Consiglio*

Dipartimento di Scienze Economiche, Aziendali e Statistiche, Università di Palermo, Palermo, Italy

[andrea.consiglio,michele.tumminello]@unipa.it

Pietro Vassallo

Banca d'Italia

pietro.vassallo@bancaditalia.it

Riccardo Cesari

Istituto per la Vigilanza sulle Assicurazioni, Rome

Università di Bologna, Bologna

Italy

riccardo.cesari@ivass.it

Fabio Farabullini

Istituto per la Vigilanza sulle Assicurazioni, Rome, Italy

fabio.farabullini@ivass.it

Abstract

Fraud is a social phenomenon, and fraudsters often collaborate with other fraudsters, taking on different roles. The challenge for insurance companies is to implement claim assessment and

*Corresponding author.

improve fraud detection accuracy. We developed an investigative system based on bipartite networks, highlighting the relationships between subjects and accidents or vehicles and accidents. We formalise filtering rules through probability models and test specific methods to assess the existence of communities in extensive networks and propose new alert metrics for suspicious structures. We apply the methodology to a real database—the Italian Antifraud Integrated Archive—and compare the results to out-of-sample fraud scams under investigation by the judicial authorities.

Keywords: Insurance Fraud Detection; Bipartite Networks; Statistically-Validated Networks.

1 Introduction and main challenges

Information and communication technologies allow for the storage of big data in very efficient, and cost effective, data warehouses. This is also possible by consolidating and integrating data with different levels of heterogeneity and from a variety of sources, including social media, email, archives and documents. In the car insurance industry, accident claims offer heterogeneous and multidimensional data as they include—not being exhaustive—the coded identity of all the subjects directly involved in an accident. This includes individuals such as, drivers, passengers, car owners, witnesses, and pedestrians; professionals, such as, doctors and lawyers. There is also data on car repairs, as well as details about injuries, fatalities, requested amount, property damage, place and time of the accident, and everything about the vehicles involved.

This variety and volume of data can be properly exploited through large-scale techniques, integrating *ad-hoc* mathematical models and fast algorithms in powerful computers capable of processing enormous amounts of data in short time frames. More specifically, it should be possible to detect organized insurance frauds with this approach. The aim is to enhance the predictive power of analytical tools by bringing to the surface hidden interconnections between subjects and events. Indeed, such interactions are usually buried under noisy or spurious relationships. We will only be able to dig out the signal content by means of targeted strategies and appropriate technologies.

The extension of the fraud phenomenon in insurance varies among countries and depends on how the product classifies: life, health, motor and benefits. Experts[1] admit that "detected and undetected fraud is estimated to represent up to 10% of all claims expenditure in Europe." In their annual report—*UK Insurance & Long Term Savings Key Facts*—the Association of British Insurer dedicates a section to the fraud phenomenon. In their 2021 report, they allege

[1] `https://www.insuranceeurope.eu/publications/492/the-impact-of-insurance-fraud/`

that "fraudulent motor claims were the most common, with over 58,000 cases in 2019" and they are valued up to £605m, which is 50% of the total volume of detected cases of attempted claims fraud in 2019 (The Association of British Insurer, 2021).

The phenomenon is widespread and it goes from one side of the spectrum where opportunists invent or exaggerate a claim, to the other extreme where highly organized criminal gangs set up sophisticated motor fraud scams. To this end, in 2012 ABI launched the Insurance Fraud Register (IFR) to convey all data on known fraudsters in a single database. The database was equipped with a comprehensive package of analytics used to provide insurance intelligence.

Along the same lines, in 2012 the Italian Parliament passed a bill[2] to entrust the IVASS[3]—the Italian Institute for the Supervision of Insurance companies—with the "fight against fraud in the motor liability insurance sector by analyzing and evaluating the information obtained from the claims data bank". The bill also gives the IVASS responsibility for managing the Antifraud Integrated Archive (AIA) an industry-wide data warehouse where insurance companies are compelled to upload a detailed description of all claims. Unlike IFR, AIA is a repository collecting information coming from heterogeneous sources about the many actors and aspects involved in a car accident. These range from the drivers to the injured parties (if any) and other information recorded includes lawyers, medical examiners, insurance adjusters, witnesses, the amount claimed, vehicles and many other aspects. It is a comprehensive and exhaustive register of the claims issued from 2011 onwards, where, however, no explicit information about fraudsters is given. Any conclusion must rely on statistical analysis and specific analytical tools.[4] Since 2011, IVASS developed a set of alerts to signal to its stakeholders unusual levels of some indicators (e.g., number of accidents of a driver, number of involved injuries, claimed amount). Typically, these kinds of indicators are binary, measuring the presence or absence of a specific claim characteristic. An alert is triggered when given thresholds are tripped based on recurrences and cross-checks criteria.

The scientific literature offers a rich set of statistical tools for identifying insurance fraud patterns. They can be usefully broken down into two wide classes whose main distinctive feature is that: they make use of training sets from the fraud and the non-fraud groups (supervised methods); or they rely on "unlabelled" data where account of frauds, together with their covari-

[2]Decree-Law No 179/2012, article 21, converted to Law 221/2012 and Law 124/2017.

[3]Istituto per la Vigilanza sulle Assicurazioni, `http://www.ivass.it`.

[4]Note that, being a governmental archive, AIA includes the data collected by all of the insurance companies operating in Italy and, therefore, it does not suffer from the limitations that companies experience due to the limited information they have access to, i.e., only the one about accidents in which they are involved.

ates, are not available (unsupervised methods). Both approaches have pros and cons, and there is no "fit-all" method. (See, Derrig (2002); Viaene and Dedene (2004) for a review and Belhadji et al. (2000); Bermúdez et al. (2008); Boyer and Peter (2018); Caudill et al. (2005); Gomes et al. (2021) for model specifications and implementations.)

As observed, fraud is a social phenomenon and fraudsters often act in collaboration, with different fraudsters having different roles. Supervised methods, although they add value to the analysis, show two main drawbacks: first, their calibration is based on a set of known frauds that are very difficult to obtain, and that are a very small sample with respect to all claims. Second, they miss a peculiar feature of frauds in motor insurance, namely the existence of "criminal infrastructures" that also encompass the professional profiles operating in this field. Network models have been proved to be a successful methodology for identifying social phenomena. In particular, networks methods are suitable for disentangling complex patterns and for obtaining hidden signals from large and noisy data sets (Easley and Kleinberg, 2010; Newman, 2010).

In the vehicle insurance context, many software companies offer products implementing social network analysis to extract fraud patterns from their databases. Nevertheless, scientific literature lacks a formal and rigorous discussion on the subject matter. To the best of our knowledge, the sole article interlacing graph theory and insurance fraud is by Šubelj et al. (2011), who describe a decision support system, to unveil odd network structures in motor insurance claims. Their approach draws on two basic characteristics of fraudulent behavior: (i) the "collaborative nature" of fraudsters, involving many different actors, and (ii) continuous innovation in fraud mechanisms that necessitates a flexible approach, so that "unlabeled relationships" can emerge as soon as they are committed. A major drawback of Šubelj et al. (2011)'s system is the limited size of the data samples it can handle. Indeed, Šubelj et al. build networks upon police records. That is very restrictive since most of the claims do not go through police investigation activities. Indeed, insurance companies typically prefer to avoid long and uncertain legal proceedings. When a data warehouse is available—as in our case—suspicious structures have to first be validated by means of a "filtering" stage, in order that only statistically-significant relationships are kept. This consideration also follows the line indicated by Bauer et al. (2021), who point to the relevance of adopting more advanced, data-driven approaches to solve this important problem in insurance.

The main contributions of our paper are threefold. First, we start by building bipartite networks to highlight the relationships between subjects and accidents or between vehicles and accidents. This is a general approach that allows for the inclusion of the whole spectrum of

actors around a claim: from the drivers to the legal professionals. The dense networks obtained has to be filtered out to cut away those connections that score a low likelihood level with respect to random chance. In this respect, only structures with very strong ties will appear, thus signalling potential fraudster groups. Clearly, we are aware that a statistical anomaly cannot be considered a sentence of guilt. However, on the one hand, statistical anomalies are already used in the literature to identify suspicious activity in the insurance sector, e.g., Li et al. (2021) propose a nonparametric method for studying the misrepresentation in insurance data, which also helps to spot suspicious individuals for the validation purpose. On the other hand, information about structures with very strong ties in the network is vital for investigating units as it strongly reduces the—virtually—uncountable number of structures, and, therefore, the cost and the time needed to liquidate honest claimants.

Second, we formalize the filtering rules through probability models and we also test specific methods for assessing the existence of communities for very large networks and we propose new alert metrics for suspicious structures.

Third, we apply the above methodology to a real data warehouse—the AIA—and compare results to out-of-sample fraud scams assessed by the judicial authorities. We carry over longitudinal analyses from 2011 to the present to assess the persistence of suspicious relationships, and cross-section analyses for collecting insights about the spatial structures of frauds throughout Italy.

This paper is organized as follows: Section 1.1 lists the challenges one has to face when dealing with big data in the context of insurance fraud detection. Section 2 introduces the basic terminology and notation for bipartite networks, which are used to model relationships amongst the agents of the car insurance system. We also formalize the notation for the statistically-validated network (SVN) and, in particular, we describe the properties of the Bonferroni SVN, which is used to adjust for multiple hypothesis testing and allows for the reduction of false positive links. Section 4 describes, in details, all the steps undertaken for the implementation of our investigation system. This section includes, too, an out-of-sample analysis comparing a set of known frauds with a set of accidents randomly picked from the AIA. The conclusions section 6 offers a discussion and rounds off the paper.

1.1 Main challenges: heterogeneity, non-stationarity, localization effects and community detection

The whole methodology is tailored to deal with a very large volume of data. Indeed, AIA is a fully-fledged *data warehouse* containing detailed information about all the car accidents in Italy since 2011: AIA recorded detailed information about more than 15 million accidents involving (with different roles during and after each event) more than 20 million subjects and companies, at the end of 2017, and it is quickly growing. (see subsection 3.1 for a more precise description). The complexity of AIA requires specific analytical tools to extract fraudulent patterns and poses challenges that need to be addressed through an advanced multi-level system. We list below the main challenges we faced in analyzing AIA during the project development:

Challenge I *Curse of dimensionality.* The complexity of AIA arises from the combination of two dimensions: to one extent, the variegate forms of its data that carries the information related to each claim; to the other extent, the massive size of records that could undermine—or make impossible—the application of methods that are effective for small–medium size samples. *Community detection* is one such example (see subparagraph 4.2).

Challenge II *Data quality.* AIA is a complex collection of data coming from several heterogeneous sources (data warehouse). The meaningfulness and effectiveness of the decisions IVASS takes based on the analysis of AIA are strictly dependent on the quality of the raw data. This data needs to be pre-processed before any analytical method can be applied. In Section 4.1 we illustrate the steps undertaken to reliably improve the quality of the data for network analysis.

Challenge III *Identification and frequency of frauds.* Labelling a claim as fraudulent is not an easy matter. The investigation units of the insurance companies usually adopt regression models based on a set of indicators sensitive to the detection of fraud. Their output is an indicator that a given instance contains elements typical of a fraud. The indicator could be a continuous one, such as the one we report in Appendix A, according to our model, or categorical one similar to one we report in Appendix B. Not all the claims deemed as *"anomalous"* are then prosecuted. In general, the decision to open an in-depth investigation depends on the cost of the claim settlement. Once triggers activate an inquiry, negotiations also start. The possible result is that an agreement is reached and the case is closed, or that the claimant withdraws their complaint, or that the case is taken to the Court. The

only information available to IVASS concerns the *claim withdrawals*–not included in AIA so far, because of information genericity. However, the number of claim withdrawals is very small compared to the whole AIA, and they cannot be assumed as confirmed frauds. Even smaller are the number of frauds assessed by the Court. The acquisition of this kind of information is not systematic because legal authorities have no obligation to inform IVASS.

Challenge IV *Heterogeneity.* The AIA data warehouse is populated with information about all the actors involved in the *"accident/claim chain"*: from the claimant to the insurance adjuster; from the witness to the lawyer; from any injured party to the physician. [5] The main consequence is that subjects with very few connections (a witness, or an injured) will *"live with"* others highly-connected (lawyers or car adjusters). The challenge here is that any statistical model used to test for anomalies has to account for this kind of heterogeneity to make sure that actors with a few connections will be marked off as not being statistically significant.

Challenge V *Detecting communities of fraudsters and monitoring their evolution over time.* Strictly linked to Challenge II, our method should be able to correctly maintain the correspondence between nodes and communities of fraudsters detected through different points in time. In fact, one big community detected today can result in two or more communities in the future, and, *vice versa*, two or more communities today can become, subsequently, an imposing connected component the needs to be monitored and vanquished in the future. In Section 5 we present three examples of real and structurally-diverse communities of fraudsters that are effectively monitored over time by the proposed tool.

Challenge VI *Time and space localization.* The data contained in AIA includes claims reported since 2011, and it covers all the accidents that took place in Italy. Any probabilistic model

[5] One may be tempted to reduce the heterogeneity by excluding professionals, i.e., elements that form many connections, at a first stage of the analysis, in order to focus on "ordinary people", that is, people directly involved in the accidents. Once some clusters of these people are identified, one could then test if some professionals or companies have anomalous interactions with them, conditional on the identified clusters. Thus, given an initial cluster of ordinary folks, it could then be amended by additional agents, if appropriate. However, some professionals (e.g., lawyers and doctors) are at the core of the fraud network and represent the link among seemingly unrelated accidents and subjects. Eliminating these professionals from the network, in most cases will reduce our ability to identify structured criminal networks that, in our view, are more relevant than single, (unstructured) ordinary-folk frauds. Therefore, in principle, no subject or professional can be excluded *a priori* from the scam investigation. However, co-occurrences between subjects with specific roles in the same insurance company are not tested for statistical significance. The rationale, is that, for example, the lawyer and the car adjusters of the same company are very unlikely to participate in a fraud together.

or data mining approach working with the whole database will run into a serious issue: a small *"perturbation"* (the statistical anomaly) in the calm of the *"sea of noise"* (the null hypothesis) will be readily highlighted, even though it is just a *"ripple"* and not a *"tsunami"*. Let us put this in practical terms: two lawyers exercising their activity in the same city could interact in a significant number of accidents, compared to all accidents in Italy. On the contrary, if we restrict ourselves on the number of accidents in the environs of the city, this relationship might lose its anomalous character. Similar examples can be found in terms of time. Note that, focusing the investigation on *ex-ante* spatial or temporal sub-samples of AIA is not a viable solution, since network of fraudsters, though they have a restricted temporal or spatial perimeter, are not confined to administrative boundaries, or limited to artificial temporal segments (years, semesters, etc.). Returning to example of the lawyers, without any spatial restriction, we run the risk that lots of relationships, like those described, are signalled as anomalies: whereas on a lower scale (region, city, etc.) these would be considered to be normal.

Challenge VII *Homophily.* "Similarity breeds connections" McPherson et al. (2001), this is in synthesis an outline of the concept of homophily. In crimes related to frauds, homophily plays a relevant role as frauds require a rather high degree of cooperation, coordination, and, therefore, trust among fraudsters. If not friends, they should be at least acquaintances, which suggests that, unless an external factor destroys the relationship, the same fraudsters are likely to be involved in several frauds together over time.

Challenge VIII *Trade-off between parsimony and the effectiveness of the fraud investigation action.* The process of fraud investigation is usually structured upon different levels of severity. At lower levels, weak requirements are needed to issue an alert. The alleged fraud is then passed to higher levels of investigation which typically require the direct involvement of human operators. Anything with human operators is a high-cost activity that needs to be pursued with parsimony. But, we also need to strike a balance between in-depth anti-fraud analyses and scarce economic resources.

2 Methods

2.1 Conceptual Framework

The complex interactions of people and vehicles through accidents find a natural representation in network modelling. In particular, it is convenient for the purpose of our study to represent the motor liability system as a bipartite network, where one set of nodes is given by people (or vehicles) and the other set of nodes consists of accidents. The idea behind the antifraud detection procedure is made very clear: people do not choose their counterpart when being involved in an accident. This means that in a fully efficient system, every accident should occur by chance. This is the null hypothesis of the work, and we search for statistical evidence to falsify such a hypothesis. In the following sections we give a detailed description of the network modelling methodology adopted to reach our objective.

2.2 Bipartite Complex Networks

Complex phenomena can be described through the relationships shared by their actors. A bipartite network is the simplest—and the most natural—way to represent interactions occurring among the entities of a system. In Fig. 1 we display a bipartite network where the entities of the system are partitioned into two sets, U and S, and the relationship between any two *nodes* of each set is reproduced through a *link* connecting the two nodes. In this work, U and S refer to respectively people (vehicles) and accidents. There is an extensive literature on (bipartite) network methodology and its application to the analysis of social systems. An illustrative, but not exhaustive, list of papers include: movies and actors Barabási and Albert (1999); Song et al. (2005); Watts and Strogatz (1998), authors and scientific papers Barabási et al. (2002); Guimera et al. (2005); Newman (2001), email accounts and emails McCallum et al. (2007), mobile phones and phone calls Onnela et al. (2007), the criminal-crime relationship for assessing generalist vs specialist behaviour in crime Tumminello et al. (2013), the GOTCHA! system which is based on a bipartite graph relating companies and resources (Van Vlasselaer et al., 2017). In graph theory, a bipartite network is a *graph* with two sets of disjoint nodes. In the next section, we provide the basic notation and definitions we will use throughout the paper.

We denote by $\mathcal{G}(V, \mathbb{E})$ a *graph* where V is the set of *vertices* and $\mathbb{E}$ is the set of *edges* connecting any couple of vertices $v_i, v_j \in V$, where $i, j = 1, 2, \ldots, |V|$ and $(v_i, v_j) \in \mathbb{E}$, where $|V|$ is the cardinality of set V. The *neighborhood* of a vertex $v_i \in V$ is the sub-graph of $\mathcal{G}$ composed

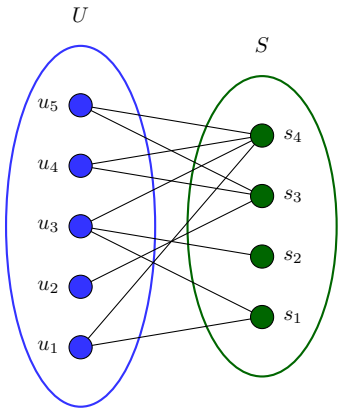

Figure 1: Bipartite network

of the vertices $v_j \in V$ and the edges $(v_i, v_j) \in \mathbb{E}$. We denote by $N(v_i)$ the neighborhood of v_i and by $\deg(v_i)$ the *degree* of v_i, i.e., the number of edges incident to the vertex v_i. If there are no loops, $\deg(v_i)$ coincides with the number of vertices of $N(v_i)$, excluding v_i itself.

A *bipartite graph* is characterized by two sets $U, S \subset V$, such that $V = U \cup S$ and $U \cap S = \emptyset$; moreover, $\forall i = 1, 2, \ldots, |U|$ and $\forall j = 1, 2, \ldots, |S|$ the edge $(u_i, s_j) \in E$ cannot have both vertex in the same set. We usually denote by $\mathcal{G}(U, S, \mathbb{E})$ a bipartite graph and we can represent it by a $|U| \times |S|$ matrix known as *bi-adjacency* matrix A, where the element a_{ij} is one when there is an edge from vertex u_i to vertex s_j, and zero otherwise,

$$(A)_{ij} = \begin{cases} 1, & \text{if} \quad (u_i, s_j) \in \mathbb{E} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The properties of bipartite networks are typically investigated by analyzing the so-called *one-mode network* or *co-occurrence network*. This is a new graph in which there is a link between two vertices of the set U if they share one or more vertices of the set S. Analogously, elements of the set S can be *"projected"* onto the set U, thus producing a new unipartite network.

2.3 Projected networks

The one-mode network is a weighted network, where the weight of a link is set according to a specific weighing function $l : U \times U \to \mathbb{R}$. Formally, given the bipartite graph $\mathcal{G}(U, S, \mathbb{E})$, the one-mode graph of U (people or vehicles) with respect to S (events) is the weighted graph denoted by $\mathcal{P}(U, \mathbb{F})$, where U is the set of vertices and $\mathbb{F}$ is the set of edges. For any $i, j = 1, 2, \ldots |U|$ and $i \neq j$, a link (u_i, u_j) is set and included in $\mathbb{F}$, if $l(u_i, u_j) > \xi$, where $\xi \in \mathbb{R}$.

The simplest weighing function imputes to each element of the weighting matrix W a value

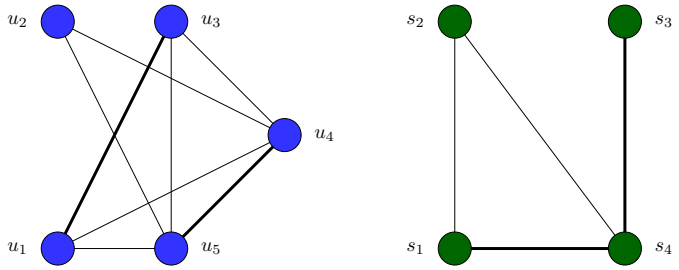

Figure 2: One-mode network

corresponding to the number of co-occurrences between u_i and u_j, i.e., $l(u_i, u_j) = |N(u_i) \cap N(u_j)|$ and $\xi = 0$:

$$(W)_{ij} = \begin{cases} |N(u_i) \cap N(u_j)|, & \text{if} \quad N(u_i) \cap N(u_j) \neq \emptyset \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Mappings like $l(u_i, u_j)$ are also know as *similarity functions* and they play a crucial role in reducing the connection density of the projected network by filtering out those links that are considered to be not significant according to given criteria (see Section 2.4).

One-mode networks can be obtained through the projection of both sides of the bipartite network onto the respective sets. In Figure 2 we show the one-mode projections extracted from the bipartite network given in Figure 1 when using the co-occurrences similarity function described above, and where edges with weights higher than one are marked by a bold line. Depending on the characteristics of the original system, a projected network can also take the form of a *directed graph*, i.e. a graph where all the edges are directed from one vertex to another. For our purposes, however, it will suffice to focus on the *undirected graph* alone.

2.4 Statistically-validated networks (SVN)

In several real-world applications, the projected network turns out to be dense, that is, the number of links is orders of magnitude larger than the number of vertices. This kind of density may hide the topological properties of the system, e.g. the presence of communities and other emergent properties.

Reducing the number of edges, by keeping those which carry the essential information about the structure of the system, is, therefore, a crucial part of our analysis. This is particularly true with respect to the arguments in Challenge I. Indeed, by setting a low value of the threshold, ξ, can lead to a poor resolution of the network and the analysis of its topological properties can

be misleading (Kenett et al. (2010), Laloux et al. (1999)).

As highlighted in Challenge IV, we deal with bipartite networks characterized with their high level of heterogeneity in terms of vertex degree. In this respect, a validation process where co-occurrences are tested against a unique threshold will lead to filtered networks where nodes (and their respective links) are validated just because they have high degrees, and, therefore, sizeable intersections with other nodes are likely to be displayed. In the claim context, that would mean that, for example, subjects like car adjusters would be over-represented in the validated network because of their *"natural"* activity within the claim process. Conversely, nodes with lower degrees (e.g., drivers) will be excluded *a priori*, thus removing interactions which can bring to light hidden anomalous behaviors.

To this end, we describe the co-occurrence between two nodes as a conditional event where the conditioning evidences are the degrees of both nodes and the total number of elements in the projecting set of the bipartite network. Formally, given the bipartite graph $\mathcal{G}(U, S, \mathbb{E})$, $\forall u_i, u_j \in U$, we define by

$$(n_{ij}|n_i, n_j, N), \tag{3}$$

the *conditional co-occurrence*, where $n_{ij} = l(u_i, u_j)$ is the unconditional co-occurrence, n_i and n_j are, respectively, the degree of u_i and u_j, and $N = |S|$ is the total number of nodes of the projecting set S[6].

Observe that the conditional co-occurrence (3) has just a symbolic meaning, however, its introduction allows comparisons with the—more substantial—*conditional threshold* that is defined as follows:

$$(\xi_{ij}|n_i, n_j, N) = Q(\alpha), \tag{4}$$

where $Q(\alpha)$ is the right-tail α-quantile of the hypergeometric distribution,

$$Q(\alpha) = \inf\left\{q \in \mathbb{Z}^+ : \alpha \geq \sum_{x=q}^{\min(n_i,n_j)} \text{Hyper}(x|n_i, n_j, N)\right\}, \tag{5}$$

and,

$$\text{Hyper}(x|n_i, n_j, N) = \frac{\binom{n_i}{x}\binom{N-n_i}{n_j-x}}{\binom{N}{n_j}}. \tag{6}$$

In practice, to avoid an *ad-hoc* choice of the acceptance threshold ξ, that would clash with the

[6]We use the conditional notation to stress the fact that the considered null hypothesis describes n_{ij} given the observed marginals n_i, n_j, and N, which are treated as parameters of the hypergeometric distribution in Eq. 6).

resolutions in Challenge I and IV, we benchmark against a random model of the co-occurences between two nodes, whose distribution function is given by (6). Finally, the level of α is chosen to guarantee that ξ_{ij} corresponds to extreme, tail, values of the random variable.

Remark 1. We specifically adopted the Hypergeometric distribution (6) to model the random co-occurences among vertices. It exactly computes the probability that k co-occurences take place when n_j links depart from node u_j and n_i links depart from node u_i. This is easily assessed if we describe the event using an urn model where, $n = n_j$ marbles are extracted without replacement from an urn with a total of $N = |S|$ marbles, and in which the urn contains $n_i = K$ marbles with a given property. In this respect, $P(X = k)$ is the probability that the sample n, drawn without replacement from the urn, shows exactly k marbles with the chosen attribute. It is worth highlighting that, the Hypergeometric distribution, in addition to being the exact probability measure, implicitly accounts for the heterogeneity of the set U. Indeed, the probability of a given intersection depends on the marginal distribution of the set U through the vertex degree n_i and n_j. In absence of heterogeneity, that is, if $\deg(u_i) \simeq \deg(u_j)$, a binomial distribution would be sufficient to approximate $P(X = k)$.

Remark 2. Similarity measures that account for the marginal distributions of u_i and u_j, i.e. that explicitly make use of n_i and n_j in their formulas, are not eligible to face the heterogeneity challenge. For example, given the bipartite graph $\mathcal{G}(U, S, \mathbb{E})$ and the associated adjacency matrix $A_{M\times N}$, where $M = |U|$ and $N = |S|$, the Pearson correlation coefficient between any two binary row vectors of A, $\rho(A_i, A_j)$, is a measure of the similarity between nodes u_i and u_j, where n_{ij} is *"adjusted by"* the degree of the two nodes, n_i and n_j. The conditional co-occurence (3) is explicitly given by

$$(n_{ij}|n_i, n_j, N) = \rho(A_i, A_j) = \frac{n_{ij} - \frac{n_i\, n_j}{N}}{\sqrt{n_i\, n_j \left(1 - \frac{n_i}{N}\right)\left(1 - \frac{n_j}{N}\right)}}. \tag{7}$$

If we consider real instances where $N \gg n_i$, n_j, for classes of nodes with almost the same vertex degree, $n_i \simeq n_j = K$, we can approximate the relationship (7) as follows:

$$\rho(A_i, A_j) \approx \frac{n_{ij}}{K}. \tag{8}$$

Equation (8) clearly shows that if we set the threshold ξ to a high level in order to reduce the complexity of the network, we will very likely be able to exclude (unless n_{ij} grows with almost

the same pace of the vertex degree K) all those nodes which characterize as *hubs*, i.e. nodes with a high vertex degree K. In fraud investigation, that would imply the exclusion, *a priori*, of subjects like lawyers or car adjusters. Conversely, a low threshold level ξ, calibrated to include node hubs of peculiar interest, would yield a very dense and uninformative network: even drivers sharing a single accident will be deemed as significant and so included in the projected network.

Armed with the conditional threshold ξ_{ij}, inferred from the null distribution $\text{Hyper}(x|n_i, n_j, N)$, the link (u_i, u_j) is statistically significant if

$$(n_{ij}|n_i, n_j, N) \geq (\xi_{ij}|n_i, n_j, N). \tag{9}$$

It is worth noticing that the validation rule in (9) is possible because both elements are conditioned to the same set of events, which, eventually, simply turns to verifying that $n_{ij} \geq \xi_{ij}$.

Finally, we denote by $\mathcal{P}(U, \mathbb{F}_\alpha)$ a *statistically-validated* projected network, where $\mathbb{F}_\alpha$ is the set of links which passed test (9), given the α-quantile of the hypergeometric distribution.

2.5 Multiple hypothesis testing

The validation test obtained from the rule (9) uses the assumption that the null hypothesis of a random co-occurrence between the couple of nodes (u_i, u_j) follows a hypergeometric distribution.

An alternative way of presenting the validation test (9) is to express it in terms of p–values. In this respect, the probability that a value larger than or equal to n_{ij} is observed by chance, under the hypothesis of a random co-occurence (6), is given by:

$$p\text{–value}(n_{ij}|n_i, n_j, N) = \sum_{x=n_{ij}}^{\min(n_i,n_j)} \text{Hyper}(x|n_i, n_j, N). \tag{10}$$

The hypothesis H_0 postulates that the link between the two nodes u_i and u_j is a noisy, random, link following a hypergeometric distribution. We reject such a hypothesis if the p–value given by the expression (10) is less than a given confidence level α.

More precisely, the p–value given in (10) tests the excess of co-occurrences between any pair of nodes linked in the projected network, and the test takes fully into account the heterogeneity of nodes u_i and u_j, since degree n_i and n_j correspond to the actual values observed in the real bipartite network. To claim that the number of co-occurrences n_{ij} between nodes u_i and u_j is too large to be consistent with the null hypothesis of random co-occurrences, we introduce a

threshold α of statistical significance to be compared with the p–value.

Using the p–value formulation allows us to better deal with a known limit of statistical validation when multiple tests are performed.

Indeed, given a projected network $\mathcal{P}(U, \mathbb{F})$, the construction of a *statistically-validated* one-mode network requires a number of tests which grows with the square of $|U|$. In particular, in the worst case, the number of tests amount to $T_U = \frac{|U||U-1|}{2}$, and the validated network will contain as many links as $|\mathbb{F}_\alpha| < T_U$. Note that, in general, $|U|$ is very high (order of millions). Given a significance level α, we then expect that at most the $\alpha \cdot 100\%$ of such repeated tests will falsely reject H_0.

There is a wide literature that deals with multiple hypothesis testing (see, for instance, Wilcox (2016) and the references therein). We control the family-wise error rate (FWER) for repeated testing: the probability of making at least one Type I Error[7].

In our fraud context, the number of tests in the family is given by T_U and the FWER controls the probability that at least one link between two subjects is significant (suspicious fraudulence), when in reality it is just a random fluctuation.

From now on, with a slight abuse of notation, we will denote by $\mathcal{P}(U, \mathbb{F}_\alpha)$ a *statistically-validated* projected network, where α is the family-wise error rate.

Among the different procedures to control FWER, we choose Bonferroni's method to validate our networks. There are two main reasons for such a choice:

- First, Bonferroni makes no assumption about the dependence structure of the p–values. We cannot exclude the possibility that a certain degree of dependence is to be found in our p–values. This is due to the validation process of the network which can involve the same node in different tests.

- Second, as is well-known, Bonferroni's control is more conservative with respect to other methods. The lower power of Bonferroni's control is a desirable property in the fraud context. Indeed, by falsely accepting the null hypothesis more often, i.e. by deeming a true link between two subjects as being due to random fluctuations, we further reduce the complexity of our network and meet the prescriptions arising from Challenge I and VIII. Needless to say, in fraud management a test with a lower power is preferable since it is implicitly in consonance with the principle of the presumption of innocence.

[7] If the tests are independent, the probability that at least one test is falsely rejected, i.e. we make a type I error, is $1 - (1 - \alpha)^{T_U}$. The latter value is virtually equal to 1, even for moderate values of T_U.

Bonferroni's correction can be obtained as corollary to the Boole's inequality Wilcox (2016). Given a *statistically- validated* one-mode network, $\mathcal{P}(U, \mathbb{F}_\alpha)$, the true positive links $\mathbb{F}_\alpha$ are the outcomes of the validation tests. The set of links $\mathbb{F}_\alpha$ is not known *ex-ante*, but, as seen above, we definitely have that $|\mathbb{F}_\alpha| < T_U$.

We denote by $L_k = L_{ij}$ the event in such a way that the hypothesis of a random linkage between (u_i, u_j) is rejected, while the alternative is true,

$$L_k = L_{ij} = (p_k < \alpha_B), \tag{11}$$

where $k \in \mathbb{F}_\alpha$, $p_k = p$–value$(n_{ij}|n_i, n_j, N)$, and α_B is the single test Bonferroni's correction.

To control the FWER at the level α, we need that

$$P\left(\bigcup_{k\in\mathbb{F}_\alpha} L_k\right) \leq \sum_{k\in\mathbb{F}_\alpha} P(L_k) = \sum_{k\in\mathbb{F}_\alpha} P\left(p_k < \alpha_B\right) < \alpha_B \left|\mathbb{F}_\alpha\right| < \alpha, \tag{12}$$

The first inequality is the Boole's inequality and the last inequality holds if we set $\alpha_B = \alpha/T_U$.

The Bonferroni correction indicates that, given a univariate global threshold of statistical significance, α, then the statistical threshold for each single test is $\alpha_B = \alpha/T_U$. The Bonferroni Statistically-Validated Network, or simply the Bonferroni Network (BN), is obtained by filtering a given real projected network, in order to only keep links that display a statistically-significant number of co-occurrences.

Notice that, if we substitute for α_B in (5), the corresponding conditional threshold is much higher than the threshold obtained using α, making the acceptance of a conditional co-occurrence n_{ij} more demanding.

A remarkable property of BN, in conjunction with the null hypothesis that random linkages follow a hypergeometric distribution, is that the co-occurrences $n_{ij} = 1$ are always excluded from the resulting one-mode networks. Such a property is highly desirable in the fraud context because all node (subject or vehicle) pairs (u_i, u_j) that are connected through only one accident represent common cases in a database of casualties (e.g the accident between two drivers), and their value is of little significance from an anti-fraud point of view.

In Proposition 1 we prove the general result that allows for the *a priori* exclusion of all the co-occurrences $n_{ij} = 1$, leading to a considerable saving in terms of computational time and storage space.

Proposition 1. *Let be $\mathcal{G}(U, S, \mathbb{E})$ a bipartite network and let $\mathcal{P}(U, \mathbb{F})$ be the one-mode projection of U with respect to S. Let also assume that $\forall u_i, u_j \in U$, and $i \neq j$, the null hypothesis of a random linkage between u_i and u_j is given by Hyper$(n_{ij}|n_i, n_j, N)$, where n_i and n_j are, respectively, the degree of the nodes u_i and u_j, and $N = |S|$. Then, a link with co-occurrences $n_{ij} = 1$ never belongs to the BN $\mathcal{P}(U, \mathbb{F}_\alpha)$, if*

$$T_U \geq \alpha N, \tag{13}$$

where T_U is the number of test in the family and α is the family-wise error ratio.

Proof. According to the hypergeometric distribution, it is trivial to show that

$$p\text{–value}(n_{ij} = 1|n_i, n_j, N) > p\text{–value}(n_{ij} = 1|1, 1, N) = \frac{1}{N}. \tag{14}$$

Recall that, a link with co-occurrence $n_{ij} = 1$ is included in the BN if p–value$(n_{ij} = 1|n_i, n_j, N) < \frac{\alpha}{T_U}$, which, in light of the inequality (14), requires that

$$\frac{\alpha}{T_U} > \frac{1}{N} \Leftrightarrow T_U < \alpha N. \tag{15}$$

□

Remark 3. Note that, the inequality (15) is never true for meaningful, real, applications, so links with co-occurrences $n_{ij} = 1$ are in no circumstance validated. This is clearly shown in Figure 3 where all the p–values labelled with $n_{ij} = 1$ are beyond Bonferroni's threshold (the vertical dashed line)[8].

To some extent, the result of Proposition 1 also partially affects co-occurrences $n_{ij} > 1$. In this case, we cannot exclude *a priori* co-occurrences $n_{ij} > 1$ or a selected part of them. However, we observe an appreciable reduction in the number of validated links for small values of the co-occurrences n_{ij}. As shown in Figure 3, a higher number of points (indicating the p–value frequency) lie to the right of Bonferroni's threshold as the number of co-occurrences decreases. Fewer and fewer links are validated as the degree of the co-occurrences gets smaller.

[8]We perform the statistical validation process over the network of subject-accident for the entire AIA. In particular, in Figure 3 we report the p-values obtained from testing the presence of links between pairs of subjects with a number of co-occurences $n_{ij} \in \{1, 2, 3, 6, 8\}$. All tests involving $n_{ij} = 1$ are not statistically significant (p-values greater than $\alpha_B = 10^{-10}$). Moreover, they represent a considerable proportion of all tests, meaning that the Bonferroni correction approach leads to a quite efficient procedure in computational terms.

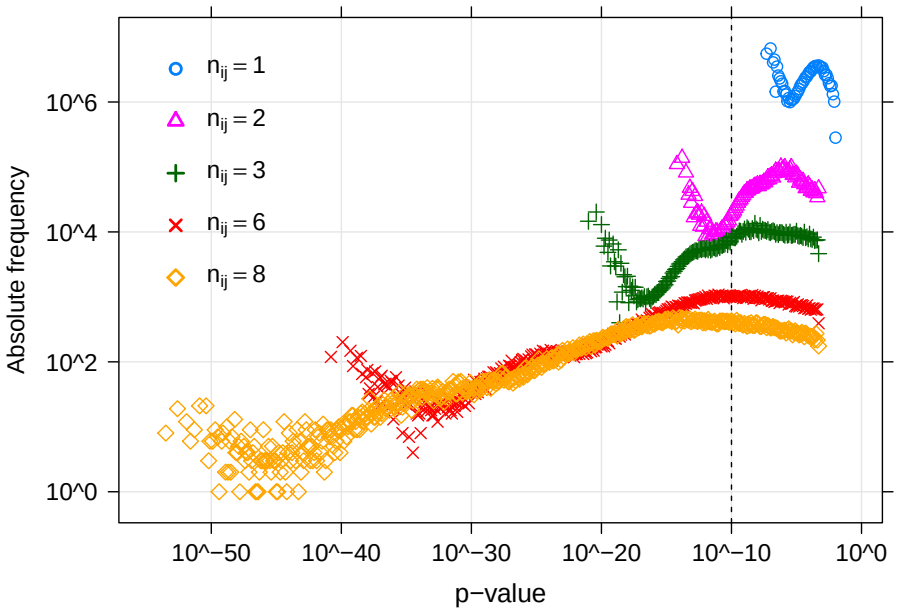

Figure 3: p–value frequencies for different levels of the co-occurrences n_{ij}. The validation process is carried over a network of subject-accident for the entire AIA. The case $n_{ij} = 1$ shows that all the p–values are beyond Bonferroni's threshold. The higher the number of co-occurrences, the higher the number of links validated.

Such a result is in line with the requirement of Challenge VIII, where the need for an effective anti-fraud activity comes up hard against limited economic resources. In this respect, it is arguably preferable to focus on networks with stronger links, by statistically pruning away less valuable bonds (in a fraudster sense).

3 Data

3.1 The IVASS Antifraud Integrated Archive

The Antifraud Integrated Archive (AIA) is the outcome of the integration of several databases, managed by both public and private bodies. The Claim Register, managed by IVASS, constitutes the central core of AIA. Insurance companies upload in real-time the features which identify each claim. AIA also embodies information from several external databases: vehicle register; driver license register; insurance coverage database; black box files; insurance expert list; and public vehicle register.

The feature data of each claim includes drivers and subjects injured (if any), lawyers, medical examiners, insurance adjusters, witnesses, amount claimed, vehicles and any other person or company directly or indirectly involved in the accidents. It is outside the scope of the present

Entity Name	Main Attributes	Definition
Accident	Id Date of the accident Place of the accident	Date and place of the accident
Claim	Date of claim Insurance company Claimant Id Claim status	Principal claim elements
Subject Claim	Claim Id Subject Id Subject role	Identification of the subject filing the claim
Subject Accident	Subject involved Id Vehicle plate Involved subject role Driver license status Flag vehicle owner Subject involved role Insurance policy status	Identification of the subject and vehicle involved in the accident. The value of the Subject involved role can be: driver, pedestrian, passenger
Claim settlement	Amount settled Settlement date Subject recipient	Settlement details of the claim
Vehicle	Plate Type Model Brand Value Ownership Damage type	Description of the vehichles involved in the accidents

Table 1: A sample of the major data entities along with their attributes. AIA contains 21 tables and each tuple of data (a single row of the database) has 81 fields.

paper to describe the relationships between objects and information in AIA. Table 1 displays a broad overview of the main entity types and the available attributes.

Being a data warehouse, AIA is not volatile: in other terms its records are not removed over time. It recorded 16,050,689 accidents and 21,574,410 subjects involved (not just claimants) at the end of January 2018, and it is quickly growing in size. Indeed, the corresponding amounts are, respectively, 18,592,317 (increase of 15.8%), and 23,943,787 (increase of 10.9%), for the end of February 2019.

The primary object of interest in assessing fraudster scams is the relational information between *subjects* and *accidents*. Likewise, association linkages between *vehicles* and *accidents* provide useful insights about unusual traits in claims. For example, in exploring the *subjects vs*

accidents relation, we are interested in answering the following questions:

- How many subjects in the database are involved in exactly $K = 1, 2, \ldots$ accidents?
- How many accidents in the database see exactly $K = 1, 2, \ldots$ subjects involved?

Similarly, the *vehicles vs accidents* relation poses the same questions with *vehicles* in place of *subjects*.

In graph theory, responding to such questions means looking at the *degree distributions* of the two sides of a bipartite network. In our context, we build and analyze the degree distributions of the two sides of the *vehicles vs accidents* and *subjects vs accidents* networks.

In Figure 4 we display the distributions of the degree of the *vehicles vs accidents* network. The left panel shows the percentage of vehicles with exactly $K = 1, 2 \ldots$ accidents, i.e., the number of links incident in each node of the vehicle's side of the bipartite network *vehicles vs accidents*. The right panel displays the degree distribution seen from the opposite front, i.e., the number of links incident in each node of the accident side (see also Figure 1 for a general representation of bi-partite networks).

We cut the tail of the two distributions to $K > 20$ for a better scaling of the two figures. However, a significant number of events (accidents and vehicles) falls in the tails of the distributions. Notice that, such events cannot be excluded *a priori* as outliers or deemed as frauds. For example, Figure 4 (left panel) clearly shows that the bulk of vehicles had incurred in 1-2 accidents. Similarly, most of the accidents link 1-2 vehicles (right panel). These observations are in line with common sense; in fact, a vehicle during its life is expected to have a few accidents, and, usually, a claim sees two vehicles involved. Less intuitive is, however, the existence of some vehicles with more than 50 accidents[9]. The coexistence of a few nodes with a high number of incident links (central nodes or *hubs*) with the great majority of nodes having few links is one of the characteristics of such networks. This wide-range dissimilarity is more pronounced in the subject-accident network where most of the subjects are involved in 3-5 accidents, (Figure 5, right panel), while roughly 4,000 subjects are involved in more than 100 accidents, and a few handfuls register the impressive number of more than 50,000 accidents. Such *massive* subjects play a technical role in the claim process. They are lawyers, insurance adjusters, physicians and all those actors who were not directly involved in the accident.

[9]Less pronounced are the extreme cases of vehicles involved in a single claim: the highest number of vehicles involved in a single claim amount to 86, and these types of claims usually concern crashes on the motorway.

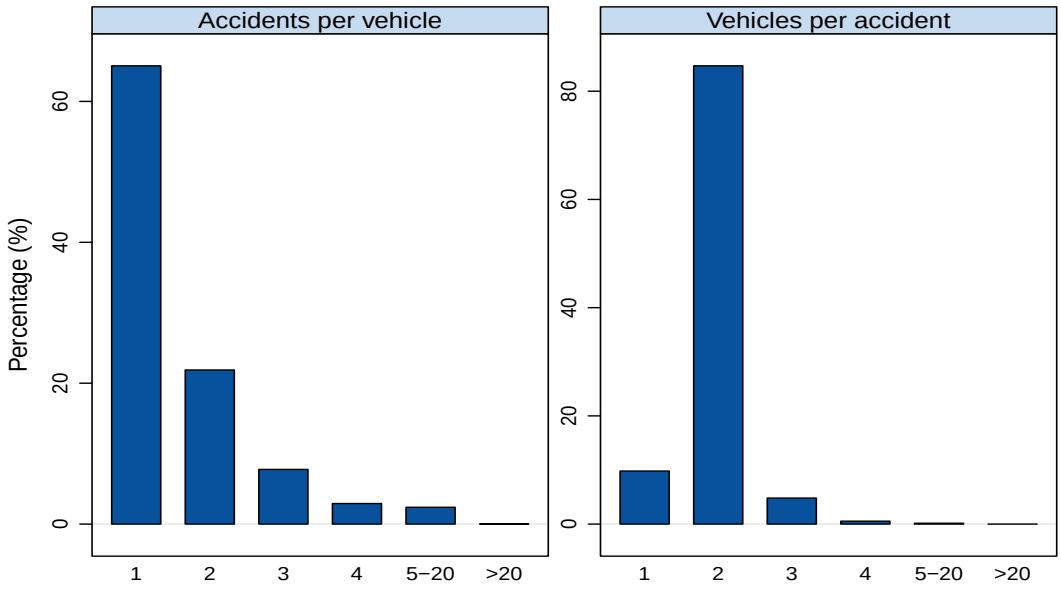

Figure 4: The left panel displays the percentage of vehicles with exactly $K = 1, 2 \ldots$ accidents. The right panel displays the percentage of accidents with exactly $K = 1, 2 \ldots$ vehicles involved. Both distributions are highly skewed (positively): most of the vehicles have few accidents, but some of them are involved in more than 20. Likewise, most of the accidents link 1 or 2 vehicles, while a small percentage of accidents involves more than 20 vehicles.

As already mentioned (see Challenge IV in the Introduction), we cannot *a priori* exclude subjects, accidents or vehicles in the tail of the distribution. The reason is that it would be an *ad-hoc* choice which would eliminate actors who are sometimes involved in frauds. On the other hand, the inclusion of highly-connected *hubs* would overshadow the relationships with (and among) those actors (drivers or witnesses) with few connecting links, who are usually directly involved in the accidents.

As remarked in Section 2.4, the heterogeneity of node degrees also reverberates in the statistical validation of the co-occurrences n_{ij}. This justifies the use of a probability model, such as the hypergeometric one, which assesses the over-representation of a given co-occurrence *conditionally* to the marginal distributions of the node degrees n_i and n_j.

Heterogeneity is a characteristic of scale-free distributions whose power-law behavior is in the tails, with exponent $2 \leq \gamma \leq 3$, gives rise to divergent moments (second and higher moments). In real networks, however, moments are always finite. But they can be very high, making them of little practical use. For a thorough discussion see, for example, Newman (2010).

It is of marginal interest for the analysis carried out in this paper to estimate with high precision the exponent of the power-law degree distribution of our networks. Although we

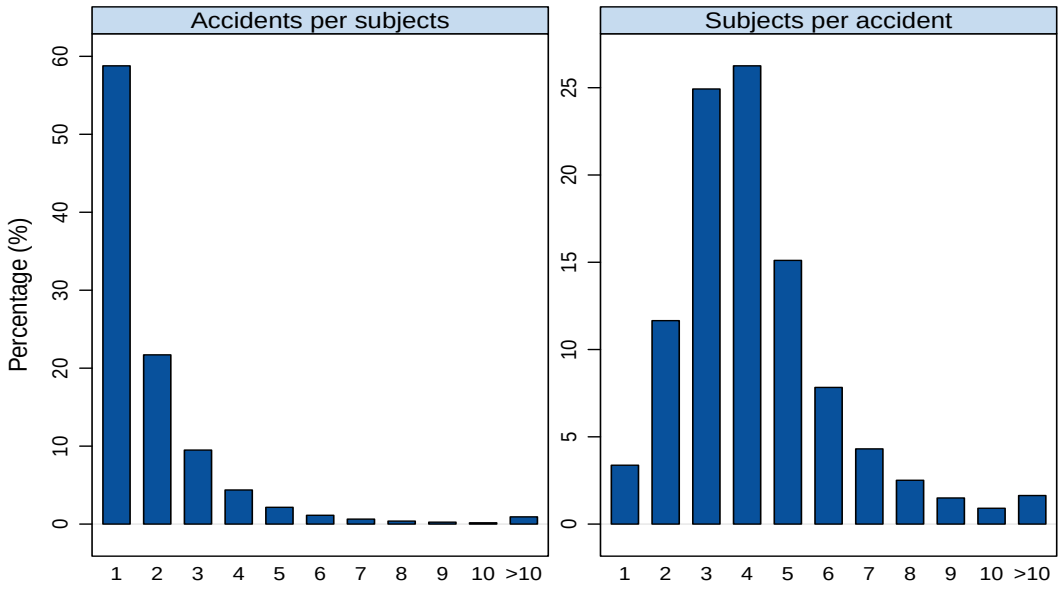

Figure 5: The left panel displays the percentage of subjects involved in exactly $K = 1, 2 \ldots$ accidents. The right panel displays the percentage of accidents in which $K = 1, 2 \ldots$ subjects are involved. The two distributions maintain the same traits of the Accidents/Vehicles distributions, but with a smoother reduction in the percentage of the events when K becomes large.

are aware that a graphical approach can lead to false conclusions about the true nature of the phenomenon, we look at the log-log chart of the complementary cumulative distribution function (CCDF) of the degrees distribution,

$$F(d) = P\left(\deg(v) \geq d\right), \tag{16}$$

and we focus our attention on the slope of the line fitting the tail of the CCDF for high values of the degree. In Figure 6, we display the CCDF of the two networks described above. The purple bullet dots indicate the data points of the distributions used to estimate the slope of the tails. As is well known, if a distribution follows a power law with degree γ, its CCDF also obeys a power law with degree $\gamma - 1$.

The estimates of γ range between 2 and 5 showing different degrees of heterogeneity. In particular, the panels on the left (top and bottom) concerns the *subjects vs accidents* bipartite network with a more pronounced power-law behaviour for the degree of the subjects side of the network. Note that, the order of magnitude in the number of accidents *per* subject spans between 1 and 5, and "node hubs" with more than 10,000 accidents attributable to professionals operating in the insurance sector.

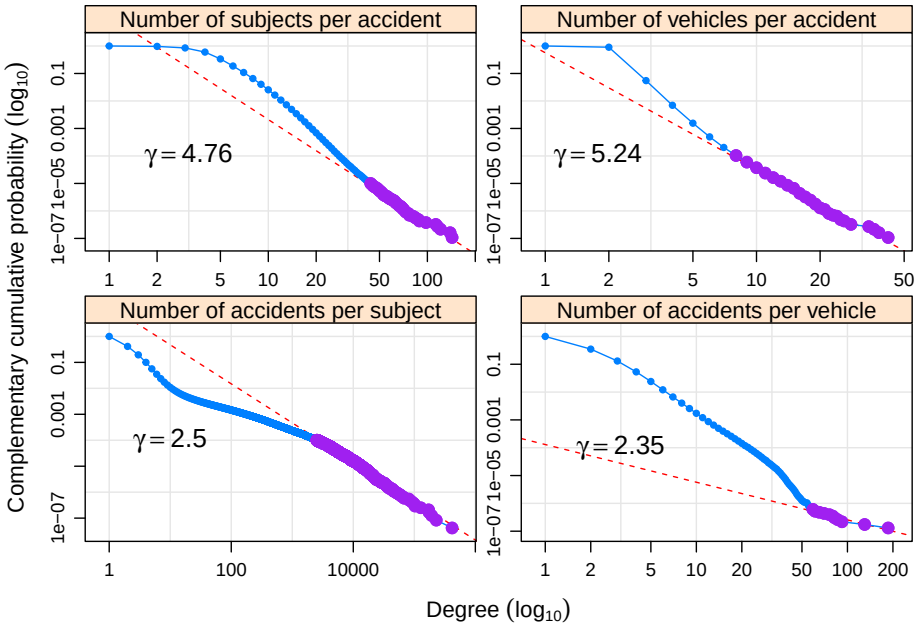

Figure 6: Complementary cumulative distribution function (CCDF–light blue) and estimate of the power-law exponent (dashed red line) on the tail of the distribution (purple). The network heterogeneity springs from the accidents side of the networks, where $\gamma < 3$ and there are "node hubs" with more than 10,000 accidents.

The panels on the right, concerning the *vehicles vs accidents* bipartite network, show similar forms of behavior. However, observe that, although $\gamma < 3$ for the degree of the vehicles, the heterogeneity between the two sides of the network is more limited due to the absence of "node hubs" with huge number of links.

The take-home message is: these kinds of empirical characteristics suggest that when filtering the network to find anomalous patterns, only the heterogeneity of the subject side has to be taken into account. A noteworthy implication is that the hypergeometric distribution (10) is a consistent, precise theoretical method for validating node bonds. Indeed, double heterogeneity requires more complex probability models, which are often more challenging from a computational point of view (Puccio et al., 2019; Tumminello et al., 2013).

4 ISAIA: an investigation system for Antifraud integrated activity in the motor insurance sector

ISAIA (**I**nvestigation **S**ystem for **A**ntifraud **I**ntegrated **A**ctivity) is a system that implements the procedures to investigate the existence of networks of fraudsters in the motor-claims sector. The framework contains various modules whose functionalities are described in Table 2. In Figure 7,

we illustrate the key relationships among the modules and their role in the fraud assessment flow.

Module	Function	Input	Output
Data pre-processing	Analyses of data integrity to remove records that are unreasonable or anomalous. Removals of claims where either the subject or the vehicle appear in white-lists.	Raw AIA records and registers of the vehicles and subjects	A sub-sample of AIA records, nearly 90% of the initial data.
SVN builder	Builds a SVN projected network using the hypergeometric test with Bonferroni correction (see Section 2.5)	A list of co-occurrences n_{ij} between the nodes of a bipartite network	A list of nodes with their validated links
Bipartite SVN builder	Reconstructs the bipartite network with statistically-validated links	A list of nodes with their validated links	A bipartite network where each link, connecting the nodes of the two layers, is statistically significant
Community detection module	Determines the communities of a large network (see paragraph 4.2.1)	A list of nodes with their validated links	A list of communities, with each community identified by a list of nodes and their validated links
Community characterization	Identifies characters that are over expressed within a given community	A community identified by a list of nodes and their validated links	A list of prevailing characters within a given community
Community pruning	Prunes links of a given community according to its over-expressed characters (see Section 4.4)	A community identified by a list of nodes and their validated links	A list of nodes with their validated links
Dashboard of network indicators	Builds and displays network and community indicators	SVN networks and communities	A set of numerical indicators and graphical network representation

Table 2: Description of the functions, inputs and outputs of the modules prescribed by ISAIA.

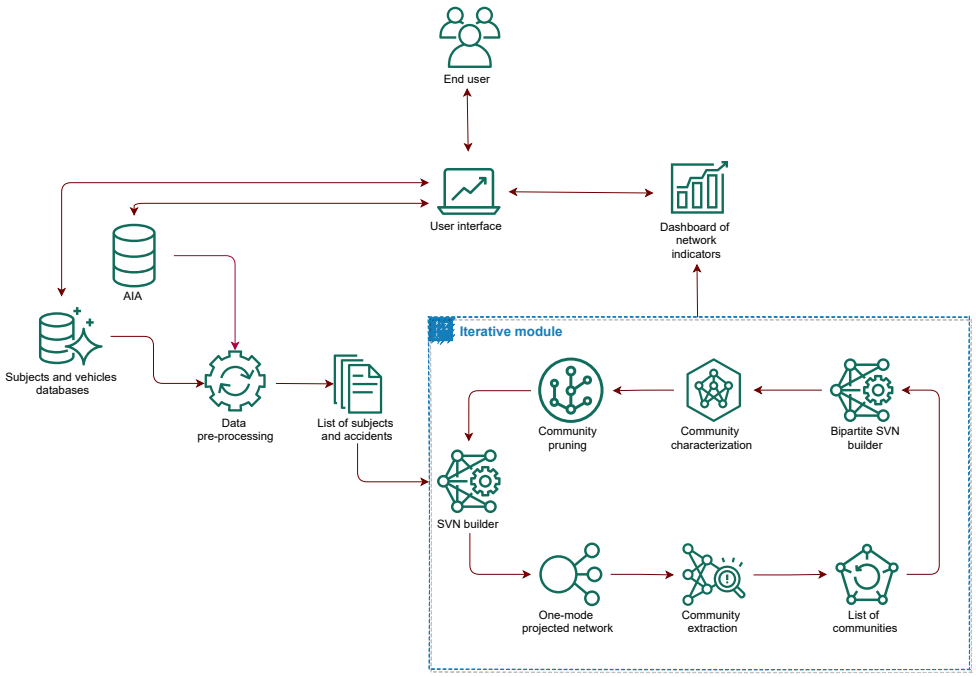

Figure 7: Map of the workflow relations among ISAIA modules.

4.1 Data pre-processing

ISAIA data pre-processing module is directed to remove two primary sources of data glitches that can produce misleading results.

One example is the presence of "subject outliers", i.e., subjects that for their role can be recipients of hundreds or thousands of accidents, but it would be inconceivable to deem them fraudsters. Another is the presence of "accident redundancies", i.e., claims deemed distinct, that actually concern the same accident and the origin of which is a loosely controlled claim input procedure.

While subject outliers can be easily removed by carefully designing "white-lists"[10] of subjects to exclude *a priori*, accident redundancies is a more subtle issue and need a special focus (see Challenge II).

One of the most common cases is claim duplication due to timestamp mistyping. In particular, at the end of each working day, new claims are added to the AIA data warehouse after a screening to assess whether they already exist. Claim records usually need to be updated (a new witness, a detailed description of the facts around the accidents, a physician report). If the operator mistypes the date, then the screening procedure files the claim as a new one.

To identify and filter out such claims, we exploit the statistical properties of the accident-

[10] As an example, white-lists usually include car hire companies, utility companies, municipalities, public transport authorities, public healthcare institutions, etc.

projected networks deriving from both the *accident vs subject* and *accident vs vehicle* networks. Both networks contain nodes (accidents) whose validated links are, respectively, the number of subjects and the number of vehicles joining them. If two accidents with a validated bond belong to both networks, and if their timestamps difference is just one digit (as measured by the Hamming distance), the two accidents are very likely identical[11]. Note that, we cannot base our decision about the integrity of the claim looking only at one projected network. Indeed, a link with two or more subjects binding two accidents, if validated, could be a fraud even if accidents very close together in time.

In general, when there is a similar accident in the two validated networks, with only a slight difference in their attributes (in the above example the timestamp), alarm bells should go off. It is likely then that these are, in fact, the same accidents repeated for administrative error.

4.2 Network community analysis

Communities are parts of the overall network that contain entities with similar features or that are closely entangled with each other.

In general, the process of structuring network communities entails two phases: *identification* and *characterization*. As illustrated in Figure 7, once communities are identified, it is critical to interpret them in terms of the characteristics shared by the elements that constitute organized groups of suspected fraudsters.

4.2.1 Community detection

Determining all the communities in a system is a very challenging matter (Challenge V). In recent years, researchers have made substantial progress in identifying communities in complex networks, and several methods have emerged to accomplish this goal (see Fortunato (2010) for a review).

In particular, the research on community detection in large bipartite complex networks is very active. Roughly speaking, there are two approaches used for community detection in bipartite networks: algorithmic approaches and model-based approaches. The former solves the problem by applying greedy searches in a heuristic way to optimize an objective function over all possible partitions of nodes; the second approach fits a generative model to the data to then assess the

[11]When two accidents involve relatively few people, we directly make sure they share at least 90% of them to validate the first check and move to the timestamp check.

relative goodness of fit. For example, Wu et al. (2022) recently showed that the modularity maximization problem can be reformulated as a spectral problem, while Yen and Larremore (2020) introduce a Bayesian non-parametric formulation of the Stochastic Block Model and a corresponding algorithm to efficiently find communities in bipartite networks: see Sun (2021) to have a rather exhaustive overview of the most recent developments in the subject.

Community detection in large networks is challenging due to the intrinsic nature of the problem. The most popular approach for community detection is modularity (Girvan and Newman, 2002). The modularity of a community is calculated as the difference between the number of links observed among community members and its expected value under the hypothesis of random connectivity (Newman and Girvan, 2004). Though we are aware that recent methods relying upon Stochastic Block Models may be more efficient than modularity optimization programs to reveal communities in large complex networks, we choose the latter approach here as a good tradeoff between efficiency and scientific consolidation.

In principle, modularity should be calculated for all possible partitions (in any number of communities) of a network's vertices. The optimal partition is that one corresponding to the maximum value of the modularity. Community detection is an NP-complete problem, and many heuristic methods have been devised to provide sub-optimal solutions in polynomial time (Fortunato, 2010; Newman and Girvan, 2004).

An other alternative approach to modularity optimization relies upon random walk processes (Rosvall and Bergstrom, 2008). However, since our network is essentially based on co-occurrence, and no information naturally flows in it, modularity optimization is more suitable. We used a combination of different heuristics, such as extreme optimization (Duch and Arenas (2005)), taboo search, etc., and introduced weak constraints on community size as well as, when appropriate, on time and geographical corrections.

In this respect, Challenge VI provides some guidelines in the search for parsimonious solutions by setting weak boundaries on the size of communities. Indeed, organized groups of fraudsters, made up of thousands of individuals, are unlikely. At the other extreme, focusing on small communities, made of two or three subjects, is costly and these kinds of investigations cannot be justified. As a rule of thumb, we set the bounds to identify communities with tens to hundreds of individuals. A 100 might appear a large number. However empirical evidence indicates that groups of fraudsters of such a dimension actually exist (Tumminello et al., 2013, 2021).

In Figure 8 we show the distribution by size of the detected communities, which is rather

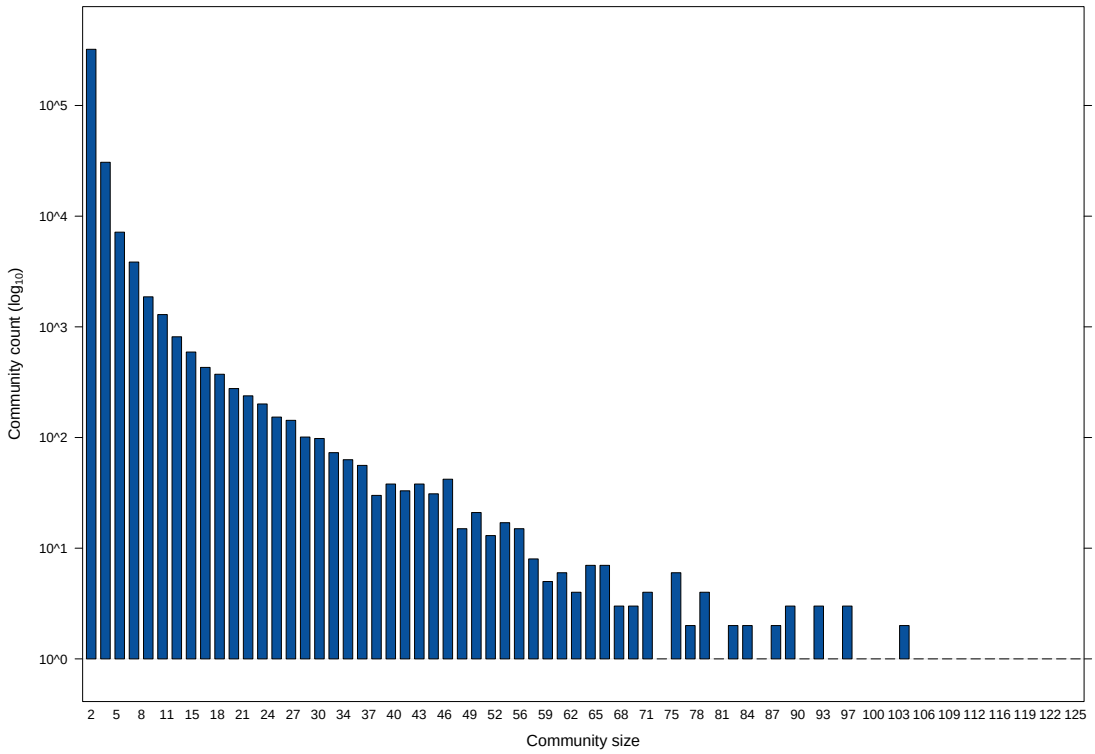

Figure 8: Distribution of the size of detected communities.

positively skewed. Communities rarely exceed one hundred vertices in size.

4.2.2 Community characterization and criminal specialization

Singling out groups in a network is not the ultimate goal. A significant issue consists in assigning to each identified group their dominant characters, i.e., those community attributes that are over-expressed compared to the whole network and other network communities. For example, communities may be over-expressed based on the role(s) that subjects of that community played in the relevant accidents, such as lawyer, doctor, car adjuster, witness and pedestrian. Such an analysis is relevant from a criminological point of view, since it provides information about the criminal specializations and the typical behavior of (suspected) fraudsters.

Indeed, the proposed system aims to identify organized groups of fraudsters, given that organization itself influences the way in which crimes are repeatedly perpetrated (Tumminello et al., 2013). Organization requires trust and synchronization among perpetrators. Participants typically play predefined roles, which valorize their attitudes, skills, and competences. Furthermore, synchronization requires communication among fraudsters, which, they should, nevertheless, minimize for the sake of secrecy.

In short, organized frauds, as perpetrated by a given group of subjects, tend to replicate a specific scheme, in order to maximize effectiveness and to reduce risks. Therefore, learning about over-represented roles in a community helps to identify the criminal strategy of members, as well as the key roles played by some of those members.

Also, communities may present an over-expression in terms of time and place where the associated accidents took place. The time and geographic localization of a community are important since they help to frame the activity of (suspected) fraudsters.

The same approach used for the validation of links (see Eq.10) is used for associating each community with one or more over-expressed attribute. In Table 3, we display some results of community characterization. For example, community 1 includes 152,906 nodes whose accidents mainly occurred around 2015-2016 and in three over-expressed regions that cover a quite wide area of the Italian territory: Sardinia (an island classified as the south of Italy), Lombardy (in the north), and Lazio (in the center). In other words, community 1 is characterized by a significant fraction of events whose attributes (years and regions) are more frequent than one should observe in a completely random case (given its size and the distribution of attributes in the population). On the other hand, community 4 is a smaller community, which does not show any over-expressed years, but its events are significantly associated with the southern region of Sicily. It is worth to note that communities reported in Table 3 are very large, as they include more than 70,000 subjects each. They are reported here not only to explain the procedure of community characterization, but also to highlight the need of the network pruning based on link robustness, which is discussed in the next sections. Indeed, all these communities, which is unrealistic to consider as organized groups of fraudsters, will break into many smaller communities after the pruning.

In general, denoting by N the number of vertices within the network, N_c the number of vertices within community c, N_ψ the number of vertices in the network that are labeled with attribute ψ, and $N_{\psi,c}$ the number of vertices of attribute ψ belonging to community c, the probability linked to $N_{\psi,c}$ is equal to Eq. 6, where $x = N_{\psi,c}$, $N_c = n_i$, and $N_\psi = n_j$. To say that an attribute ψ is over-expressed for a certain community c, we apply the hypergeometric test of Eq. 10.

Let us say that $N_{\psi,c}$ is statistically greater than what we would observe in a situation of completely uniform distribution of attributes in the system. In that case we will say that attribute ψ is over-expressed, and therefore, characterizes community c. That is, if $P(N^{obs}_{\psi,c} \geq$

Community ID	Size	Years over-expressed	Regions over-expressed
1	152,906	2015, 2016	SARDINIA, LOMBARDY, LAZIO
2	117,396	2011, 2012	CAMPANIA
3	100,036	2015, 2016	LAZIO
4	73,537	-	SICILY
5	71,974	-	EMILIA ROMAGNA

Table 3: Example of communities by size (number of nodes), over-expressed years and regions. A community is characterized by an attribute related to either years, regions, or provinces when that attribute appears in that community more than it would be if it was randomly distributed across all the events in the population.

$N_{\psi,c}^{0.05}) < 0.05$, then we will say that attribute ψ is over-expressed in community c, where $N_{\psi,c}^{obs}$ and $N_{\psi,c}^{0.05}$ are, respectively, the observed and the 95^{th} threshold quantile of the hypergeometric distribution.

In the particular situation in which communities have few vertices or where the attribute we study is rare in the system, the hypergeometric test leads to unreliable results due to its discrete nature. Therefore, in these cases we say that an attribute ψ characterizes community c if at least 90% of its nodes are of attribute ψ. [12]

4.3 Assessing the robustness of links

In the phase of construction of the subjects' SVN, for each pair of nodes we test the hypothesis of random co-occurrences. Note that one needs to pay attention and be aware of the effects that the time- and geo-localization of accidents may have on the rate of false positive links, i.e. links between subjects that are deemed potentially fraudulent, but who, in reality, were not.

This aspect is apparent, for instance, when two professionals work in the same restricted area. They could show a lot of co-occurrences because of their normal activity rather than because of any fraudulent activity. This would be especially so when they operate close to one another. As a result, these people have a higher likelihood of being involved in the same accidents, in a certain time window.

To address this issue, we introduce a robustness score (R-score) R_{ij}, computed for each validated link. Given the pair of subjects i and j,

$$R_{ij} = \log_{10} T - \log_{10} m_{ij}^{*} \tag{17}$$

[12] *Example*: community c has 3 subjects, all witnesses. The test for the value of $N_{\psi,c}$ may not be statistically significant but, since the attribute $\psi =$"*witness*" is the role of all subjects in the community, we say that attribute *witness* characterizes community c.

where T is the total number of accidents in the system regardless of the time and place of occurrence, and m^*_{ij} is the minimum value of T such that link between subjects i and j is statistically validated. Fig.s9 and 10 show the rationale behind, respectively, the computation and the distribution of the R-score.

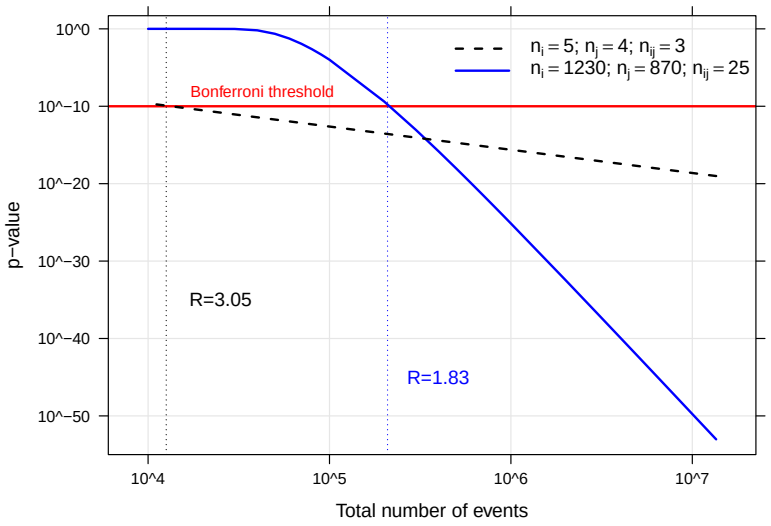

Figure 9: Rationale of the computation of the R-score for two pairs of subjects i and j with a different intensity of activity. Both the blue solid line and the dashed black line show how the statistical significance of the link connecting i and j increases (the p-value decreases) as the total number of events T in the population increases. However they represent two cases with quite different subject activity intensity: in the first case (blue solid line) subjects i and j have in common 25 events and have marginal values of, respectively, 1230 and 870; instead in the second case (black dashed line) subjects i and j have 3 events in common and had, respectively, 5 and 4 accidents. The R-scores are computed according to formula (17).

The lower m^*_{ij}, i.e. the higher R_{ij}, the more robust the link between subjects i and j will be. Once the R-score has been assigned to every link in the SVN, decision about whether they must be discarded or not comes after a community detection procedure.

4.4 Community pruning

Once a first detection of communities is completed, it is possible to associate each of these communities with a value of overall robustness (R-score):

$$R_k = \log_{10} T - \log_{10} n^*_k \tag{18}$$

where T is the total number of accidents in the system and n^*_k the number of accidents occurred in the place/s and in the year/s that characterize community k.

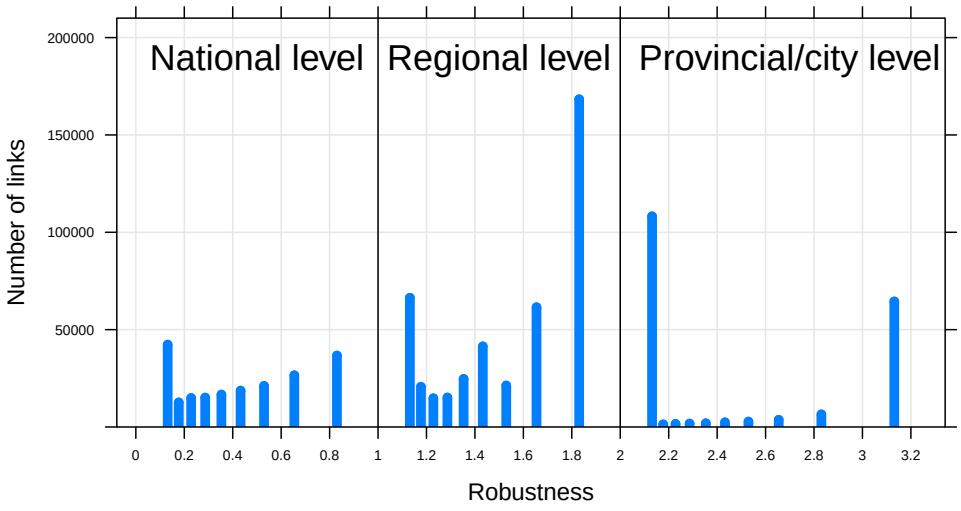

Figure 10: Distribution of the R-score: robustness values increase as the range of action of people gets smaller.

We compare the R-score (R_{ij}) in (17) of a link between a generic pair of nodes i and j with the R-score (R_k) in (18) computed for the community they belong to.

In fact, this kind of measure of the robustness of the overall infrastructure of a community is a reference point to be aware of the variability of the robustness of internal links. Ultimately, this approach provides a way to remove links that are not robust compared to other links belonging to the same community in the network.

Indeed, remembering that m^*_{ij} is the minimum value of T such that link between subjects i and j is statistically validated,

$$R_k - R_{ij} = \log_{10} \frac{T}{n^*_k} - \log_{10} \frac{T}{m^*_{ij}} = \log_{10} \frac{m^*_{ij}}{n^*_k} \quad \Rightarrow \quad 10^{R_k - R_{ij}} = \frac{m^*_{ij}}{n^*_k} \tag{19}$$

On one hand, if $m^*_{ij} < n^*_k$, then $R_k - R_{ij} < 0$ meaning that the link between i and j is very robust and should be kept within community k. On the other hand, if $m^*_{ij} > n^*_k$, then it means that the link between i and j is not validated when considering a number of accidents that exceed the number of accidents characterizing community k. It is, therefore, less robust than expected within the community. Specifically, we remove the link between nodes i and j if

$$R_k - R_{ij} > t^* \quad \forall i \neq j : \{i, j\} \in \text{community } k$$

The threshold t^* is fixed to 0.1, that is, when m^*_{ij} is about 26% greater than n^*_k. The choice

of t^* is made in order for us to be not too restrictive when deleting links from the validated network. Also, apparently there is no unique way to choose this threshold.

Eventually, community pruning will bring forth the benefit of reducing potential false positive links from the validated network. After this step is completed, the community detection algorithm used before is again performed to find the new community structure in the SVN, together with the characterization of its communities.

4.5 Bipartite and enlarged SVN

While the BN depicts statistically anomalous relationships between subjects or vehicles, it does not give explicit information about the accidents these subjects or vehicles were involved in. In fact, accidents usually represent the units of interest from the viewpoint of authorities to further investigation activity. Starting from the BN of subjects one can define the bipartite SVN, linking subjects to the accidents that contributed to the statistical validation of their relationships. Moreover, if we include all the subjects that were directly involved in the accidents included in the bipartite SVN, then the resulting network is referred to as the enlarged bipartite SVN. This leads to an average increase of two people *per* person.

The approach used for the construction of the BN of subjects, aimed at the detection of anomalous relationships between subjects, can be replicated to filter the bipartite network vehicles-accidents in order to detect anomalous relationships between vehicles.

Unlike that of subjects, the BN of vehicles is much less structured as in general a vehicle is linked to a limited number of subjects (see Tab. 4). Therefore, community detection and the correction for time-space localization are not needed in this case and the focus is given to small highly connected components.

	Nodes	Links	Connected Components (CC)	Dimension of the biggest CC
BN of subjects	2,016,505	1,919,897	638,878	651,267
BN of vehicles	112,771	61,311	54,563	12

Table 4: Dimension of SVN of subjects and SVN of vehicles.

The information carried by the BN of vehicles can be usefully integrated with that of BN of subjects. Its complementary inclusion in the detection of fraud activity will allow for a more integrated and complete set of knowledge on the complex linkages involved in the system.

4.6 Network structure and properties

The network structure emerging from the validated relationships among accidents and subjects (i.e. macro information of the system) improves the quality and efficacy of the antifraud activity performed by IVASS, which had always been based on accidents individually considered (i.e. exploiting only the information at the micro level)[13].

Table 5 shows a set of network measures and the motivation behind their use. Integrated indicators that summarize the network information and highlight possible anomalies are discussed in the Appendix B

Relying on the data stored in AIA at the end of February 2019, the number of communities detected within the BN reached 488,362. Here we set the univariate level of statistical significance at $\alpha = 0.01$ and use the Bonferroni correction for multiple hypothesis tests to be very conservative on the control of false positive links in the validated network. About 60.2% of these communities is made up of only four nodes (two subjects and two accidents), while about 9,767 communities (the highest 2%) has a number of nodes between 26 and 13,778.

In Table 6 we display the number of communities belonging to each category according to the macro-groups formed based on the characterization of roles of subjects and time/space localizations.[14].

Further analyses based on machine learning algorithms provide a better discrimination of the fraudster elements in a network (see Appendix A).

5 Three case studies of detected communities of fraudsters

This paragraph remarks the positive impact ISAIA has on the fraud detection activity performed by the IVASS. Specifically, we illustrate three empirical case studies of fraudulent organizations, which are structurally different in terms of link formation, the nature of nodes and scale.

The first case study is related to the data on three fiscal codes belonging to three out of

[13]SVN approach has been adapted to analyze criminal data already in ref. [Tumminello et al. 2013], though the original bipartite network was made of criminals linked to crime types, not events (i.e., accidents), like in the present case. As far as we know, the broad concept of SVN has never been applied before in the insurance sector, as well as for the sake of fraud detection. Furthermore, the reconstructed "Statistically Validated Bipartite Network" described in Table 2 represents a new type of SVN, which we devised to link significantly associated nodes (subjects or vehicles) in the SVN to the events (accidents) that determined such a significant association. The method we used to detect communities does not . Instead, it is used to Furthermore, the characterization of the communities in the "Statistically Validated Bipartite Network" is performed by controlling for the FWER, which extends the method used in [Tumminello et al. 2013] to bipartite networks.

[14]Communities characterized only by time and/or space attributes show a limited variability in the network indicators, as shown in Table 6 under column $\overline{P}$-$\overline{NP}$

Subject-level indicators	
Degree (K)	How active a subject is in organizing or participating in frauds, where s/he may be acting in either a local or more extended area.
Betweenness centrality	its value tells us about how central/marginal the role of a subject is in the network. It is important to understand whether a person bridges two or more communities, therefore plausibly being among the most active as a criminal leader.
Event-level indicators	
Degree (H)	number of subjects that are involved in a specific event: in general, the bigger the event, the higher the claim requested from insurance companies. Furthermore, a high value of the number of involved subjects indicates a high degree of complexity of the accident, which likely requires a set of coordinated fraudsters, and, therefore, makes the event suitable for further investigation.
AIA score	an integer in the interval [0; 100] that takes into account information unrelated to the network (information on events at the micro-level).
Mixed subject-event indicators	
H-K score	it summarizes the level of centrality and connectivity of subjects and events lying in the same region of the network. A high H-K score suggests a persistent level of coordinated criminal behavior, which helps to identify at once the key events for the most important subjects (e.g., according to the degree and/or the betweenness centrality).
Link-level indicators	
Robustness score	a measure of robustness of links in the statistically-validated network of subjects (see Section 4.3). This measure allows us to determine how sensitive the connections are to time and space localization. A high level of link robustness indicates that the link is unlikely to be a false positive due to the time-space localization of events, and subjects' activity.

Table 5: Motivation of the main network measures.

five components of a family. The father, who had divorced his wife, was the one claiming the insurance and the ex-wife and one of their children were organizing the fraud.

We first check for the presence of mother and son (and father) in the validated network, and after that, we observe in how many accidents they were involved. Consequently, we add all the subjects that were involved in those accidents, obtaining the enlarged bipartite BN.

	P	NP	P-NP	$\overline{P}$-$\overline{NP}$	None	Overall
# of communities	15,403	112,103	310	300,564	59,982	488,362
# accidents (average)	58.5	2.3	45.3	3	3	4.6
# subjects (average)	6.2	2.1	10.1	2.2	2.3	2.3
# links (average)	123.2	4.7	97.3	6.2	6.2	9.6

Table 6: Number of communities and average of nodes, subjects and links, according to community characterization: professional roles only (P); non-professionals only (NP); both professionals and non-professionals (P-NP); only time and/or space attributes ($\overline{P}$-$\overline{NP}$); no characterization.

Fig. 11 shows the fraudulent sub-network with accidents involving at least one member of the family, which highlights the connections between the mother, the father, their three sons (one of them three-years old), two relatives of the mother and two professionals, specifically a physician and a technical expert.

The first case study proves that ISAIA is effective with spotting quite small groups of fraudsters. Indeed, it is important to notice that the method is able to detect fraudulent organizations acting on very different scale dimensions, thanks to the SVN approach. In fact, ISAIA manages to integrate the available information with that which is not *a priori* known: two out of three children and two relatives of the mother were not initially claimed by the father. But they are there in the validated network. Moreover, six out of seven (85.7%) accidents are highly anomalous in statistical terms and one is associated with a medium level of anomaly.

The second case study consists of a network on a larger scale. The data on this case refers to nineteen fiscal codes reported by the prosecutor office of an Italian city, and it describes the fraudulent activity of people belonging to organized crime (Fig. 12). Also in this case, the integrated indicator manages to associate the majority of accidents with a high level of statistical anomaly (60% and most of them in the deepest and most connected part of the network), a 20% of accidents is associated with a medium level, and therefore the remaining 20% with a low level of statistical anomaly. Note that no accident in the validated network is associated with a null level of anomaly.

Finally, the third case study consists of a network on an even larger scale. In particular, this network connects people and accidents involving 313 car plates in the context of a legal identity theft reported by enforcement authorities. The number of events and subjects linked to the 313 plates are, respectively, 874 and 3,004. When we look at the bipartite validated network, a group of 1,313 of those subjects are involved in 88,672 car accidents, forming a total of 979

communities. One of the subjects (marked with a bigger black node in Fig. 13) is linked to the VAT number of the robbed company, covering a central position/role in the network. The integrated indicator classifies as high potential frauds the 42.2% of the accidents, while 19.4% and 38.2% are classified as having, respectively, a medium and a low level of statistical anomaly.

Therefore, starting with external information about a set of claimed subjects/accidents/car plates, and despite the relatively low proportion of subjects and accidents in the validated network (8.4% and 13.3% of respectively subjects and accidents that are in the validated network), the method is apparently able to detect frauds. It is also able to integrate them with other useful information.

5.1 Life-cycle of communities

One interesting point about the usage of ISAIA concerns the temporal domain or evolution of fraudulent communities.

We evaluated the persistence of the communities being detected in the validated network over five consecutive months, in particular from September 2019 until January 2020. The validated network grows monthly by around 2% and its communities are rather persistent over time, as can be seen from Table 7. The table shows both the Jaccard and the Szymkiewicz-Simpson coefficients to quantify the overlap of communities in two consecutive months, which are rather high, slightly above or below 90%, depending on the metrics.[15].

Month	(a) % variation	(b) Jaccard	(c) Szymkiewicz-Simpson
10/2019	2.05	0.867	0.938
11/2019	2.07	0.870	0.940
12/2019	1.92	0.872	0.940
01/2020	1.73	0.870	0.938

Table 7: Percentage variation from month t to month $t-1$ of the size of the validated network is in column (a); the Jaccard and the Szymkiewicz-Simpson coefficients are in columns (b) and (c) respectively.

The principled idea is that any community has to have a starting point, a phase of proliferation, and a progressive decline, e.g., since fraudsters are discovered. We analyzed the dynamics of the communities of fraudsters considered in this section. Fig. 14 shows the time series of the yearly average of the integrated indicator for the three communities of fraudsters reported

[15]The Jaccard coefficient is the ratio between the intersection and the union of the two sets A and B: $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$. The Szymkiewicz-Simpson coefficient is the ratio between the intersection of the two sets A and B and the minimum of their sizes: $SS(A,B) = \frac{|A \cap B|}{\min\{|A|,|B|\}}$

as case studies. The network of family members (in blue) lasts four years, starting in 2012 and ending in 2015. It is rather cohesive and every accident has a high level of statistical anomaly leading to a high average value each year of its existence. The fraudulent activity of this community suddenly stopped in 2015, following legal prosecution. The organized-criminality network (in purple) starts in 2011 and its statistical anomaly begins to slightly decrease starting from 2014. That's because the authorities prosecuted some of the subjects in this community since then, though the overall network was not dismantled. Finally, the legal identity theft network (in green) starts in 2014, and again, after about three years of activity and proliferation, its anomaly begins to decrease in 2017, when some of the people disappeared from the network, likely for the intervention of the authorities.

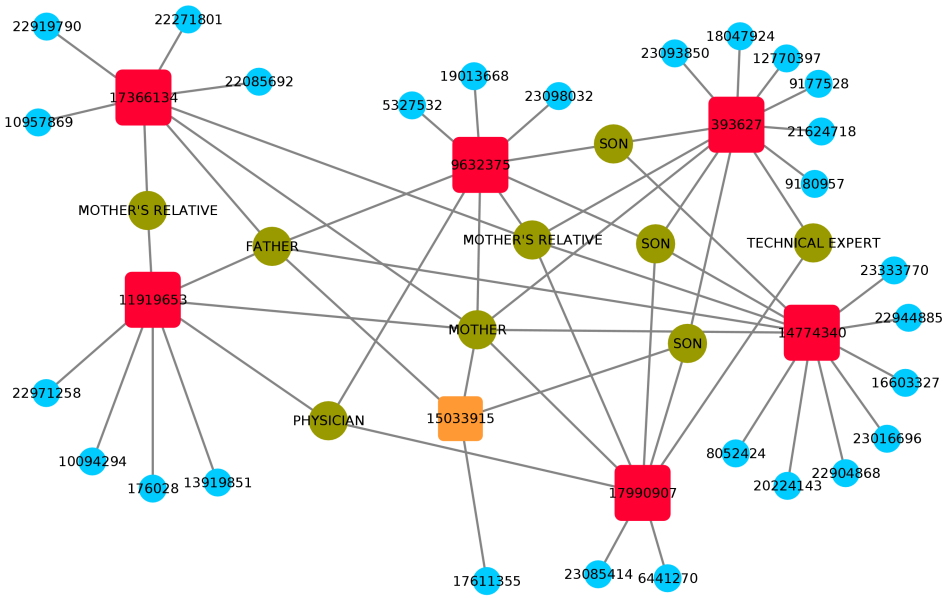

Figure 11: An enlarged bipartite sub-network, which includes accidents involving the reported fraudsters of the first case study. Rectangular nodes are accidents while circular nodes are subjects. Accidents are coloured in red if they have been assigned a "high" level of statistical anomaly according to the integrated indicator, and in orange if they have been assigned a "medium" level of anomaly. Subjects in light blue are people involved in the accidents but that are not directly included in the validated network.

6 Conclusions

In this work, we developed a novel statistical tool for the detection of communities of fraudsters that uses micro-level data of subjects and vehicles involved in the same accidents. In particular, we used a statistically-validated network approach to detect anomalous communities of subjects and vehicles in AIA, the comprehensive and exhaustive Antifraud Integrated Archive managed by the Italian Insurance Supervision Authority (IVASS).

The method proved to be very effective in uncovering anomalous patterns among subjects in the bipartite complex system *subjects-accidents* and between vehicles in the bipartite complex system *vehicles-accidents*. We construct an integrated indicator that synthesizes the information at node (micro) and network (macro) level to define a degree of statistical anomaly of car accidents, and so communities, subjects, and vehicles linked to them.

The introduction of the statistically-validated network approach improves the ability of the model to detect frauds with respect to the case where only the micro-level AIA score is considered. Based on the evidence that emerges from the new tool, IVASS can inform all the competent authorities: police and prosecutor offices. In this way fraudulent activities are restrained and the efficiency of the car insurance market in Italy is improved.

6.1 Remarks

Our methodology is general enough to be applied to similar micro-level datasets, at varying degrees of detail, as recorded in other countries. The core information required is the one that allows the system to (i) univocally associate subjects and vehicles with the accidents they were involved in and (ii) locate accidents both geographically, at least at a regional level, and in time. Our approach can also deal with incomplete datasets, e.g., the data that a single insurance company has at hand, though with some caveats. Indeed, an over-expressed co-occurrence of subjects or vehicles in the same accidents, as revealed in the incomplete dataset, already indicates a statistical anomaly. The most crucial caveat consists in the control of false negatives, that is, co-occurrences that are not statistically significant in the incomplete dataset but may be otherwise in an enlarged dataset.

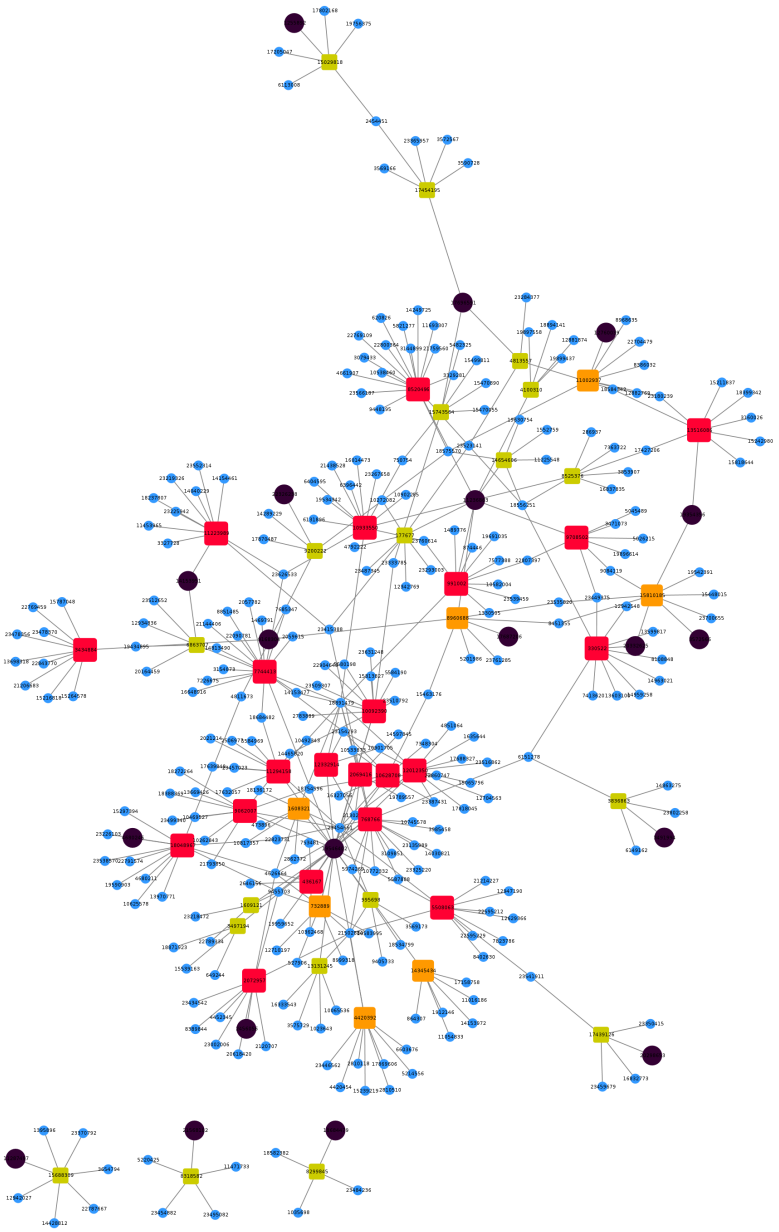

Figure 12: An enlarged bipartite validated network with accidents involving the reported fraudsters (coloured in black). Rectangular nodes are accidents while circular nodes are subjects. Accidents are coloured in red if they have been assigned a "high" level of anomaly; in orange if they have been assigned a "medium" level of anomaly; in light-green if the have been assigned a "low" level of anomaly. Subjects in light blue are people involved in the accidents who are not directly included in the validated network.

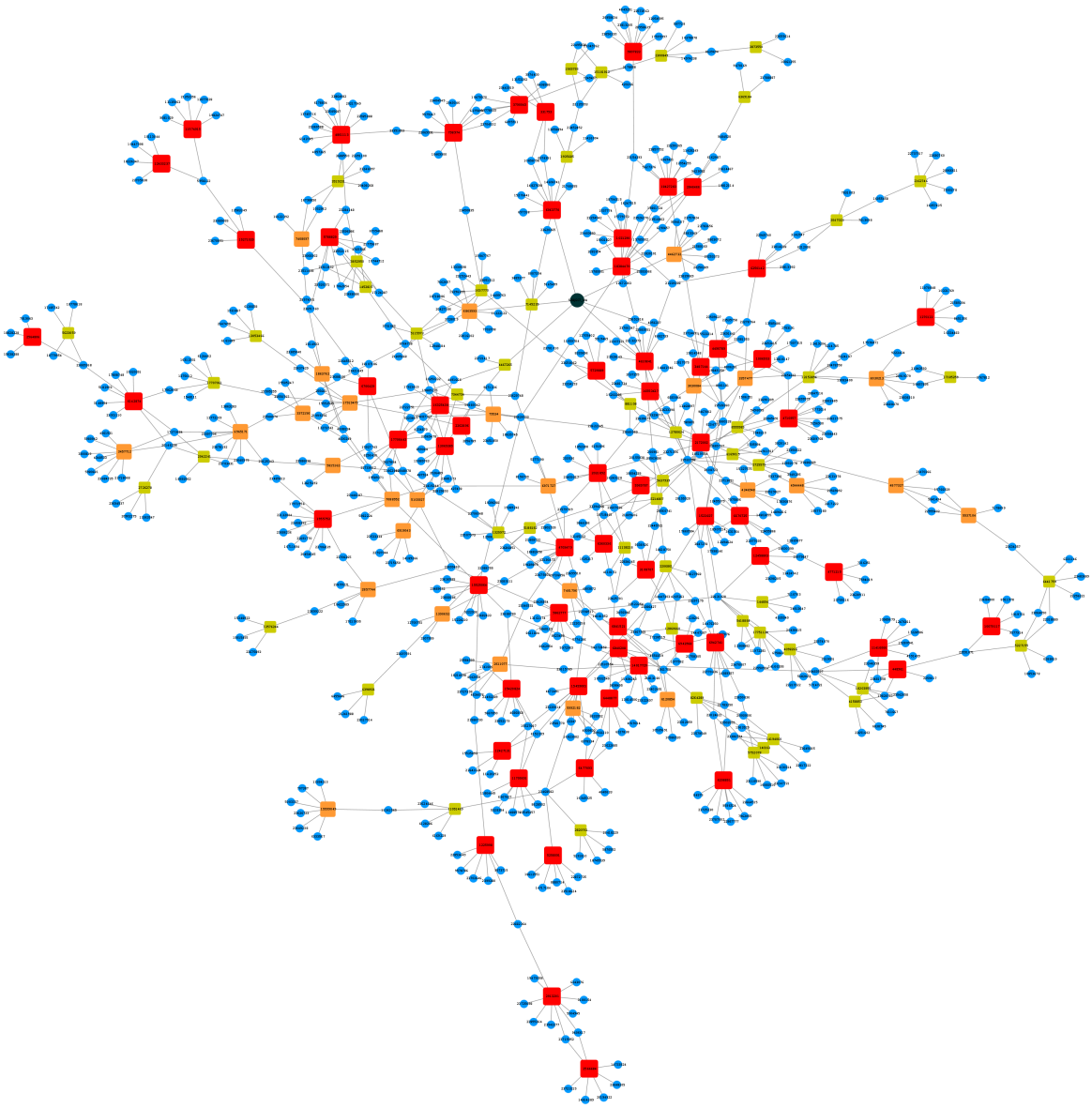

Figure 13: An enlarged bipartite validated network with accidents involving the reported car plates. Rectangular nodes are accidents while circular nodes are subjects. Accidents are in red if they have been assigned a "high" level of anomaly; in orange if they have been assigned a "medium" level of anomaly; in light-green if the have been assigned a "low" level of anomaly. Subjects in light blue are those involved in the accidents, who are not directly included in the validated network. The black node represents the subject that is linked to the VAT number of the robbed company.

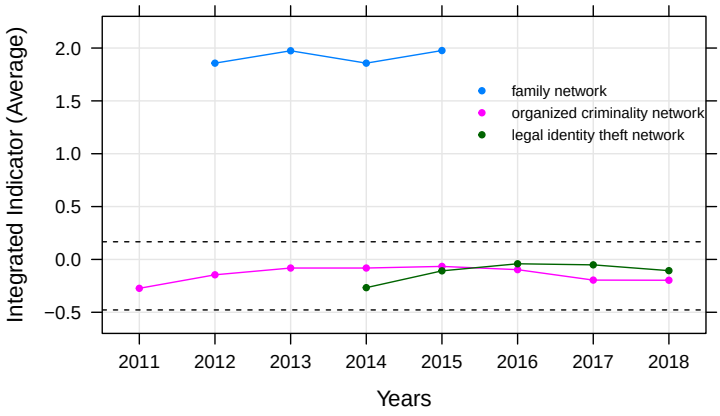

Figure 14: Yearly average values of the integrated indicator for the three case studies. The horizontal dashed black lines represent the thresholds separating low-medium and medium-high classes of statistical anomaly. The family network (in blue) lasts from 2012 to 2015. The organized criminality network (in purple) started in 2011, and its statistical anomaly decreased in 2014. The legal identity theft network (in green) started in 2014 its anomaly was slightly reduced in 2017.

6.2 Future research: testing triplets

Triadic closure is a social mechanism that lies on the more fundamental concept of homophily (see Challenge VII), which is relevant for fraud contexts (Rapoport (1953)). Indeed, triadic closure represents a simple mechanism through which fraudsters learn to work together. Let's suppose that fraudster A cooperates, separately, with fraudster B and fraudster C, and nonetheless, B and C don't even know each other. Triadic closure suggests that the presence of A as a common associate provides the *opportunity* (that B and C come to know each other), the *trust* (due to the common trust in A) and the *incentive* (A may want to perpetrate a fraud with both B and C together) not to mention the possibility that B and C become associates (in frauds). As a future research advancement, the presence of a series of frauds in which the same subjects appear as involved in triplets and triangles of cooperation should both be taken into account to spot potential frauds in car accidents.

References

A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

A. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social

network of scientific collaborations. *Physica A*, 311:590–614, 2002.

D. Bauer, J. Tyler Leverty, J. Schmit, and J. Sydnor. Symposium on insure?tech, digitalization, and big?data techniques in risk management and insurance. *Journal of Risk and Insurance*, 88(3):525–528, 2021.

E. Belhadji, G. Dionne, and F. Tarkhani. A model for the detection of insurance fraud. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 25(4):517–538, 2000.

L. Bermúdez, J. Pérez, M. Ayuso, E. Gómez, and F. Vázquez. A bayesian dichotomous model with asymmetric link for fraud in insurance. *Insurance: Mathematics and Economics*, 42(2): 779–786, 2008.

M.M. Boyer and R. Peter. Insurance fraud in a rothschild–stiglitz world. *Journal of Risk and Insurance*, 87(1):117–142, 2018.

S. Caudill, M. Ayuso, and M. Guillén. Fraud detection using a multinomial logit model with missing information. *Journal of Risk and Insurance*, 72(4):539–550, 2005.

R. Derrig. Insurance fraud. *Journal of Risk and Insurance*, 69(3):271–287, 2002.

J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72:027104, 2005.

D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

M. Girvan and M. Newman. Community structure in social and biological networks. *PNAS*, 99: 7821–7826, 2002.

C. Gomes, Z. Jin, and H. Yang. Insurance fraud detection with unsupervised deep learning. *Journal of Risk and Insurance*, 88(3):591–624, 2021.

R. Guimera, B. Uzzi, J. Spiro, and L. Amaral. Team assembly mechanisms determine collaboration network structure and team performance. *Science*, 308:697–702, 2005.

D. Kenett, M. Tumminello, A. Madi, G. Gur-Gershgoren, R. Mantegna, and E. Ben-Jacob. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PLOS ONE*, 5:12, 2010.

L. Laloux, P. Cizeau, J. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Physical Review Letters*, 83:1467–1470, 1999.

H. Li, Q. Song, and J. Su. Robust estimates of insurance misrepresentation through kernel quantile regression mixtures. *Journal of Risk and Insurance*, 88(3):625–663, 2021.

B. Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure.*, 405(2):442–451, 1975.

A. McCallum, X. Wang, and A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research*, 30:249–272, 2007.

M. McPherson, L. Smith-Lovin, and J. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27:415–444, 2001.

M. Newman. The structure of scientific collaboration networks. *PNAS*, 98:404–409, 2001.

M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69:n.026113, 2004.

J. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. Argollo de Menezes, K. Kaski1, A. Barabási, and J. Kertész. Analysis of a large-scale weighted network of one-to-one human communication. *New Journal of Physics*, 9:179, 2007.

E. Puccio, P. Vassallo, J. Piilo, and M. Tumminello. Covariance and correlation estimators in bipartite complex systems with a double heterogeneity. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(5):053404, 2019.

A. Rapoport. Spread of information through a population with socio-structural bias: I. assumption of transitivity. *The bulletin of mathematical biophysics*, 15(4):523–533, 1953.

M. Rosvall and C. Bergstrom. Maps of random walks on complex networks reveal community structure. *PNAS*, 105(4):1118–1123, 2008.

C. Song, S. Havlin, and H. Makse. Self-similarity of complex networks. *Nature*, 433:392–395, 2005.

Yidan Sun. *Bipartite Network Community Detection: Development and Survey of Algorithmic and Stochastic Block Model Based Methods*. University of California, Los Angeles, 2021.

The Association of British Insurer. UK Insurance & Long Term Savings Key Facts. `https://www.abi.org.uk/globalassets/files/publications/public/key-facts/abi_key_facts_2021.pdf`, 2021.

M. Tumminello, C. Edlin, Liljeros F., R. Mantegna, and J. Sarnecki. The phenomenology of specialization of criminal suspects. *PLOS ONE*, 8(5):1–8, 05 2013.

M. Tumminello, F. Petruzzella, C. Ferrara, and S. Miccichè. Anagraphical relationships and crime specialization within "Cosa Nostra". *Social Networks*, 64:29–41, 2021.

V. Van Vlasselaer, T. Eliassi-Rad, Akoglu L., M. Snoeck, and B. Baesens. Gotcha! network-based fraud detection for social security fraud. *Management Science*, 63(9):3090–3110, 2017.

S. Viaene and G. Dedene. Insurance fraud: Issues and challenges. *The Geneva Papers on Risk and Insurance - Issues and Practice*, 29(2):313–333, Apr 2004.

D. Watts and S. Strogatz. Collective dynamics of small-world networks. *Nature*, 393:440–442, 1998.

R.R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, 2016.

Guolin Wu, Changgui Gu, and Huijie Yang. A spectral method of modularity for community detection in bipartite networks. *Europhysics Letters*, 137(3):31001, 2022.

Tzu-Chi Yen and Daniel B Larremore. Community detection in bipartite networks with stochastic block models. *Physical Review E*, 102(3):032309, 2020.

L. Šubelj, S. Furlan, and M. Bajec. An expert system for detecting automobile insurance fraud using social network analysis. *Expert Systems with Applications*, 38(1):1039–1052, 2011.

Appendix

A Machine learning to discriminate frauds

We use machine learning to discriminate fraudulent from random events. The event features that are deemed useful for the classification problem are listed in Table 5. In particular, we extract the first four principal components found from this set of features[16] and accompany them with a dummy variable that indicates whether the accidents belong to the SVN or not. These form the set of explanatory variables of the model.

Several popular machine learning supervised algorithms could be used to deal with this kind of binary classification problem. Here we consider the logistic regression model, the support vector machine (SVM), and the random forests.

We run a ten-fold cross-validation by exploiting the information carried by a balanced sample of 4,566 random and 4,566 fraudulent accidents[17] (see Challenge III). The ROC curves of Figure A.1 show the performance of the logistic model, which works quite well in discriminating fraudulent accidents from the random ones, and this applies to both balanced and unbalanced test sets. The general performance (AUC) of three widely-used machine learning models is reported in Table A.1. The random forest reaches a slightly higher out-of-sample AUC of about 0.83-0.84, followed by the logistic regression and the support vector machine (AUC about 0.80-0.82) for both balanced and unbalanced test sets.

	Area Under the Curve (AUC) Balanced and Unbalanced test sets			
Method	4,566 vs 450	4,566 vs 4,566	4,566 vs 45,000	4,566 vs 450,000
LR	0.803 *(0.777-0.822)*	0.810 *(0.801-0.819)*	0.812 *(0.802-0.815)*	0.812 *(0.804-0.815)*
SVM	0.809 *(0.785-0.827)*	0.797 *(0.792-0.802)*	0.821 *(0.810-0.824)*	0.822 *(0.811-0.824)*
RF	0.832 *(0.810-0.855)*	0.832 *(0.828-0.838)*	0.843 *(0.838-0.849)*	0.844 *(0.839-0.850)*

Table A.1: Out-of-sample AUC through a 10-fold CV. LR=Logistic Regression; SVM=Support Vector Machine (Radial kernel); RF=Random Forest. For the RF, the optimal model is chosen based on accuracy maximization, and the optimal value of the number of variables that are randomly sampled as candidates at each split (tuning parameter *mtry*) is 2. 95% confidence intervals (DeLong) are in parentheses.

[16] We select the optimal number of components through a Random Matrix Theory (RMT) approach; Laloux et al. (1999).

[17] The R software has been used for the analysis, and in particular the functions *trainControl* and *train* of the library *caret* to perform ten-fold CV and to estimate the models.

Lastly, it is worth noting that we are aware of the fact that, in principle, random events cannot be *a priori* deemed as non-fraudulent. Nevertheless, we frame the problem to find a tool to discriminate fraudulent events from events that might present some elements of frauds with a certain unknown probability[18]. In this way we further follow our initial line for which we should avoid false positives as much as possible[19].

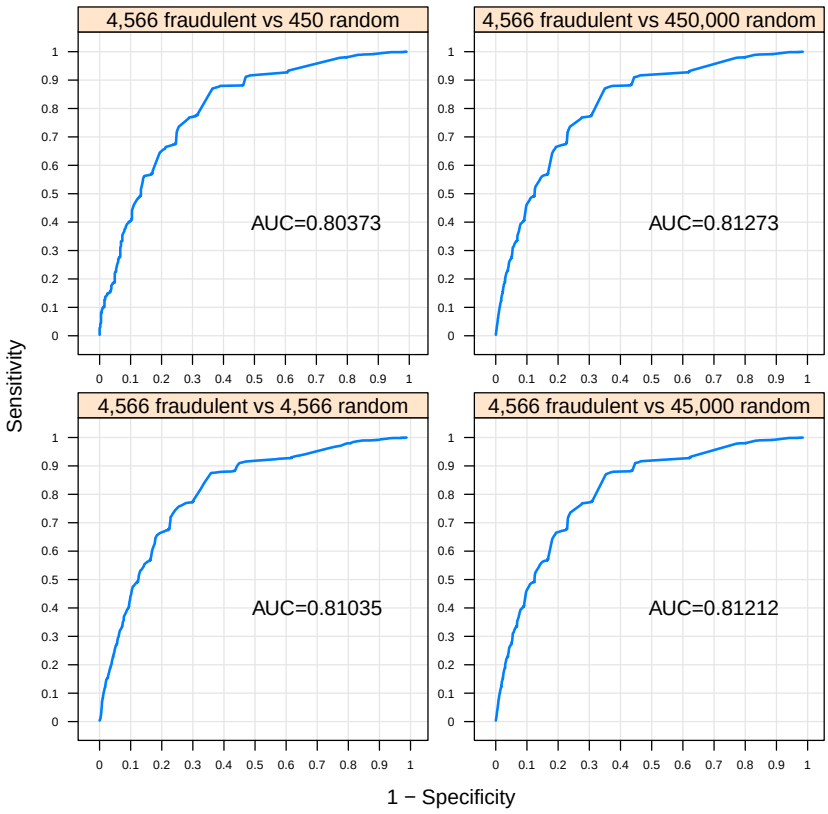

Figure A.1: ROC curves from predicted class probabilities (LR model) in balanced (bottom left) and unbalanced (top left and right panels) 10-fold CV test sets.

B Integrated indicators

Providing IVASS with an integrated indicator of the statistical anomaly of accidents is an important step towards the implementation of an effective and efficient anti-fraud task.

Indeed, on the one hand, this kind of indicator will highlight the degree of anomaly of any accident in a succinct way. On the other hand, the indicator will let the IVASS take swift action, so that the most anomalous accidents are prioritized and so that targeted investigations can be

[18] Which we could assume being close to the population counterparty.

[19] In fact, if some frauds (we would assume they reflect the proportion in the population) were contained in the control group, then the classifier would be calibrated towards the discrimination of frauds that show even stronger signs of statistical anomaly.

put in place by the competent authorities. Moreover, IVASS can proceed to communicate the integrated indicator to insurers, which will use it as a central driver to develop targeted risk management policies.

The integrated indicator is calibrated through a balanced supervised analysis, which compares a group of 4,566 accidents that were recognized as frauds by competent authorities after investigation and a group of 4,566 accidents randomly picked from AIA. These were, in turn, sampled based on an opportune stratification of AIA according to geographical and time localization reflecting that of fraudulent reported accidents. We deploy the main network measures of Table 5 to obtain a set of principal components through a spectral decomposition of the correlation matrix and a reasonable selection of eigenvalues[20].

In the activity of fraud detection, one has to face both time and cost constraints, which force to limit the actual number of anomalous accidents that will eventually be communicated to insurers, and this has to be done in a succint way. Therefore, it is reasonable to categorize the indicator into categories of risk.

In particular, we define four categories, namely *null*, *low*, *intermediate*, and *high* risk of fraudulent activity[21]. The four categories are identified through the tertiles of the sample distribution of the indicator, which is flanked by a check for the presence/absence of the accident in the validated network (see Tab. B.1). The thresholds are chosen based on the percentiles of the distribution of the integrated indicator, namely the 33^{th} percentile, that is approximately the mode of the distribution, and the 66^{th} percentile, that is approximately the value for which the Matthews Correlation Coefficient (Matthews (1975)) is maximized.

Thresholds	$a \notin$ SVN	$a \in$ SVN
$X(a) \leq t_{33^{rd}}$	null	low
$t_{33^{rd}} < X(a) < t_{66^{th}}$	low	medium
$X(a) \geq t_{66^{th}}$	medium	high

Table B.1: Categories of statistical anomaly according to the value of the integrated indicator and to whether the accident a belongs to the Bipartite BN or not.

The thresholds are used to associate all accidents of the AIA with a category of statistical

[20]The number of eigenvalues to extract is chosen relying on the Random Matrix Theory (RMT): in this application we find that the first four principal components are statistically significant. Refer to Laloux et al. (1999) for a complete treatment of RMT.

[21]IVASS already used this communication strategy even before the introduction of statistically-validated networks.

anomaly. In turn, this implies that any community, subject, and vehicles related to these accidents can also be associated with a level of statistical anomaly. This is possible by applying an aggregating procedure to the scores of the group of accidents under scrutiny: for instance, one way of associating a community with a "high" statistical anomaly could be to require that the community contains a given proportion of accidents with high statistical anomalies.

Carrying out an in-depth investigation of accidents is costly and time consuming. Therefore, resources should be allocated in an economic and sustainable way. Surely, bigger and more persistent groups of fraudulent perpetrators should take precedence. For example, only communities that include at least 4 accidents should be of interest. In this case, we say that a community is statistically highly anomalous when at least 66.7% of its accidents shows a high score on the integrated indicator. Also, we take into account the presence of accidents that belong to two or more communities. Indeed, these multi-community accidents are more frequently associated with a high score on the integrated indicator, 70% (175,304 out of 250,370) against a percentage of 54% characterizing the accidents belonging to only one community (1,092,222 out of 2,014,525). In total, 6.1% of communities (29,965 out of 488,362) are associated with a "high" level of statistical anomaly.