Workshop on Complex Networks in Banking and Finance (CoNBaF 2024)
Fields Institute, Toronto, June 27th 2024

Parenclitic Networks applied to credit card fraud detection

Miguel Romance del Río

Universidad Rey Juan Carlos (Spain)

Universidad Rey Juan Carlos
Instituto Tecnológico de Data, Networks and Cybersecurity

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de Recuperación, Transformación y Resiliencia

incibe_
INSTITUTO NACIONAL DE CIBERSEGURIDAD

Colaborators & Institutions

The results presented in this talk are part of some joint works with:

- David Aleja (Universidad Rey Juan Carlos, DCNC, Spain)
- Stefano Boccaletti (Inst. of Complex Systems, Florence, Italy)
- Regino Criado (Universidad Rey Juan Carlos, DCNC, Spain)
- Karan K. Hanumanthappa (Scuola Sup. Merid., Naples, Italy)
- Massimiliano Zanin (CSIC-IFISC, Spain)

They have been developed in collaboration with:

- National Institute of Cybersecurity - INCIBE (León, Spain),
- Rey Juan Carlos University - URJC (Madrid, Spain)
- Technological Institute of Data, Networks and Cybersecurity - DCNC (Madrid, Spain)

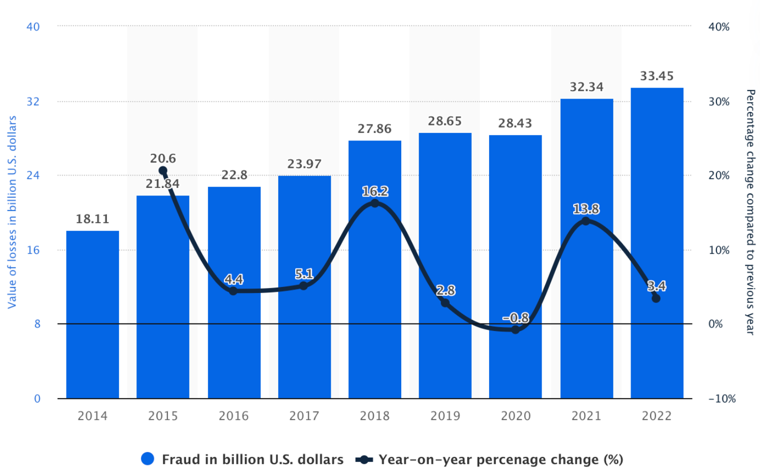
Universidad Rey Juan Carlos
Instituto Tecnológico de Data, Networks and Cybersecurity
Financiado por la Unión Europea NextGenerationEU
GOBIERNO DE ESPAÑA
MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA
SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL
Plan de Recuperación, Transformación y Resiliencia
incibe_ INSTITUTO NACIONAL DE CIBERSEGURIDAD

Roadmap

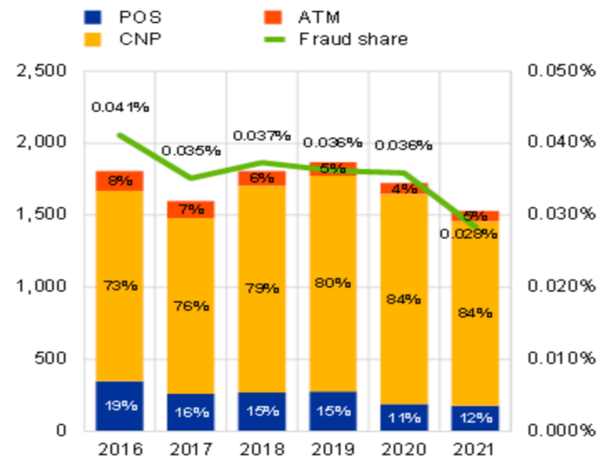
1. Setting the problem: credit card fraud detection
2. Parenclitic Networks' Analysis
   - 2.1. Original model
   - 2.2. Parenclitic networks for supervised classification
3. A real application to credit card fraud detection
   - 3.1. Description of the dataset
   - 3.2. Ingredients and workflow
   - 3.3. Some results
4. Conclusions and related works

The exponential growth in e-commerce has resulted in an increasing number of credit card frauds and therefore significant financial costs.

The exponential growth in e-commerce has resulted in an increasing number of credit card frauds and therefore significant financial costs.



NILSON REPORT



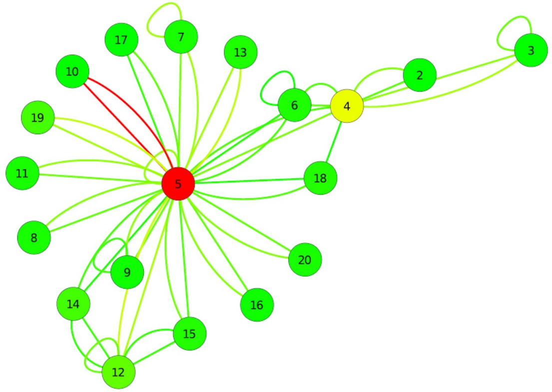
EUROPEAN CENTRAL BANK

Techniques used in order to detect such fraudulent credit card transactions include:

- Classic learning (Rule-based and known fraud patterns).
- Machine learning.

Question

Can we use complex networks to detect credit card frauds?



2. Parenclitic networks' Analysis

Parenclitic networks allow representing time independent, scalar data sets as complex networks, as follows:

- We start from a dataset D with a pre-labelled subset of subjects $S \subset D$.
- For every subject $x \in D$, we will construct a complex network $PN(x)$ that measures the divergence between x and the test set S.

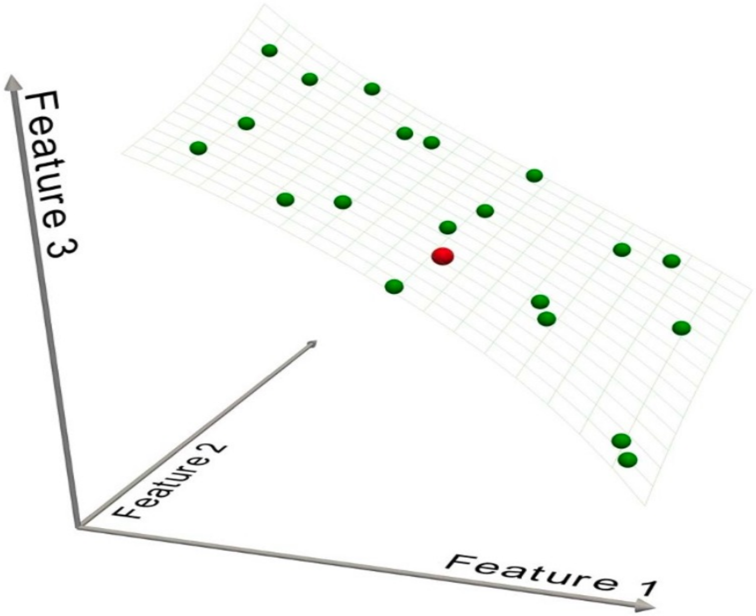
In fact the term parenclitic comes from the word $\pi\alpha\rho\varepsilon\gamma\kappa\lambda\iota\sigma\iota\varsigma$, the Greek term for 'deviation", originally used by the Greek philosopher Epicurus to designate the spontaneous and unpredictable swerving of free-falling atoms.

2. Parenclitic networks' Analysis

Given a dataset $D = \{t_1, \cdots, t_n\}$:

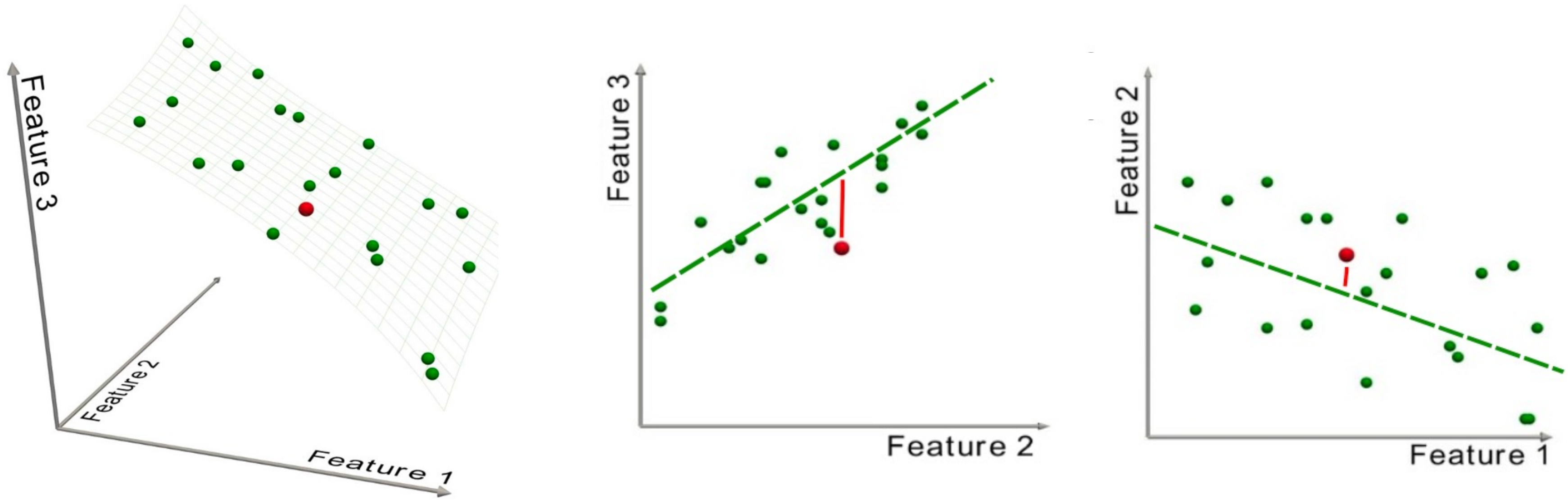
- each subjet $t_i \in D$ is a feature vector $t_i = (f_{i1}, \cdots, f_{id})$,
- we have fixed a pre-labelled subset of subjects $S \subset D$ (sane subjects).

We construct a function $H : \mathbb{R}^d \longrightarrow \mathbb{R}$ and $\delta > 0$ such that if $|H(t_i)| \leq \delta$, then $t_i \in S$.



2. Parenclitic networks' Analysis

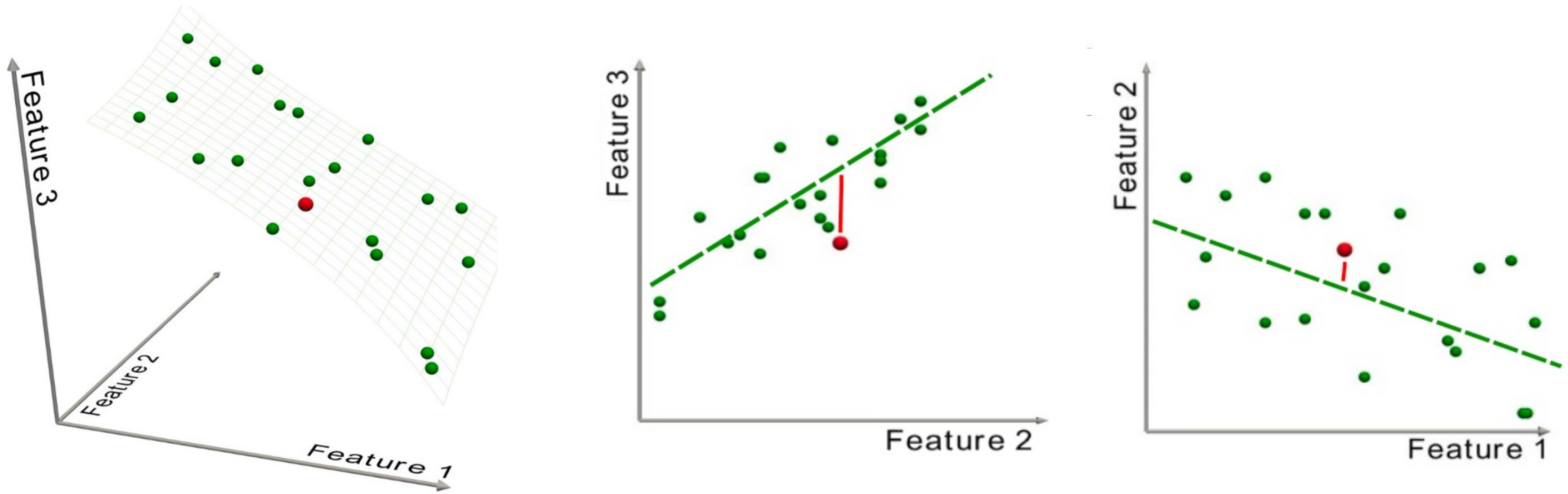
In general, computing such function $H : \mathbb{R}^d \longrightarrow \mathbb{R}$ and $\delta > 0$ is not possible, since it is a high dimensional problem, so we project the problem into $\mathbb{R}^2$.
For every $1 \leq i, j \leq d$, we consider the projection $\pi_{ij} : \mathbb{R}^d \longrightarrow \mathbb{R}^2$ and construct a function $H_{ij} : \mathbb{R} \longrightarrow \mathbb{R}$ which is approximately the (i, j)-projection of H.



2. Parenclitic networks' Analysis

In practice, functions $H_{ij} : \mathbb{R} \longrightarrow \mathbb{R}$ are constructed separately from H, and they can be computed by any classic curve-fitting technique. For example, later we will use linear regression formalism and consider

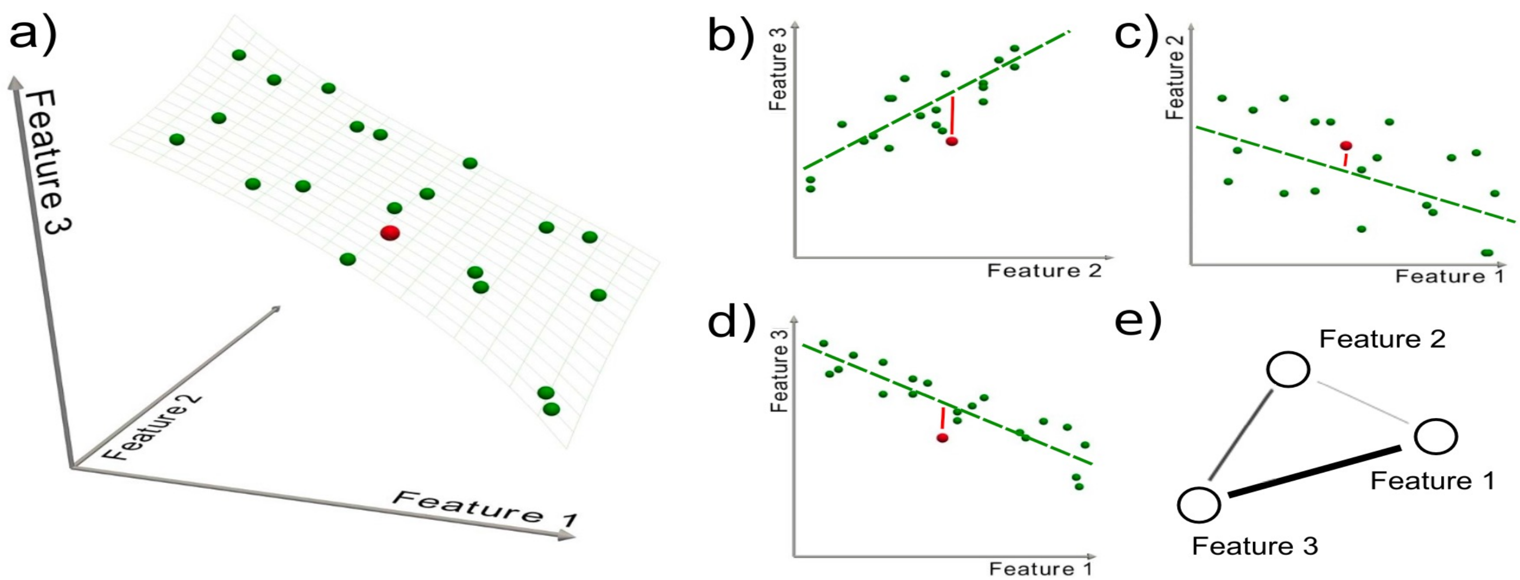
$$H_{ij}(t_i) = a_{ij} t_i + b_{ij}.$$



Definition 2.1.

For every subject $t_i = (f_{i1}, \cdots, f_{id}) \in D$, its (weighted) parenclitic network $PN(t_i) = (V, E)$, given by

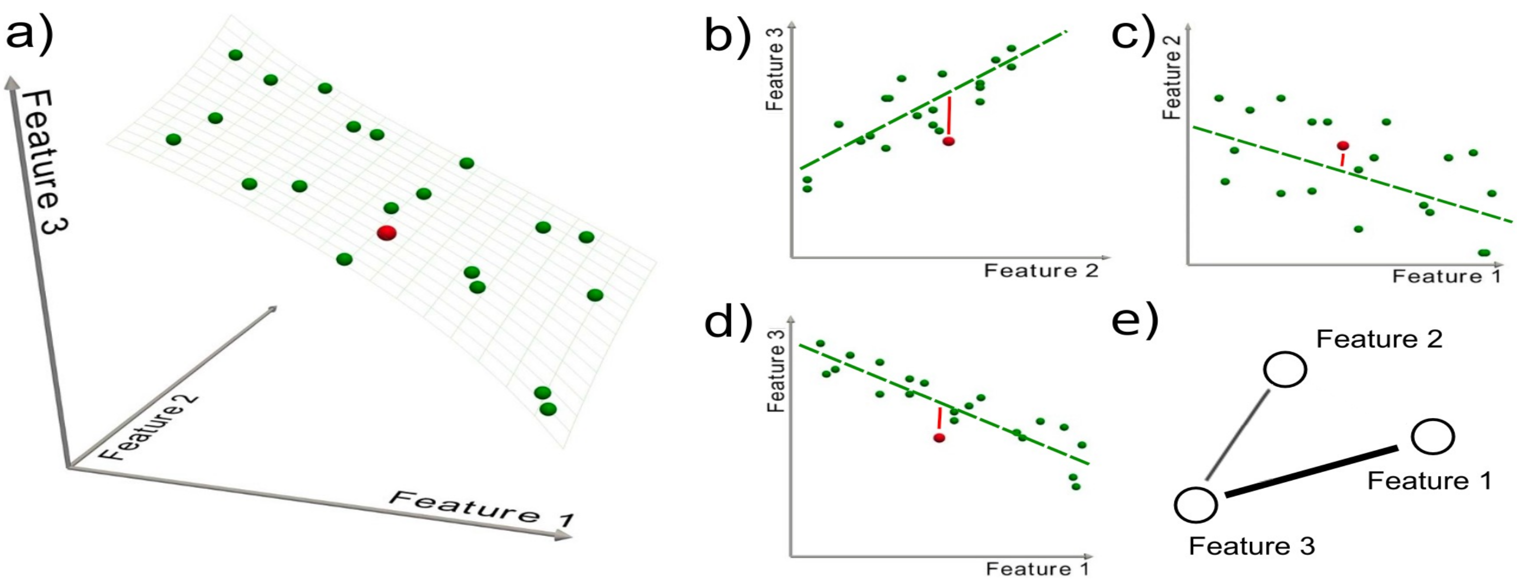
- $V = \{1, \cdots, d\}$ (i.e. nodes are the features).
- $w_{jk} = d(\pi_{jk}(t_i), H_{jk}(f_{ij}))$ (i.e. link weights are deviations from the expected features).



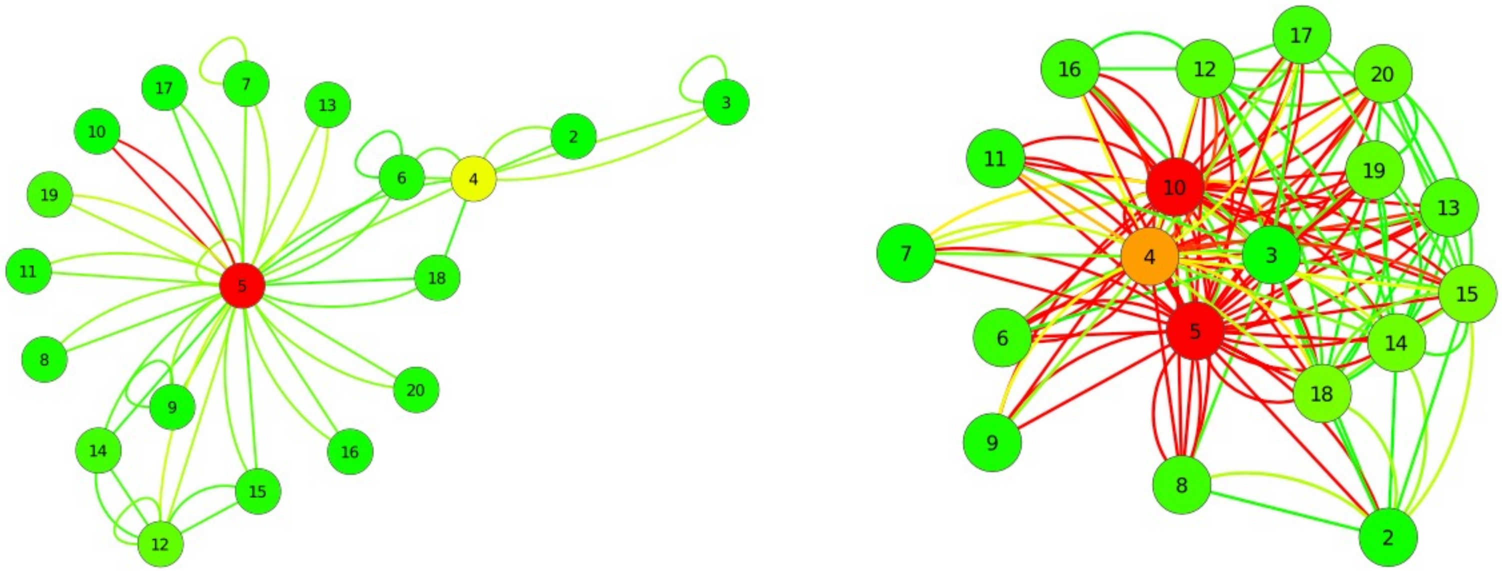
Definition 2.2.

Given a threshold $\theta \geq 0$ and a subject $t_i = (f_{i1}, \cdots, f_{id}) \in D$, its parenclitic network $PN(t_i, \theta) = (V, E)$ is given by

- $V = \{1, \cdots, d\}$ (i.e. nodes are the features).
- $\{j, k\} \in E$ if $d(\pi_{jk}(t_i), H_{jk}(f_{ij})) \geq \theta$ (i.e. links are pairs of features with abnormal correlation).



- If $\theta = 0$, then $PN(t_i, \theta)$ is a complete graph for all subject $t_i \in D$.
- If θ is big enough, then $PN(t_i, \theta)$ is an empty graph for all subject $t_i \in D$.
- for θ in-between, $PN(t_i, \theta)$ differentiates S and $D \setminus S$.



2. Parenclitic networks' Analysis

- Originally, parenclitic networks were considered for identifying critical genes in genomics.
- In particular, it was used to spot critical genes of the plant Arabidopsis thaliana responsible of the plant response under osmotic stress (Zanin2014).

Question

Can we use such parenclitic networks in order to work on supervised classification problem?

Yes, it has been used in biomedical problems (see, for example Whitwell2018 and Zhang2022).

Question

Can we use parenclitic networks for credit card fraud detection?

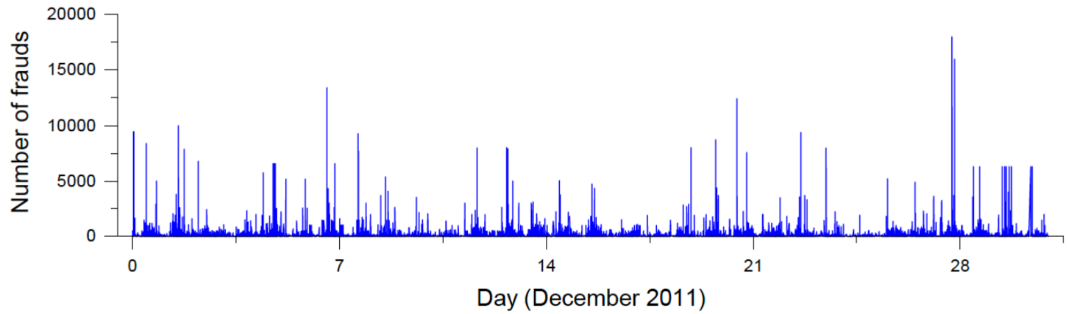
- We consider the problem of credit card fraud detection as a supervised classification problem.
- We start with a dataset such that each record is a (past) credit card transaction labeled as *"fraudulent"* or *"licit"*.
- We enrich the features of each record/transaction t_i, by adding its corresponding parenclitic network $PN(t_i, \theta)$.
- Threshold θ can be considered as a variable of the problem or as an external parameter.

We have test the proposed method in two scenarios:

- Synthetic datasets based on random models (Zanin2016).
- A real dataset obtained from a Global Spanish Bank.

Description of the dataset

- It includes all credit and debit card transactions of clients of a major Spanish bank from January 2011 to December 2012.
- Includes all in-person and virtual (on-line) card transactions.
- Each month, an average of 15 million operations were realized by 7 million cards.
- Around 40,000 transaction per moth were reported as fraud.



3. Application to fraud detection

For each transaction, the raw features considered were:

- Transaction size: Size, in Euro, of the transaction.
- Hour of the day: Hour at which the operation was realized.

3. Application to fraud detection

For each transaction, the raw features considered were:

- Transaction size: Size, in Euro, of the transaction.
- Hour of the day: Hour at which the operation was realized.

In addition to this, the following features were constructed:

- Time since last transaction: Time since the last transaction of the same card.
- Last transaction size: Size of the previous transaction by the same card.
- Average transaction size: Average size of the transactions executed by the card in the last month.
- Average time between transactions: Average time, in seconds, between consecutive transactions of the same card.
- Same shop: 1 if the shop is the same of the last transaction.
- Fraud rate: Average rate of illegal operations, for all cards, in the last 50.000 transactions.

3. Application to fraud detection

By using previous 8 features of each transaction t_i, it is observed that the parenclitic network $PN(t_i, \theta)$ of a fraudulent transaction:

- it has a star-like structure,
- is shows a high clustering coefficient (number of triangles),
- it has high efficiency,
- it exhibit a low Information Content.

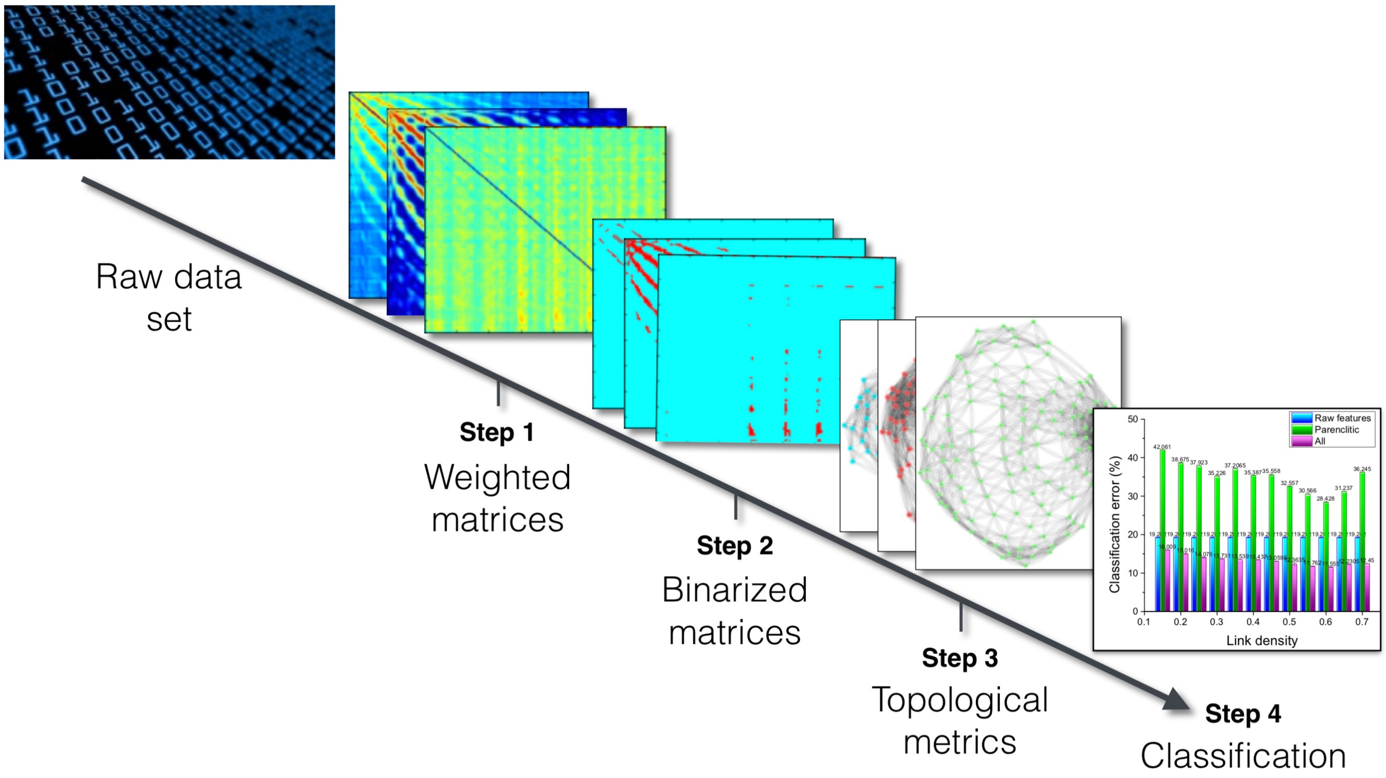
Finally, for every transaction t_i , its parenclitic network $PN(t_i, \theta)$ have been added (for a fixed θ). For this step, there are two alternatives:

- Adding directly $PN(t_i, \theta)$.
- Adding only several structural parameters of $PN(t_i, \theta)$.

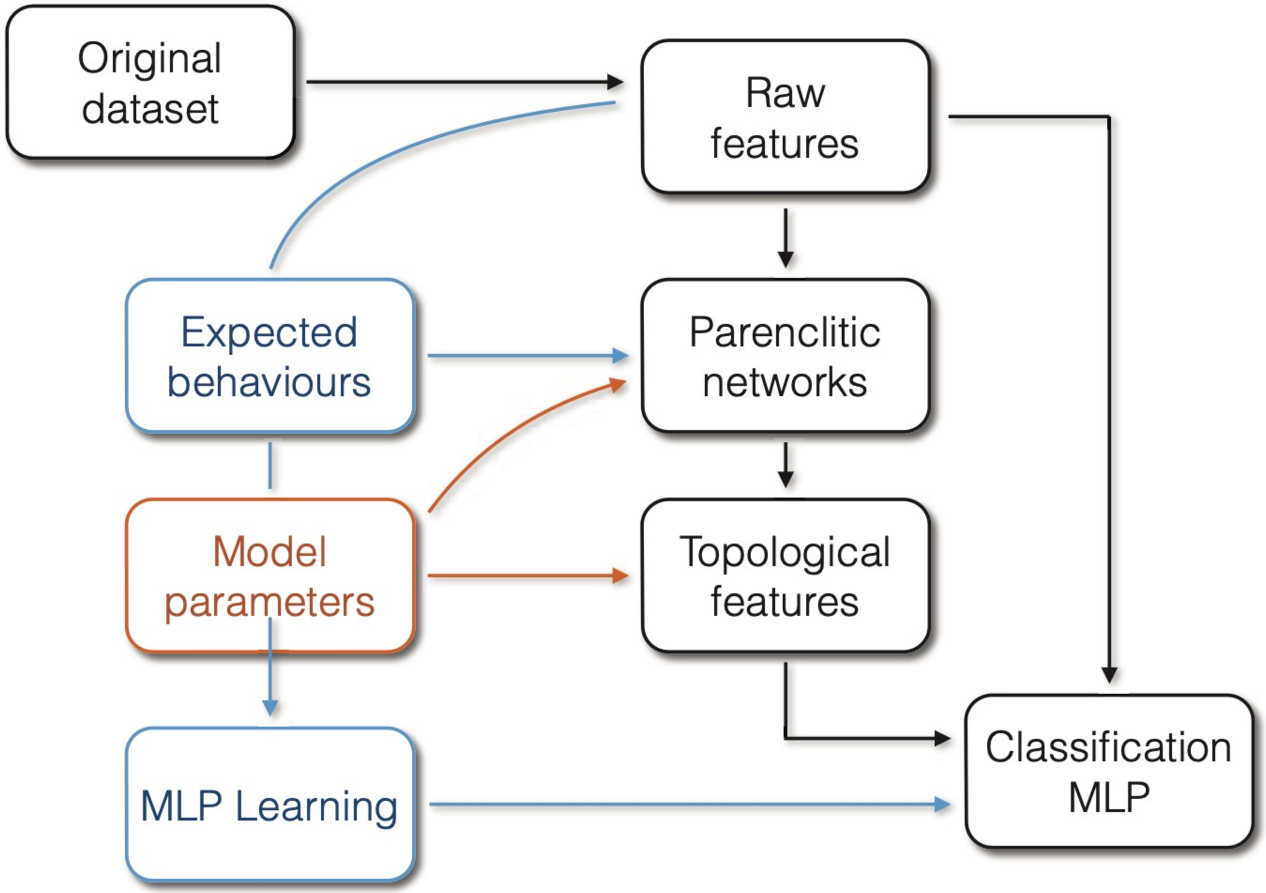
The final structural parameters considered were:

- Maximum node degree.
- Entropy of the degree distribution.
- Assortativity coefficient.
- Clustering coefficient.
- Average Geodesic distance.
- Efficiency.
- Information Content.

3. Application to fraud detection



3. Application to fraud detection



In order to analyze the contribution of parenclitic networks, we consider three different scenarios:

- Classic approach: Taking only the (8) non-parenclitic features.
- Pure-parenclitic approach: Taking only the (7) parenclitic features.
- Combined approach: Taking the (8) non-parenclitic features and the (7) parenclitic features.

3. Application to fraud detection

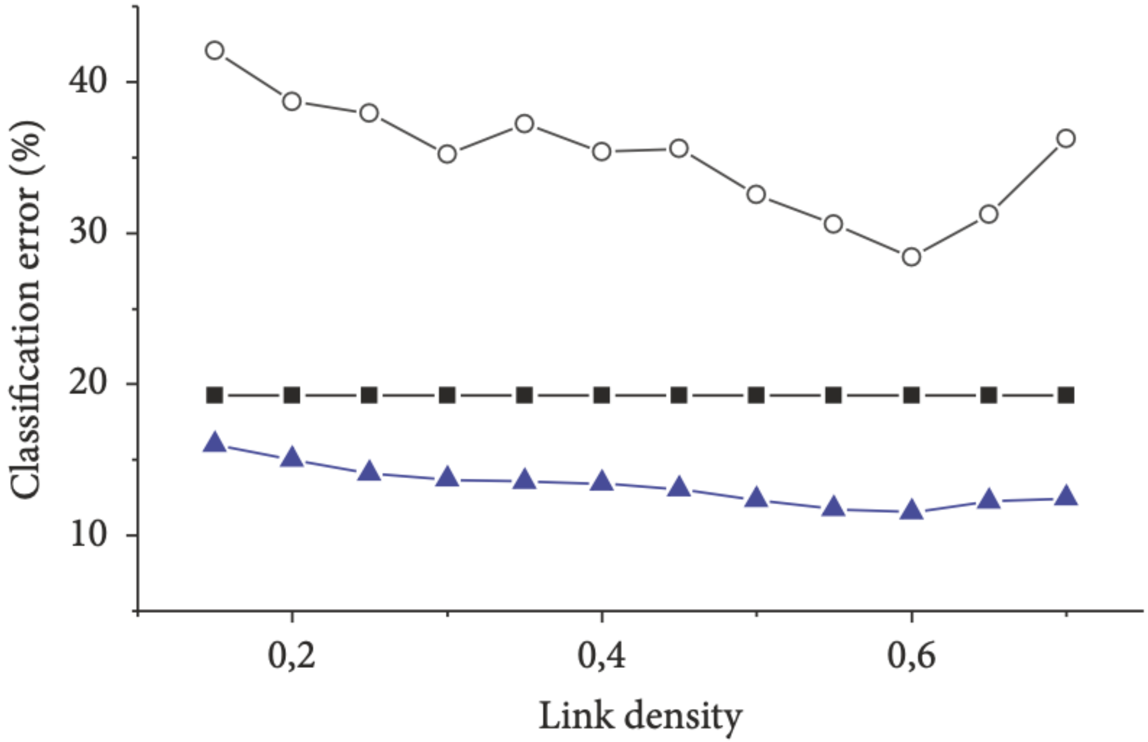
In order to analyze the contribution of parenclitic networks, we consider three different scenarios:

- Classic approach: Taking only the (8) non-parenclitic features.
- Pure-parenclitic approach: Taking only the (7) parenclitic features.
- Combined approach: Taking the (8) non-parenclitic features and the (7) parenclitic features.

For the classification model, we have tested several classic method including Random Forest, SVM and Logistic Regression, but the (best) results presented used Multi-Layer Perceptrons with sigmoid activation function

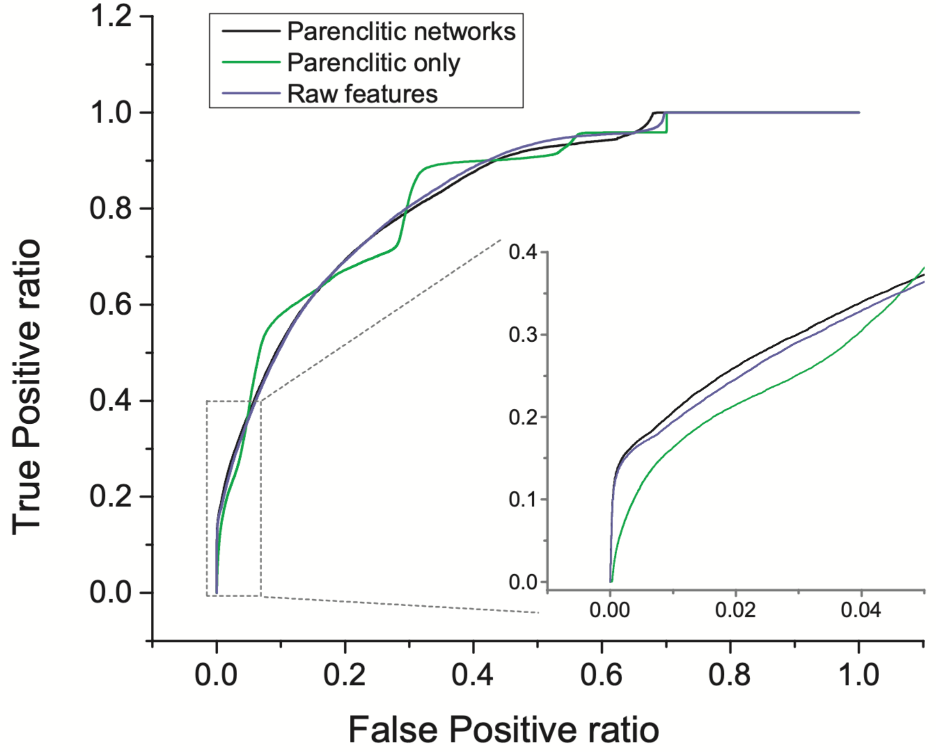
$$f(x) = \frac{1}{1 + \exp(-x)}.$$

The evolution of the classification error as a function of the considered link density (related to θ), for the classic ($\blacksquare$), pure-parenclitic ($\triangle$) and combined approach ($\bigcirc$) is the following:



3. Application to fraud detection

Similarly, the Receiver Operating Characteristic (ROC) that depicts the TPR against the FPR at various threshold settings for the classic (in blue), pure-parenclitic (in green) and combined approach (in black) is the following:



4. Conclusions and future works

- Parenclitic networks' Analysis construct a network for each entry in a dataset that measures the divergence of the relations between its features.
- They give complementary information about relationships useful for supervised classification problems.
- In particular, in credit card fraud detection, a combined approach gives better detection rates.

4. Conclusions and future works

- Parenclitic networks' Analysis construct a network for each entry in a dataset that measures the divergence of the relations between its features.
- They give complementary information about relationships useful for supervised classification problems.
- In particular, in credit card fraud detection, a combined approach gives better detection rates.

- We are also considering directly $PN(t_i, \theta)$ in the classification model (GNN).
- Higher order networks parenclitic networks are being introduced for multi-features interactions.

Some References

[1] H.J. Whitwell, et al. *"Parenclitic networks for predicting ovarian cancer"*, Oncotarget, 9(32), 22717 (2018).

[2] M. Zanin, et al. *"Parenclitic networks: uncovering new functions in biological data"*, Sci. Rep. 4, 5112 (2014).

[3] M. Zanin, M. Romance, S. Moral, and R. Criado. *"Credit card fraud detection through parenclitic network analysis"*, Complexity, 2018(1), 5764370 (2018).

[4] Zanin, et al. *"The topology of card transaction money flows"*, Physica A, 462, 134-140 (2016).

[5] H. Zhang, et al. *"Prognosis and Survival Modelling in Cirrhosis Using Parenclitic Networks"*, Front. Netw. Physiol. 2:833119 (2022).

Parenclitic Networks applied to credit card fraud detection

Miguel Romance del Río (miguel.romance@urjc.es)

Universidad Rey Juan Carlos (Spain)

Universidad Rey Juan Carlos

Instituto Tecnológico de Data, Networks and Cybersecurity

Financiado por la Unión Europea NextGenerationEU

GOBIERNO DE ESPAÑA

MINISTERIO PARA LA TRANSFORMACIÓN DIGITAL Y DE LA FUNCIÓN PÚBLICA

SECRETARÍA DE ESTADO DE DIGITALIZACIÓN E INTELIGENCIA ARTIFICIAL

Plan de Recuperación, Transformación y Resiliencia

incibe_
INSTITUTO NACIONAL DE CIBERSEGURIDAD