# Pattern Mining for Anomaly Detection in Graphs

## Application to Fraud Detection in Public Procurement

Lucas Potin[1] Rosa Figueiredo[1] Vincent Labatut[1]
Christine Largeron[2]

[1]Laboratoire Informatique d'Avignon – LIA EA 4128
{firstname.lastname}@univ-avignon.fr

[2]Laboratoire Hubert Curien – LabHC UMR 5516
christine.largeron@univ-st-etienne.fr

**ECML PKDD – Torino**
20 September 2023

LABORATOIRE INFORMATIQUE D'AVIGNON

ANR DeCoMaP

LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE

anr

AVIGNON UNIVERSITÉ

# Summary

## Public Procurement

Process by which public authorities (buyers) purchase work, goods or services from companies (winners)[1].

## Current situation

- **14%** of European Union Gross Domestic Product
- Public data available at European level (**T**enders **E**lectronic **D**aily[2])
- Over 2.1M award notices → last 10 years
- Over 75 attributes available to define a contract

[1] https://single-market-economy.ec.europa.eu/single-market/public-procurement_en

[2] https://ted.europa.eu/TED/browse/browseByMap.do

# Overview
## DeCoMaP

- **DeCoMaP**[3] **project**: Predicting fraud in public procurement.
- Project funded by the French agency
- Collection and analysis of data related to French public procurement
- Develop automatic methods for fraud detection


[3] https://decomap.univ-avignon.fr/

- Very few cases of fraud available in TED → no proper ground truth
- Current approach: using **red flags** [2]
- **Problem**: Red flags are sometimes impossible to compute

## Our assumption

We can use **relational** information to identify fraud when red flags are missing

Relation ⟶ Graph-based approach

# Problem definition

## Data

Let $\mathcal{G}$ be a collection of attributed graphs $G(V, E, \mathbf{X}, \mathbf{Y}, L)$ composed of a set of vertices $V$, a set of links $E$, a vertex attribute matrix $\mathbf{X}$, an edge attribute matrix $\mathbf{Y}$ and a graph label $L$ which can be `Anormal` or `Normal`.

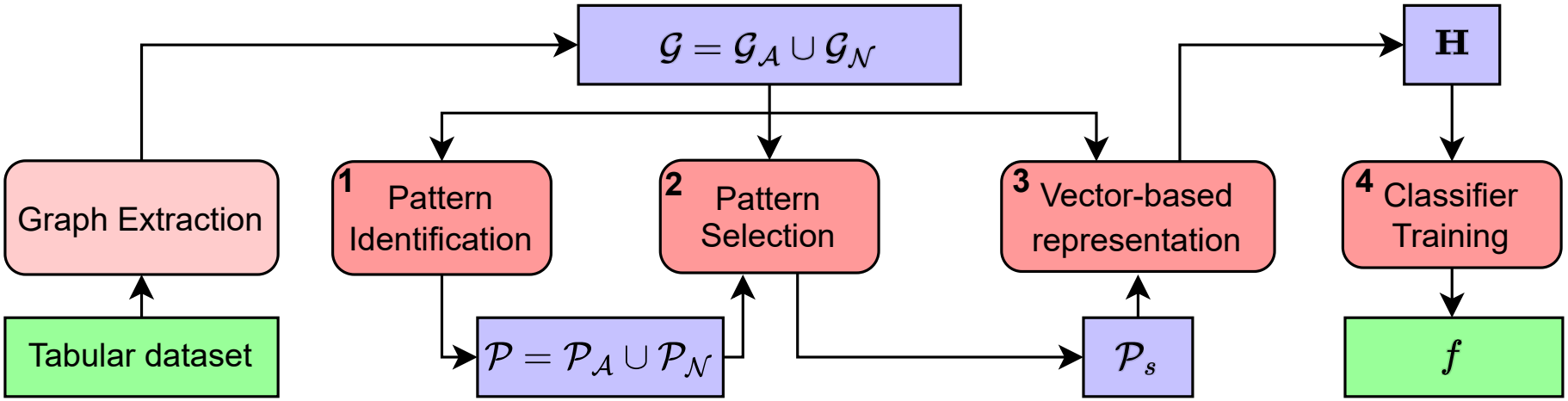
## Goal

Predicting unknown graph labels

## Main idea

Represent graphs according to their subgraphs called **patterns**.

# General architecture of our framework PANG

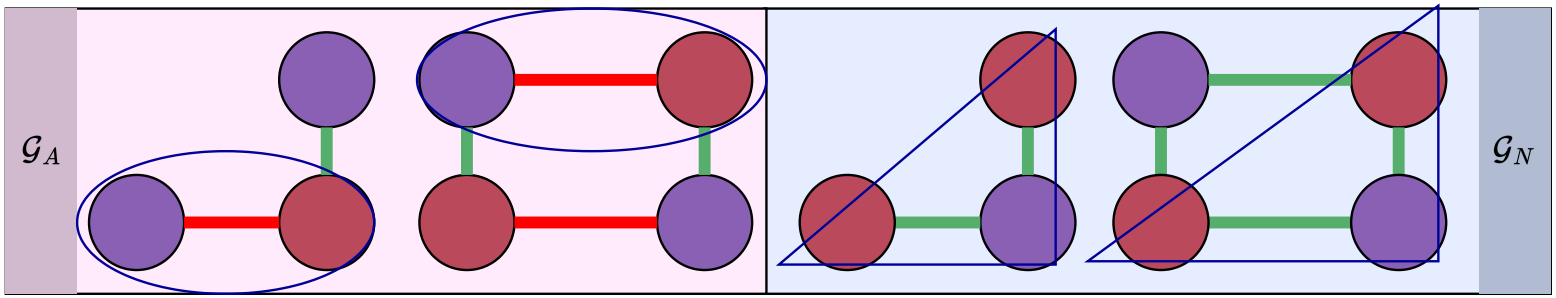


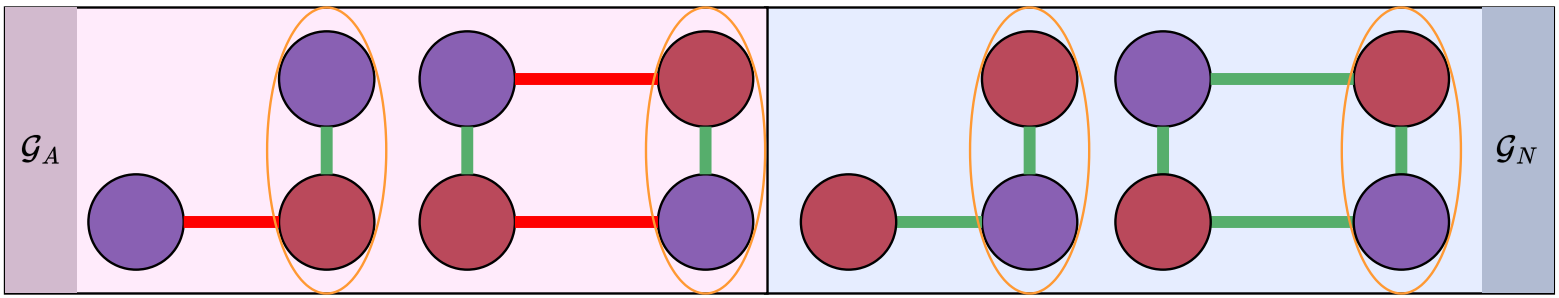

3 main parts :

- Pattern identification and selection
- Vector-based representation
- Classifier training

# Pattern Identification and Selection

What is an useful pattern for classification?




Some patterns are class-related, while others are more general

# Pattern Identification and Selection

## Pattern Identification

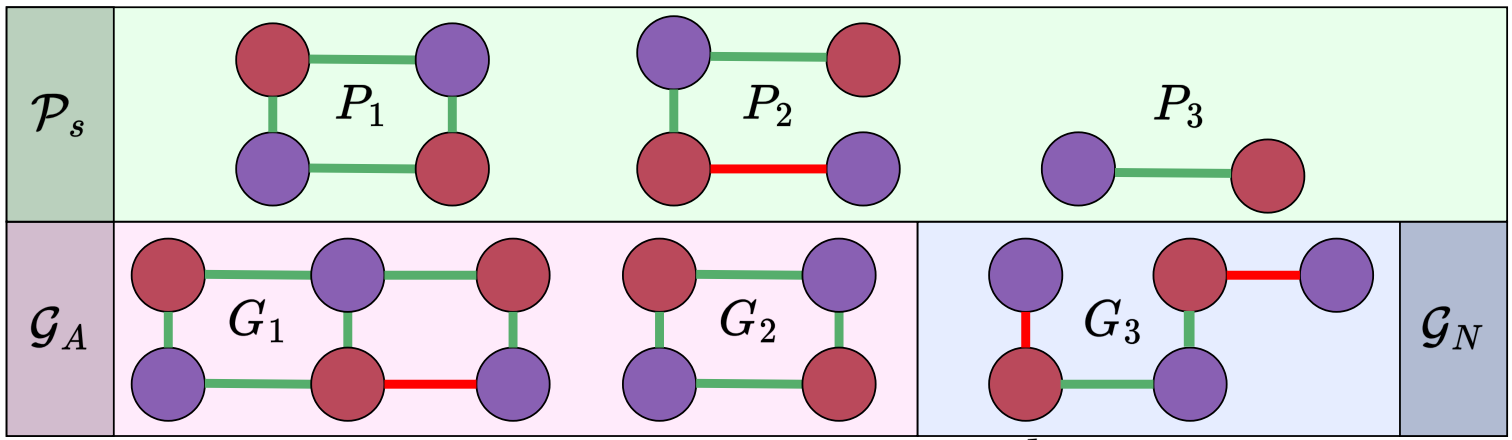
- Mine a set of candidate patterns to represent our graphs
- Choice of pattern type
  - All patterns
  - Induced patterns
  - Closed patterns

## Pattern selection

- Compute their frequency in each class $\longrightarrow$ Graph frequency
- Rank the candidates patterns $\longrightarrow$ Discrimination Score

# Vector-based representation

- Construction of a vector-based representation $h_i$ for each graph $G_i$
- Each component $h_{ij}$ in this vector is the weight of pattern $P_j$ in graph $G_i$
  - **Binary** representation ($h_.^b$): presence or absence of the pattern
  - **Integer** representation ($h_.^z$): number of subgraph isomorphism of the pattern




$$h_1^b = [1,1,1] \quad h_2^b = [1,0,1] \quad h_3^b = [0,1,1]$$
$$h_1^z = [1,3,6] \quad h_2^z = [1,0,4] \quad h_3^z = [0,2,2]$$

# Classifier training

Last step: use of a standard classifier for predicting the label $L$ using the representation $h$

| | $h$ | $L$ |
|---|---|---|
| $G_1$ | 1 1 1 | A |
| $G_2$ | 1 0 1 | A |
| $G_3$ | 0 1 1 | N |
| $G_4$ | 1 0 0 | ? |
| $G_5$ | 0 0 0 | ? |

# Evaluation setup

## Datasets

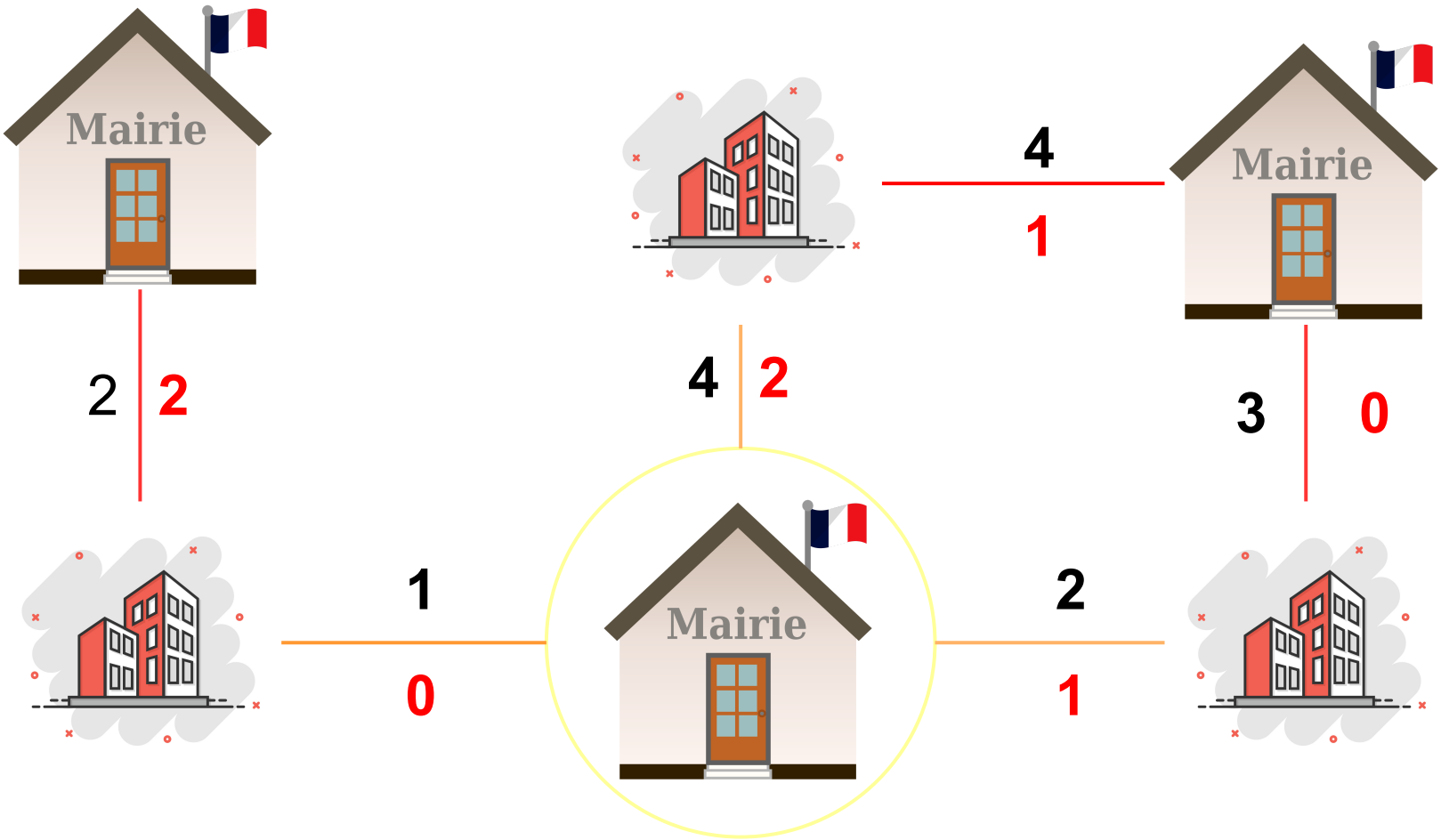
- 4 standard proteins datasets (NCI1, PTC, MUTAG, D&D)
- 1 public procurement dataset (**FOPPA**)

## Graph dataset

- Based on **FOPPA**[4], a French public procurement database [4].
- Each graph relies on a municipality
- Nodes are economic agent, edges are contracts.
- Graph label based on the number of red flags in the contracts of this municipality.

[4] https://zenodo.org/record/7808664

# Example of graph




**Anormal**

# Evaluation setup

## Baselines

- **Pattern-Based**: CORK [5]
- **Graph Kernels**: WL [1]
- **Whole graph embeddings**: Graph2Vec [3]
- **GNN**: DGCNN [6]

## Parameters

- Classifier: C-SVM[5]
- 10-fold cross-validation

[5] https://scikit-learn.org/stable/

# Results

$F$-Scores ($\pm$ standard deviation) for the *Anomalous* class.

| Representation | **MUTAG** | **NCI1** | **D&D** | **FOPPA** |
|---|---|---|---|---|
| PANG_GenBin | 0.85 (0.05) | 0.79 (0.02) | 0.77 (0.03) | 0.93 (0.02) |
| PANG_GenOcc | 0.87 (0.04) | 0.77 (0.02) | 0.75 (0.02) | 0.91 (0.03) |
| PANG_IndBin | 0.87 (0.05) | 0.79 (0.02) | 0.76 (0.03) | **0.95 (0.01)** |
| PANG_IndOcc | 0.87 (0.03) | 0.79 (0.01) | 0.75 (0.03) | 0.92 (0.02) |
| PANG_CloBin | 0.86 (0.05) | 0.78 (0.03) | 0.75 (0.03) | 0.94 (0.03) |
| PANG_CloOcc | **0.88 (0.04)** | 0.76 (0.02) | 0.71 (0.04) | 0.92 (0.02) |
| CORK | 0.66 (0.08) | 0.78 (0.02) | 0.73 (0.03) | 0.63 (0.05) |
| WL | 0.86 (0.06) | **0.83 (0.01)** | **0.82 (0.01)** | 0.90 (0.05) |
| WL_OA | 0.86 (0.06) | 0.81 (0.03) | 0.77 (0.03) | 0.90 (0.05) |
| Graph2Vec | 0.84 (0.07) | 0.82 (0.01) | 0.72 (0.03) | 0.91 (0.04) |
| DGCNN | 0.86 (0.04) | 0.74 (0.01) | 0.79 (0.01) | 0.89 (0.01) |

# Results for FOPPA

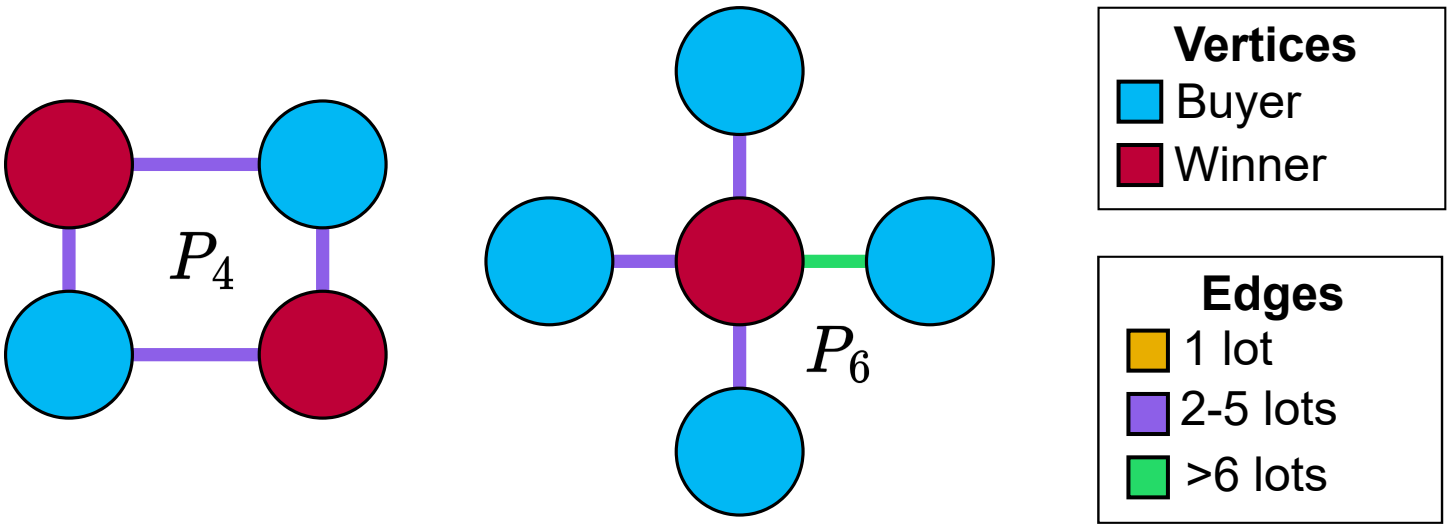
$F$-Score ($\pm$ sd) depending on parameter $s$, the size of $\mathcal{P}_s$.

| Representation Size | *Anomalous* Class | *Normal* Class |
|---:|---:|---:|
| 10 | 0.66 (0.05) | 0.73 (0.05) |
| 50 | 0.74 (0.05) | 0.77 (0.04) |
| 100 | 0.81 (0.05) | 0.83 (0.04) |
| 150 | 0.88 (0.03) | 0.88 (0.03) |
| (*all*) 15,793 | **0.93 (0.02)** | **0.93 (0.02)** |

Top 1% of the patterns $\longrightarrow$ 95% of **total performance**.

# Analysis

Benefits of our method: linking **patterns** to **economic behaviors**.



# Conclusion

## Contributions

- PANG[6]: A generic, multi-class framework for graph anomaly detection
- Application to a dataset based on public procurement
- Experimental results confirm good fraud identification in the absence of red flags

## Perspectives

- Improvement of the **discrimination** score
- Identify the **optimal** parameters of PANG

[6] https://github.com/CompNet/Pang/

# Any questions ?

AVIGNON
UNIVERSITÉ

# Bibliography I


[1] N. M. Kriege, P. L. Giscard, and R. Wilson. "On Valid Optimal Assignment Kernels and Applications to Graph Classification". In: *30th International Conference on Neural Information Processing Systems*. 2016, pp. 1623–1631. URL: https://proceedings.neurips.cc/paper_files/paper/2016/hash/0efe32849d230d7f53049ddc4a4b0c60-Abstract.html.

[2] N. Modrušan, K. Rabuzin, and L. Mršic. "Review of Public Procurement Fraud Detection Techniques Powered by Emerging Technologies". In: *International Journal of Advanced Computer Science and Applications* 12.2 (2021). DOI: 10.14569/ijacsa.2021.0120272.

[3] A. Narayanan, M. Chandramohan, R. Venkatesan, L. Chen, Y. Liu, and S. Jaiswal. "graph2vec: Learning Distributed Representations of Graphs". In: *13th International Workshop on Mining and Learning with Graphs*. 2017, p. 21. URL: https://arxiv.org/abs/1707.05005.

[4] L. Potin, V. Labatut, C. Largeron, and P. H. Morand. "FOPPA: an open database of French public procurement award notices from 2010–2020". In: *Scientific Data* 10 (2023), p. 303. DOI: 10.1038/s41597-023-02213-z.

[5] M. Thoma, H. Cheng, A. Gretton, J. Han, H. Kriegel, and A. Smola. "Near-optimal supervised feature selection among frequent subgraphs". In: *SIAM International Conference on Data Mining*. 2009, pp. 1076–1087. DOI: 10.1137/1.9781611972795.92.

# Bibliography II

[6] M. Zhang, Z. Cui, M. Neumann, and Y. Chen. "An End-to-End Deep Learning Architecture for Graph Classification". In: *AAAI Conference on Artificial Intelligence*. Vol. 32. 2018, pp. 4438–4445. DOI: 10.1609/aaai.v32i1.11782.

# Pattern Selection

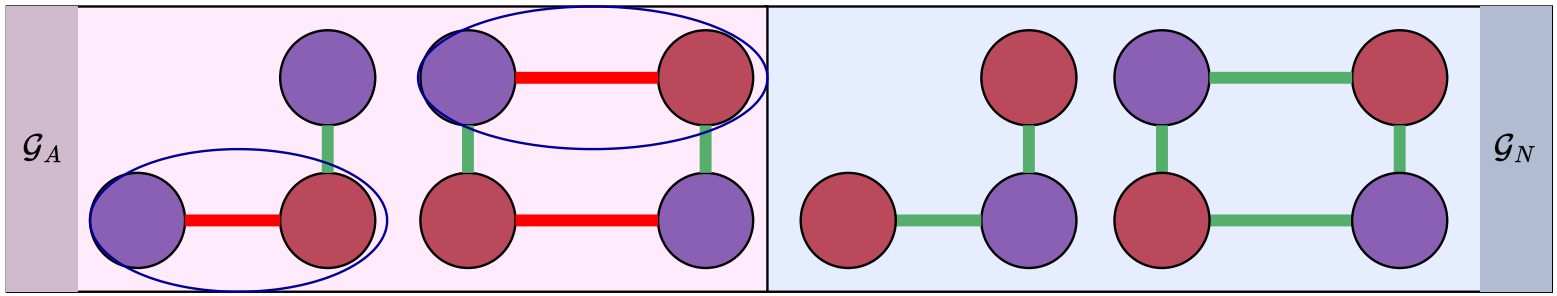
## Graph Frequency

The graph frequency $GF(P, \mathcal{G})$ of a pattern $P$ in $\mathcal{G}$ is the number of graphs in $\mathcal{G}$ having $P$ as a pattern:
$GF(P, \mathcal{G}) = |\{G \in \mathcal{G} : \exists H \subseteq G \text{ s.t. } P \cong H\}|$.

## Discrimination Score

The discrimination score of a pattern $P$ of $\mathcal{G}$ is defined as
$disc(P) = |GF(P, \mathcal{G}_A) - GF(P, \mathcal{G}_N)|$.

# Pattern Selection




## Discrimination Score

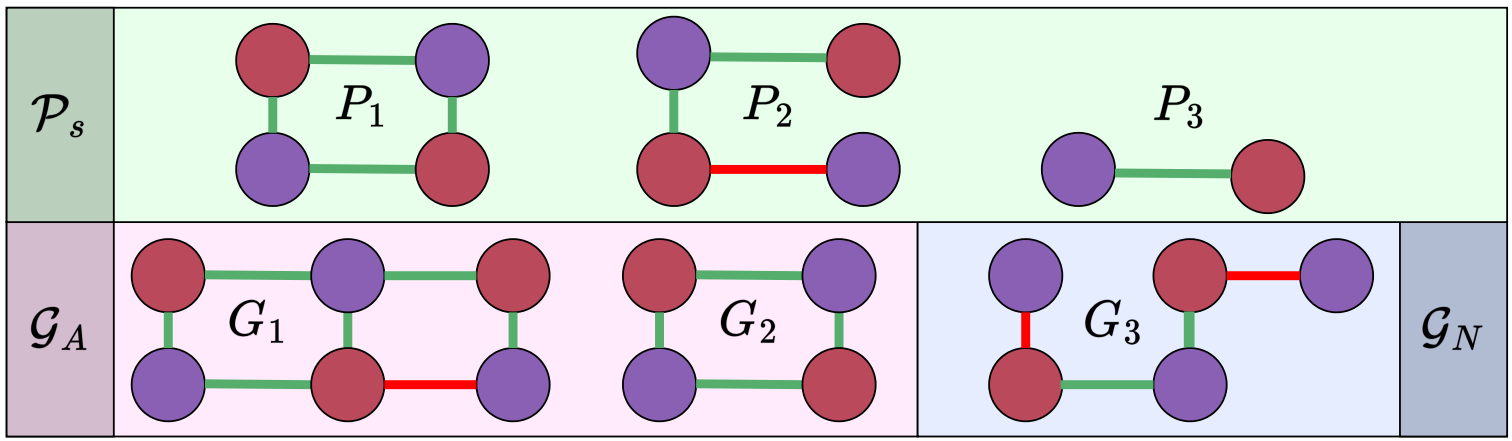
The discrimination score of a pattern $P$ of $\mathcal{G}$ is defined as $disc(P) = |GF(P, \mathcal{G}_A) - GF(P, \mathcal{G}_N)|$.

$$disc(P) = |2 - 0| = 2$$

Construction of a set of discriminant patterns ($\mathcal{P}_s$)

# Vector-based representation

- Construction of a vector for each graph.
- Each component is the weight of pattern $P_j$ in graph $G_i$:
  - **Binary** representation ($h^i$): presence or absence of the pattern
  - **Integer** representation ($h^z$): number of subgraph isomorphism of the pattern




$h^i_1 = [1,1,1]$ $h^z_1 = [1,3,6]$

$h^i_2 = [1,0,1]$ $h^z_2 = [1,0,4]$

$h^i_3 = [0,1,1]$ $h^z_3 = [0,2,2]$

]

# Classifier training

Last step : use of a standard classifier

| | h | L |
|---|---|---|
| $G_1$ | 1 0 1 1 | A |
| $G_2$ | 1 0 0 1 | A |
| $G_3$ | 0 1 0 0 | N |
| $G_4$ | 0 1 0 0 | N |
| $G_5$ | 0 1 1 0 | ? |