MOTIFS TRANSACTIONNELS

ARPI 2024 REVENU QUÉBEC

De gauche à droite: Gilles Caporossi, Kian Karimi, Mathieu Gervais-Dubé, Nicolas Goulet, Hugues-Étienne Moisan-Plante, Karine Dufresne, Abdelmouksit Sagueni.

MEMBRES DE L'ÉQUIPE – FORCES COMPLÉMENTAIRES!

RAPPEL

Objectif

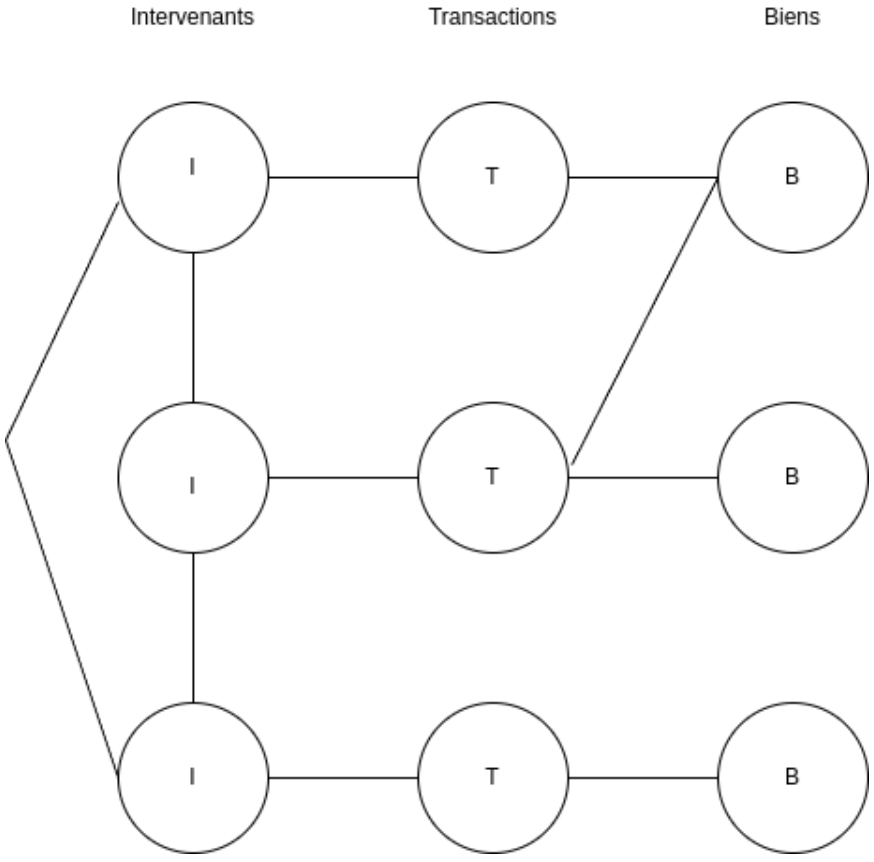
- Détection de motifs transactionnels suspicieux dans une banque de données de transactions immobilières.

Données

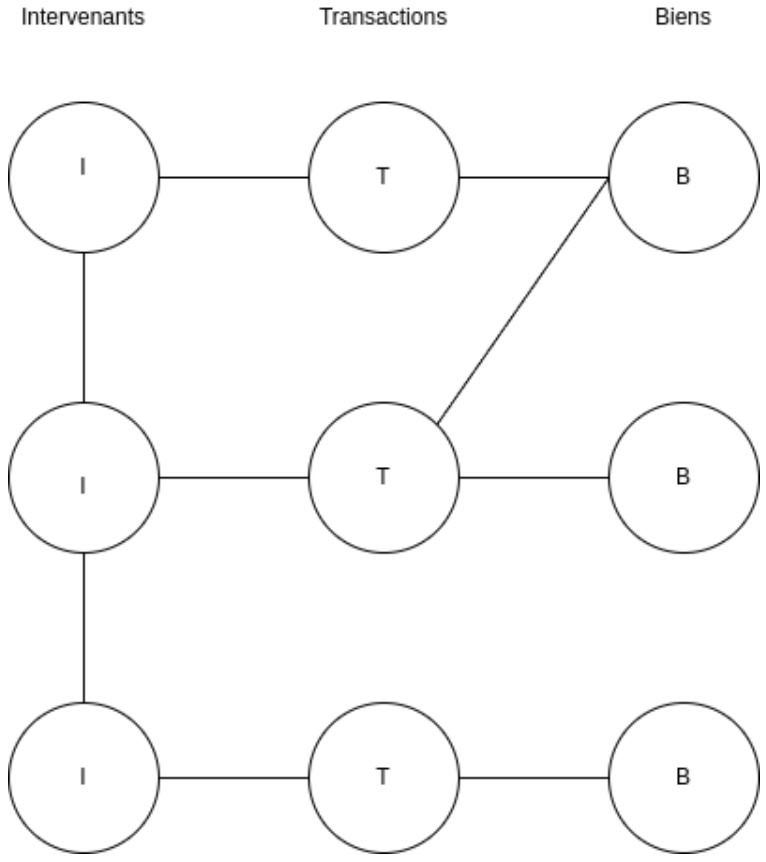
- Transactions (évaluations municipales, prix de vente actuel, intervenants impliqués).
- Relations personnelles entre les intervenants.
- Type des biens transigés, location approximative.

STRUCTURE DES DONNÉES

- Aborder les données avant le problème
- Deux structures de réseaux distinctes



- 3 types de noeuds : Intervenant, transaction, biens
- 2 ou 3 types d'arcs : I-I, I-T, I-B



CARDINALITÉS DES DONNÉES BRUTES

- 40gb de ram (liste d'adjacence)

Réseau1

Nb Sommets	17 096 773
Nb arcs	63 326 327
Nb sommets i	11 253 698
Nb sommets t	3 663 648
Nb sommets b	2 179 427
Nb arcs t-i	11 435 723
Nb arcs t-b	3 663 648
Nb arcs i-i	48 226 866

Réseau 2

Nb Sommets	9 424 925
Nb arcs	15 099 371
Nb sommets i	3 581 850
Nb sommets t	3 663 648
Nb sommets b	2 179 427
Nb arcs t-i	11 435 723
Nb arcs t-b	3 663 648
Nb arcs i-i	NA

COMPOSANTES CONNEXES PRINCIPALES

Composante Connexe1

Nb Sommets	17 007 724
Nb arcs	63 327 521
Nb sommets i	11 203 385
Nb sommets t	3 642 175
Nb sommets b	2 162 164
Nb arcs t-i	11 389 110
Nb arcs t-b	3 642 175
Nb arcs i-i	48 206 236

Composante Connexe 2

Nb Sommets	8 773 121
Nb arcs	14 498 213
Nb sommets i	3 259 578
Nb sommets t	3 486 526
Nb sommets b	2 027 017
Nb arcs t-i	11 011 687
Nb arcs t-b	3 486 526
Nb arcs i-i	NA

EXPLORATION DES RÉSEAUX

- Preuves de concept pour les pipelines suivant :
  - Mise en mémoire des listes d'adjacences de chaque réseau et de chaque composantes connexes
  - Distributions selon les types de noeuds (i-b-t) de plusieurs métriques (degrés, centralités, etc...)
  - Listes d'adjacences comme squelette pour analyses ultérieures (parallélisation avec *cugraph*)
  - Utiliser un *graph auto-encoder* pour détecter les anomalies (voir conclusion)
  - Il faut trouver des *features* maintenant!

FCC 1 (n=17 007 724)

	Qté Abs	Ratio
Noeuds I	30 992	1: 19%
Noeuds T	98	4 : 40%
Noeuds B	573	1 : 59%

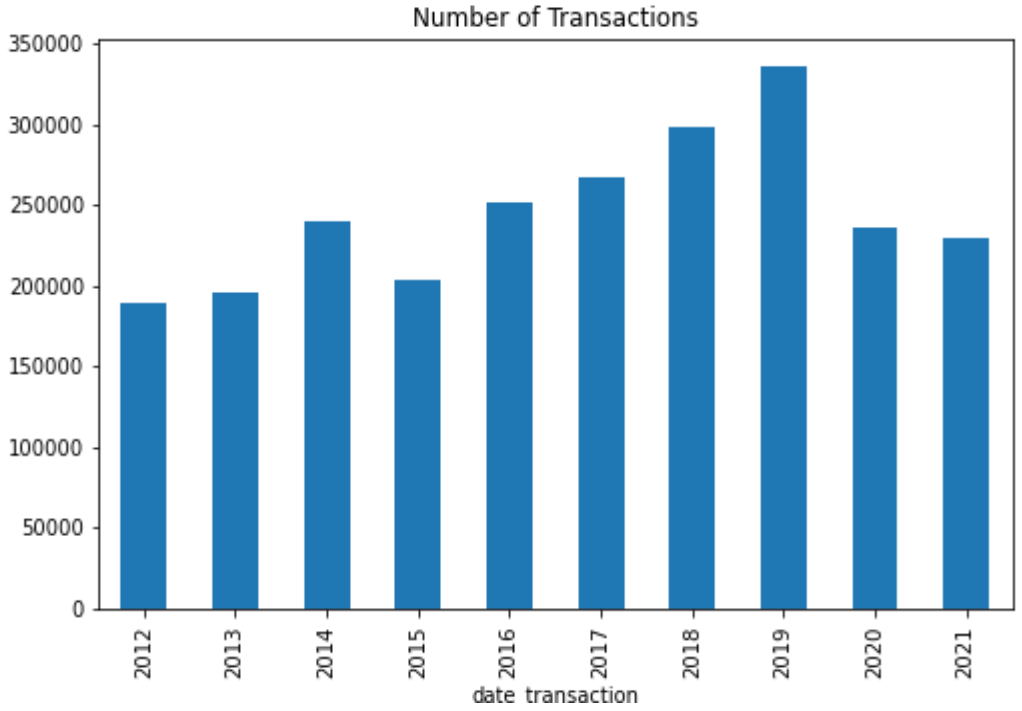
FCC 2 (n=8 773 121)

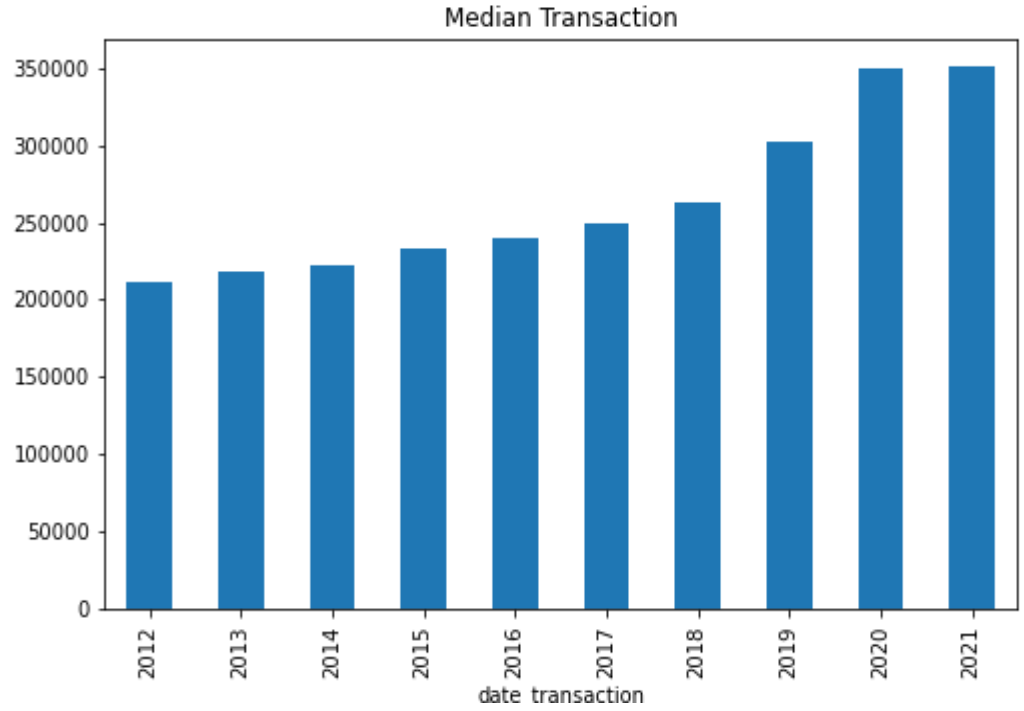
	Qté Abs	Ratio
Noeuds I	30 983	1 : 48%
Noeuds T	98	4 : 40%
Noeuds B	573	1 : 57%

APPROCHE DÉVELOPPÉE

Étape 1:	Étape 2:	Étape 3:	Étape 4:
Détection de transactions atypiques.	Création d'un graphe intervenants-intervenants de liens transactionnels.	Création d'un graphe biparti intervenants-transactions atypiques.	Détection de communautés par énumération de bicliques maximales.

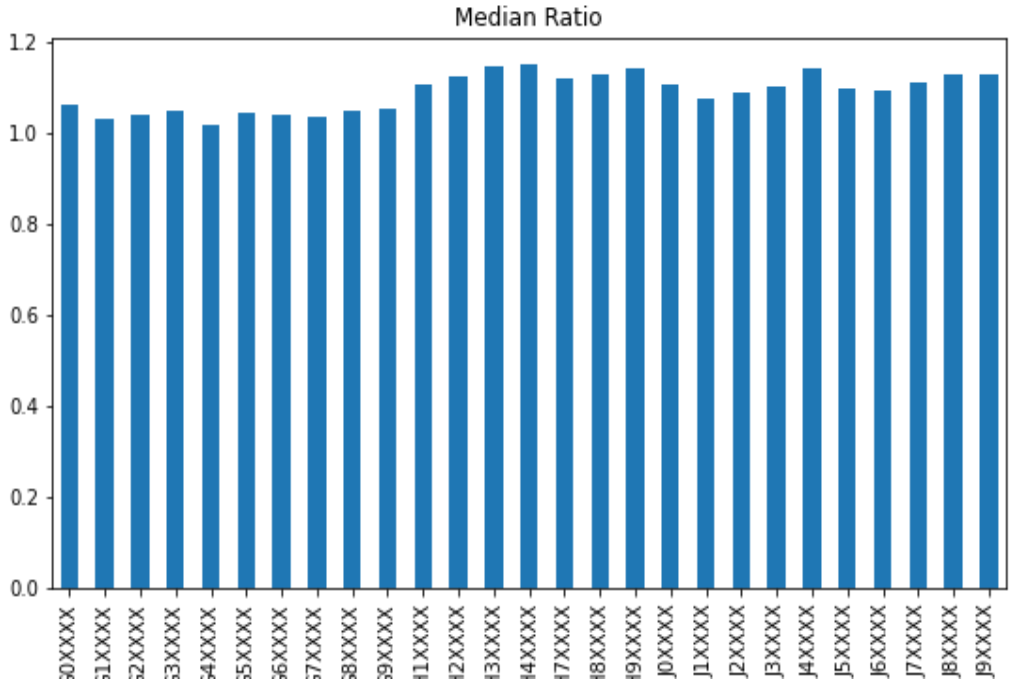
ÉTAPE 1 – DÉTECTION DE TRANSACTIONS ATYPIQUES

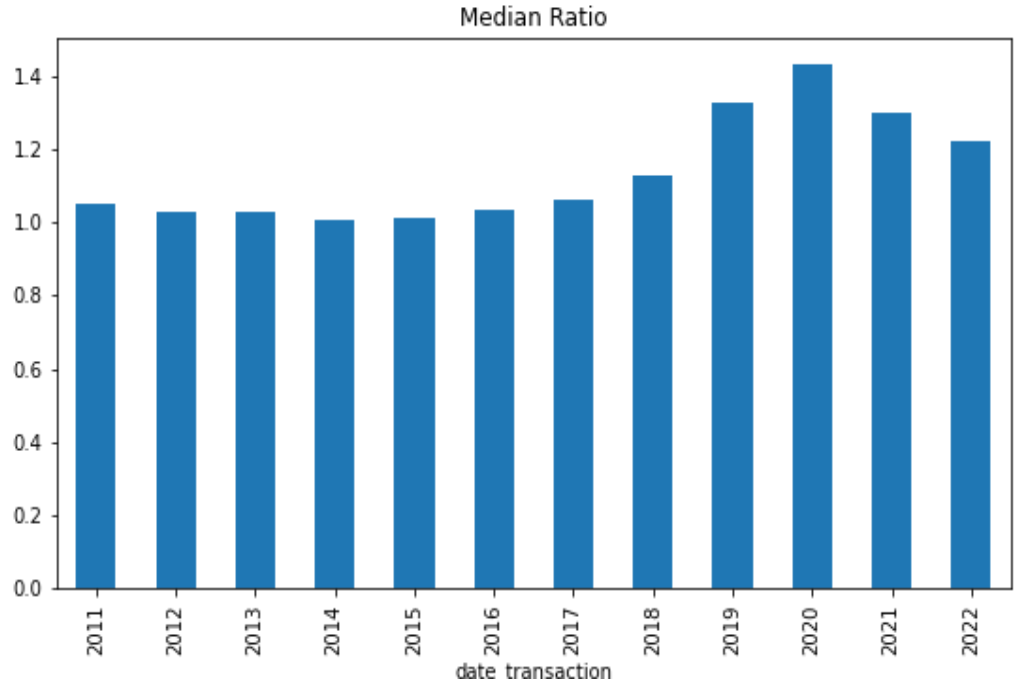




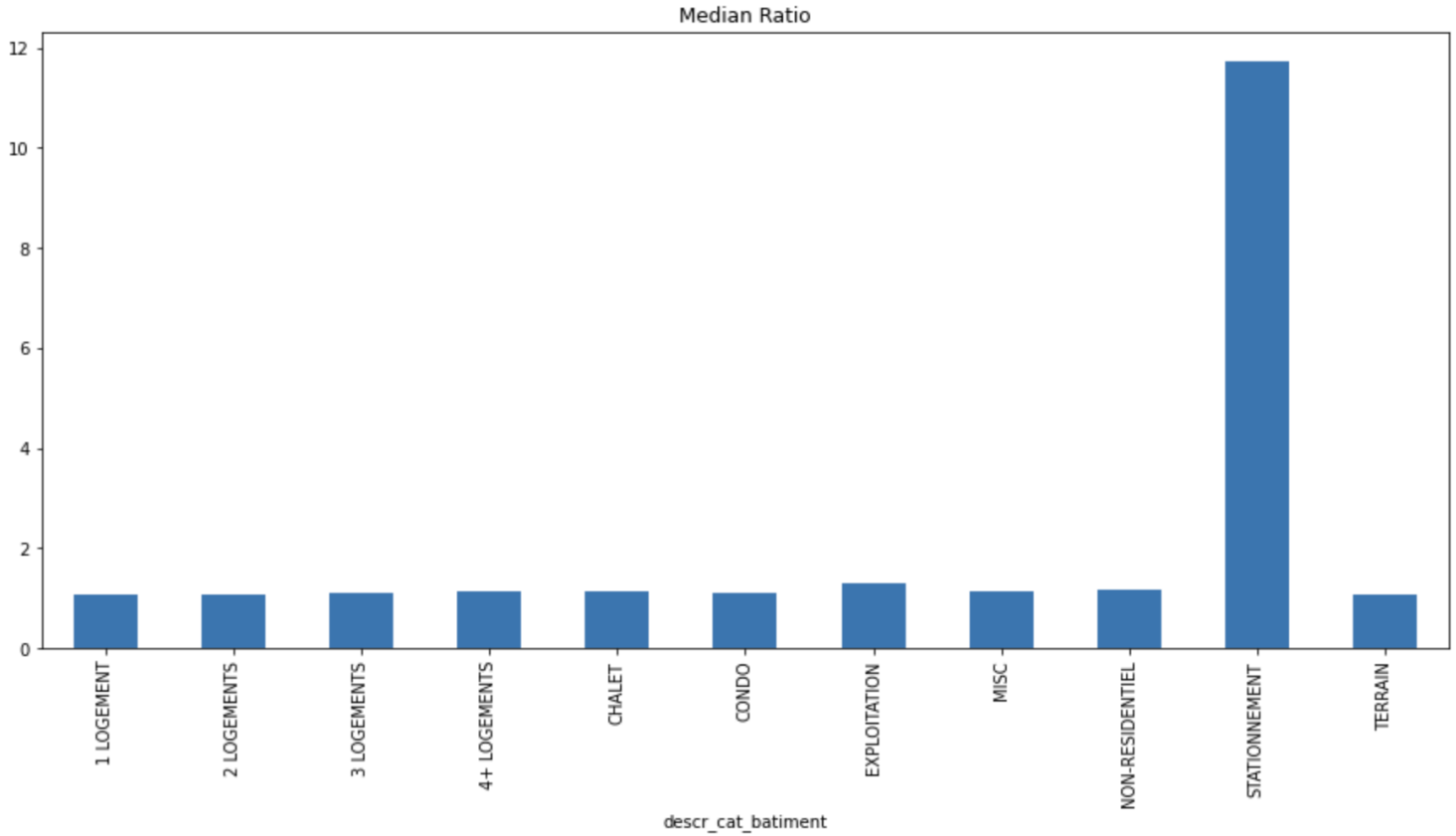
ÉTAPE 1 – DÉTECTION DE TRANSACTIONS ATYPIQUES

Ratio = montant / évaluation





ÉTAPE 1 – DÉTECTION DE TRANSACTIONS ATYPIQUES

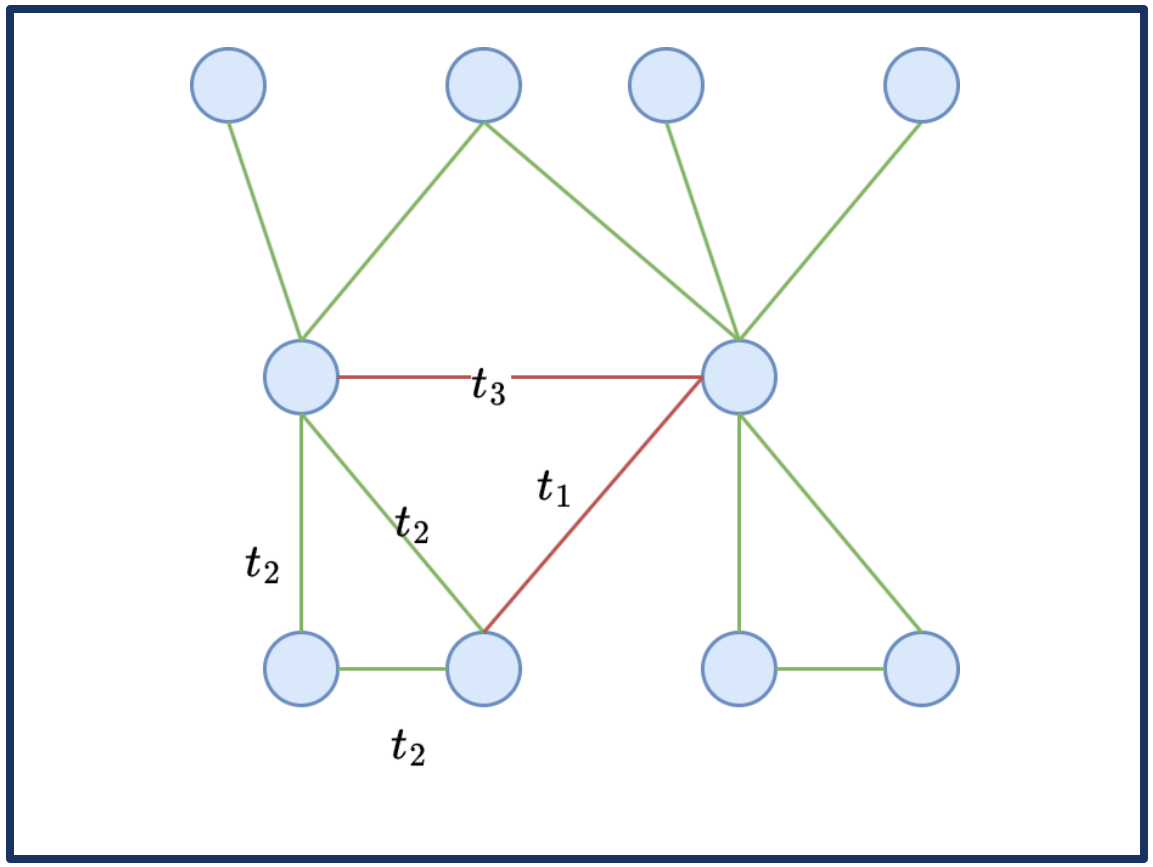


ÉTAPE 1 – DÉTECTION DE TRANSACTIONS ATYPIQUES

Méthodes à explorer:

- Isolation forest.
- K-means clustering.

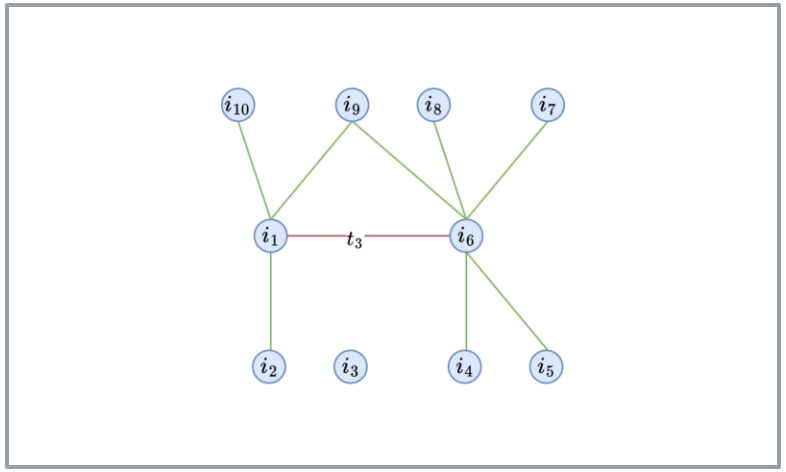
ÉTAPE 2 - CRÉATION D'UN GRAPHE INTERVENANTS-INTERVENANTS DE LIENS TRANSACTIONNELS.

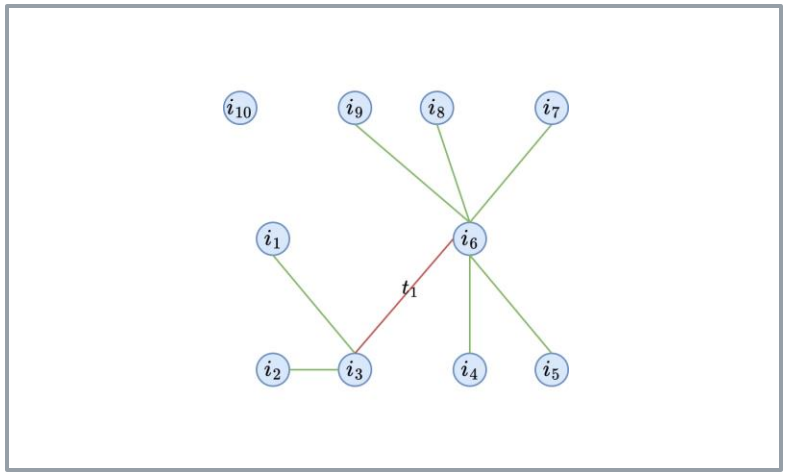


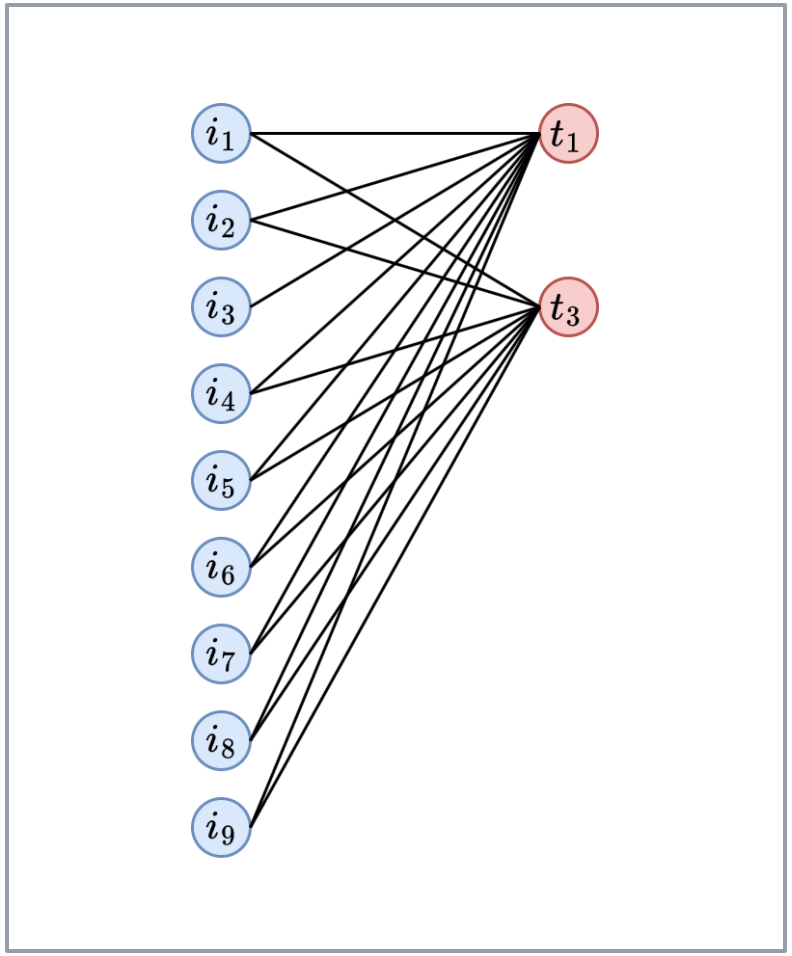
- Création d'un graphe.
- Les sommets sont les intervenants.
- Les arêtes sont les transactions.
- Les arêtes rouges sont les transactions suspicieuses fournies par l'étape 1.

ÉTAPE 3 - CRÉATION D'UN GRAPHE BIPARTI INTERVENANTS-TRANSACTIONS ATYPIQUES.

- Pour chaque transaction on regarde à **«distance» deux** les sommets.
- Problème: graphe biparti dense.

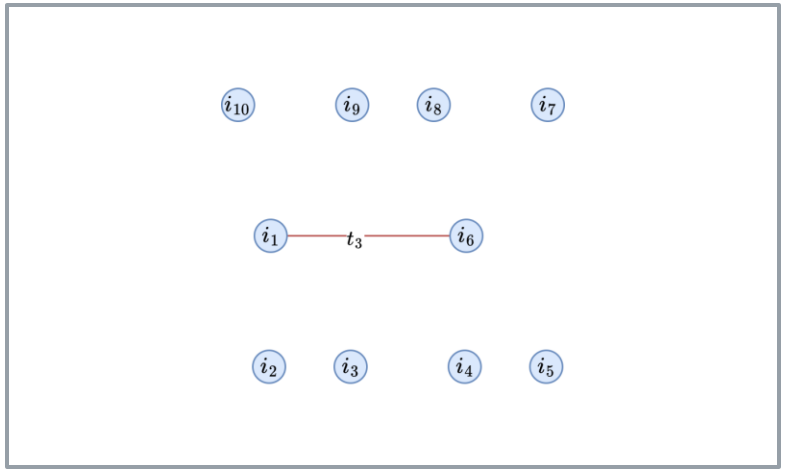


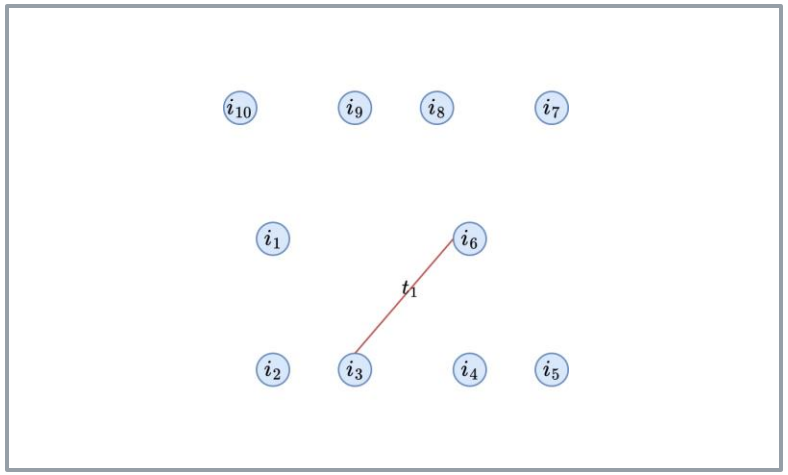


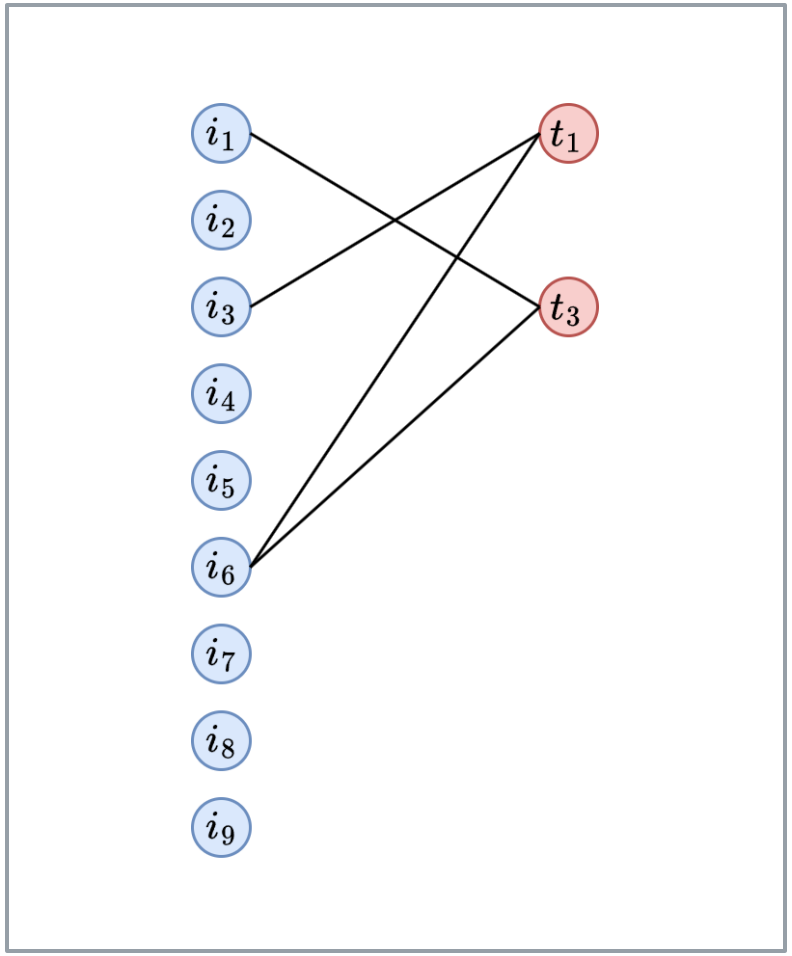


ÉTAPE 3 - CRÉATION D'UN GRAPHE BIPARTI INTERVENANTS-TRANSACTIONS ATYPIQUES.

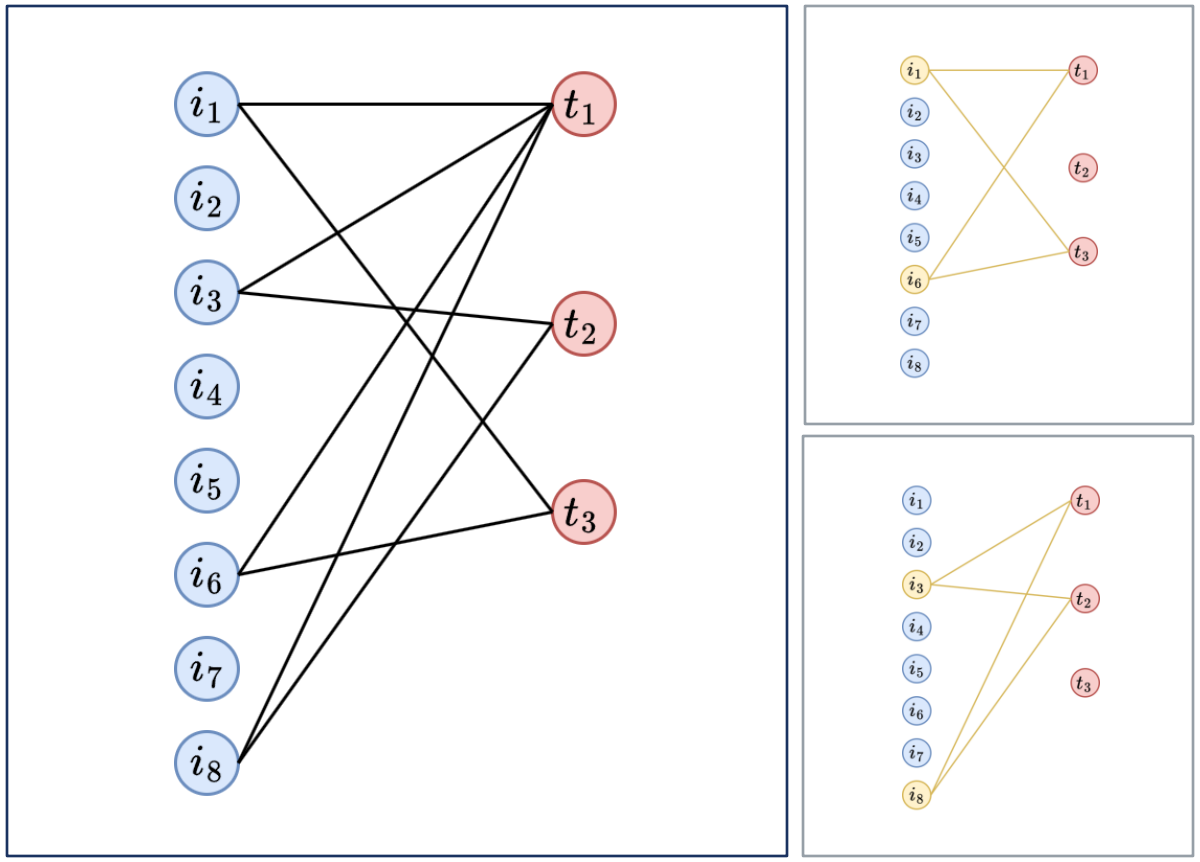
- Pour chaque transaction on regarde à **«distance» un** les sommets.
- Graphe biparti épars.
- On cherche des communautés d'intervenants qui ont des transactions suspicieuses en commun.







ÉTAPE 4 - DÉTECTION DE COMMUNAUTÉS PAR ÉNUMÉRATION DE BICLIQUES MAXIMALES.



Exemple de **bicliques maximales (en jaune)** dans un graphe biparti.

- On énumère les bicliques maximales dans le graphe biparti.
- Tiens compte de la fréquence des transactions suspicieuses en commun.

RÉSULTATS

Quantile	Nombre de communautés
0.95	11 174
0.98	4 926

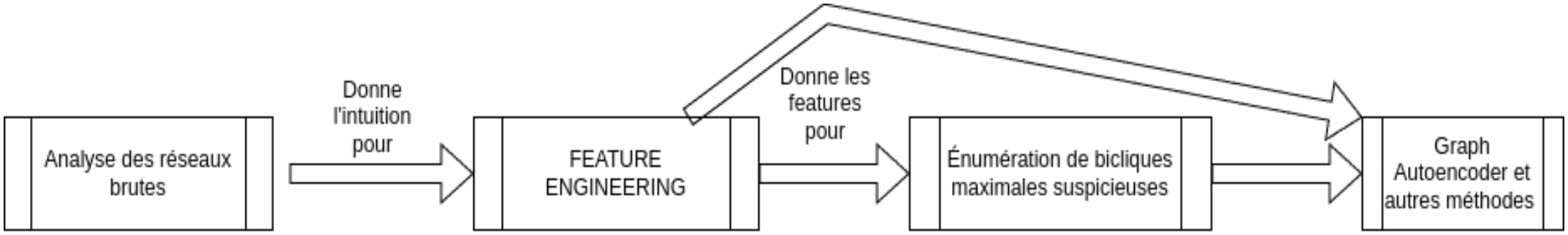
*Quantile par préfixe, par type de biens et par année.

Idée future:

- Considérer la distance 2 et ordonner les intervenants (sommets de gauche) selon une fonction $T(G, v) = n_1(G, v) + n_2(G, v)$ où n_1 et n_2 donnent respectivement le nombre de transactions suspicieuses à distance 1 et 2 de l'intervenant v.
- Agrégation des intervenants de plus grande fonction $T(G, v)$ selon leur localité transactionnelles. Les intervenants suspicieux proche dans le graphe des transactions sont regroupés pour former des communautés.

CONCLUSION

- Complémentarité dans les approches développées :
    - L'étude initiale de la structure a débuté les discussions
    - Les discussions ont alimenté le *feature engineering*
    - Le *feature engineering* a alimenté la détection des bicliques
    - Le tout sera pertinent aux avenues futures, dont le *graph autoencoder (GAE)*



CONCLUSION – AVENUES À EXPLORER

- Entraîner un GAE de façon non-supervisée sur un réseau.
- Les différences représentent les anomalies!
- Plus les données sont riches, meilleure est la performance.