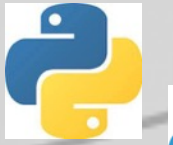
REVENU QUÉBEC

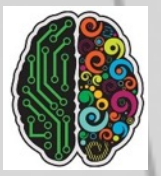
Énumération de bicliques

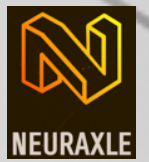
ARPI

AOÛT 2022

Revenu Québec et l'IA







Contexte du projet

Basés sur les articles *Gotcha!** et *Guilt-by-Constellation***

- Hypothèses
  - Une entreprise partageant plusieurs ressources avec une entreprise ou plusieurs ayant fait une faillite frauduleuse est plus à risque d'en faire une elle-même
  - Extraire des attributs d'un graphe pondéré à travers le partage de ressources permet d'enrichir un algorithme d'apprentissage supervisé

- Objectif
  - Améliorer l'évaluation du risque d'une entreprise de faire une faillite frauduleuse

*Van Vlasselaer, V., Eliassi-Rad, T., Akoglu, L., Snoeck, M., Baesens, B., "GOTCHA! Network-Based Fraud Detection for Social Security Fraud", Management Science 63(9):3090-3110, July 2016.

**Van Vlasselaer, V., Akoglu, L., Eliassi-Rad, T., Snoeck, M., Baesens, B., "Guilt-by-constellation: fraud detection by suspicious clique memberships". In: 2015 48th Hawaii International Conference on System Sciences (HICSS), pp. 918–927. IEEE, January 2015.

Contexte du projet

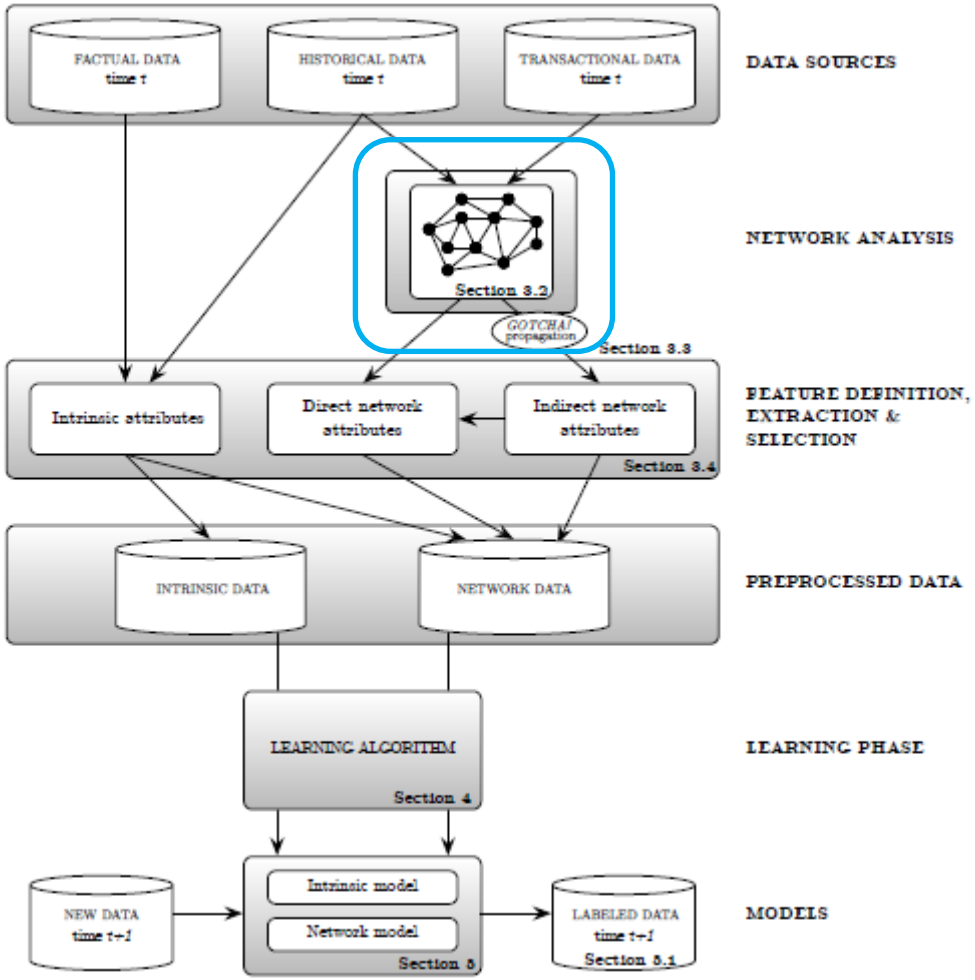
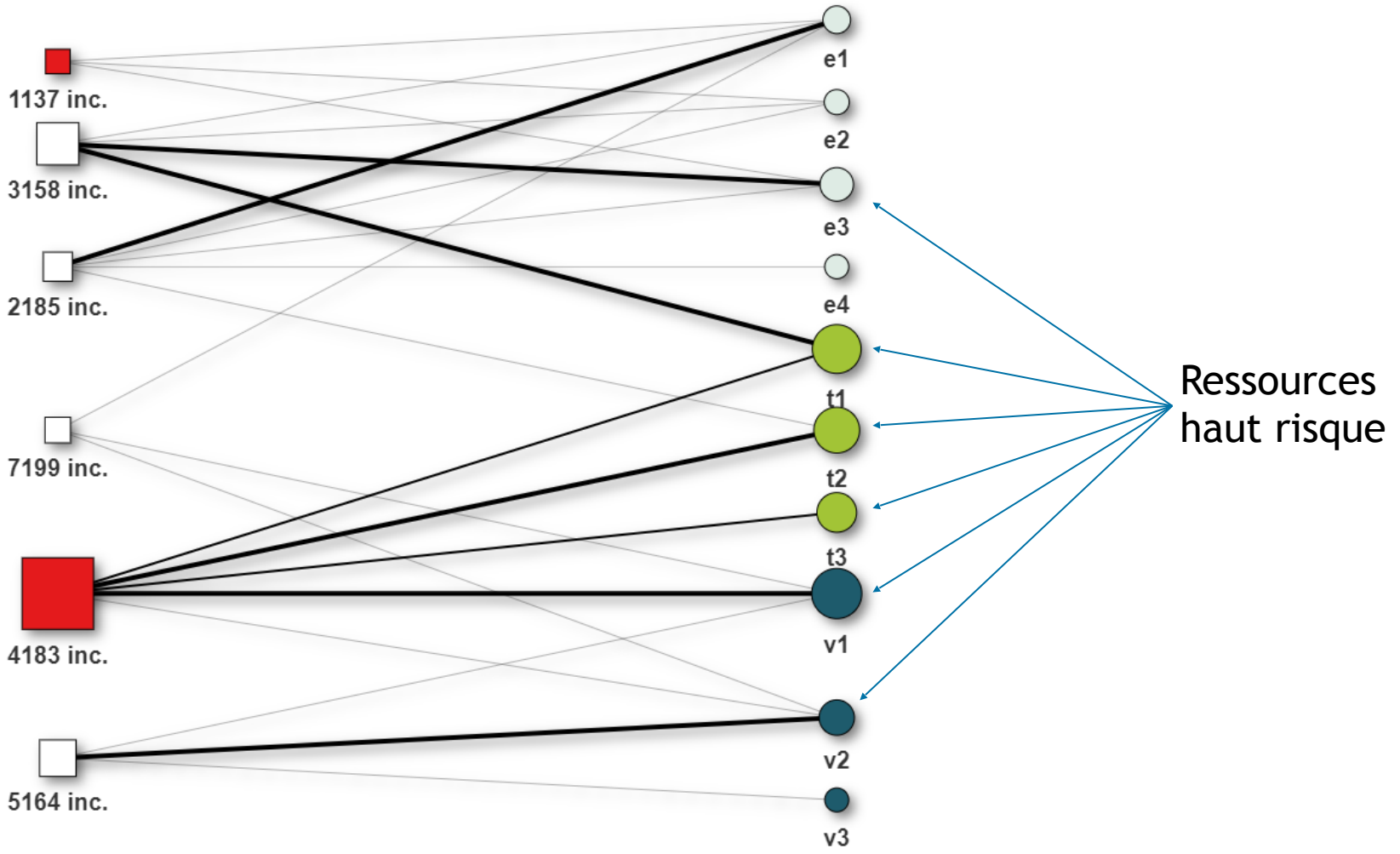


Figure 5 Proposed GOTCHA! framework for social security fraud detection.

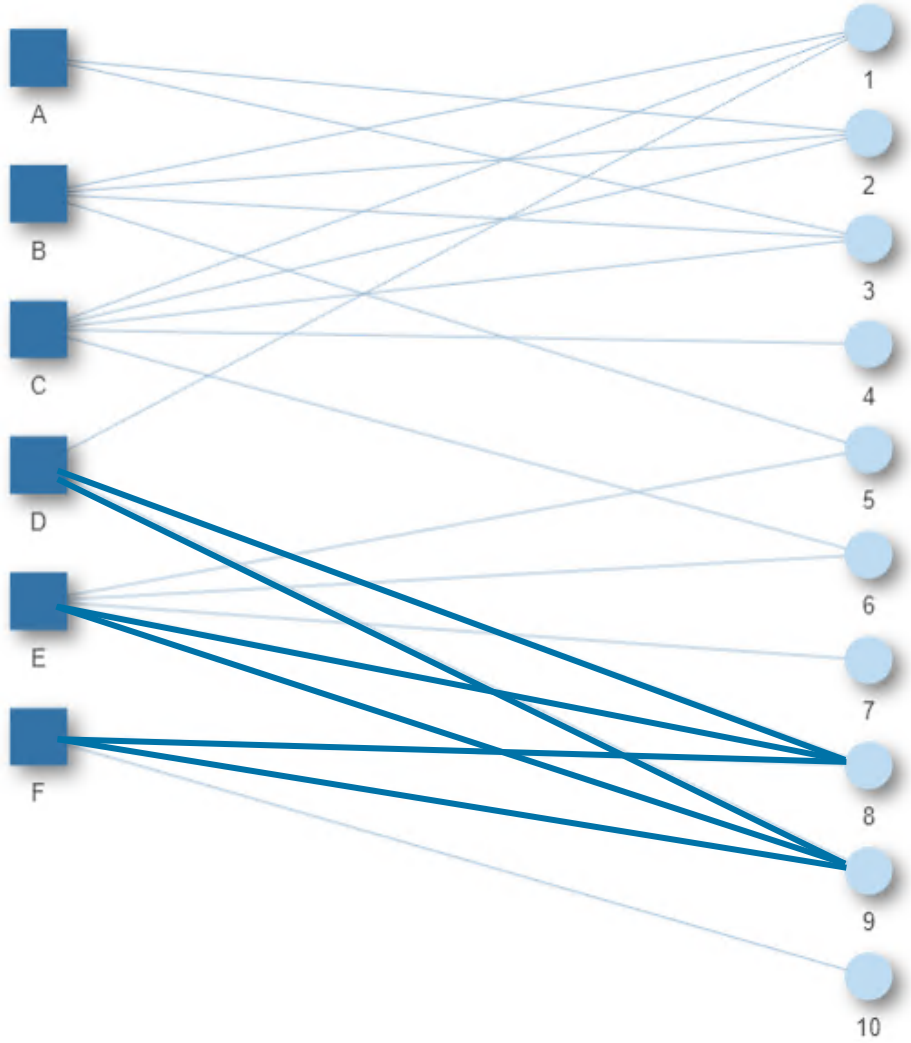
Énoncé du problème

- Entreprises
  - Indicateur non-conformité fiscale
  - Score d'exposition
- Ressources
  - Score d'exposition
  - Identification haut risque
- Arêtes
  - Pondération



Énoncé du problème

- Étant donné un graphe biparti G où des sommets de type 1 sont reliés à des sommets de type 2
- **On cherche l'ensemble des bicliques maximales**
- Biclique : sous-graphe de G dans lequel chaque sommet de type 1 est lié à chaque sommet de type 2
- Biclique maximale : biclique à laquelle aucun sommet de G ne peut être ajouté pour former une biclique dans G



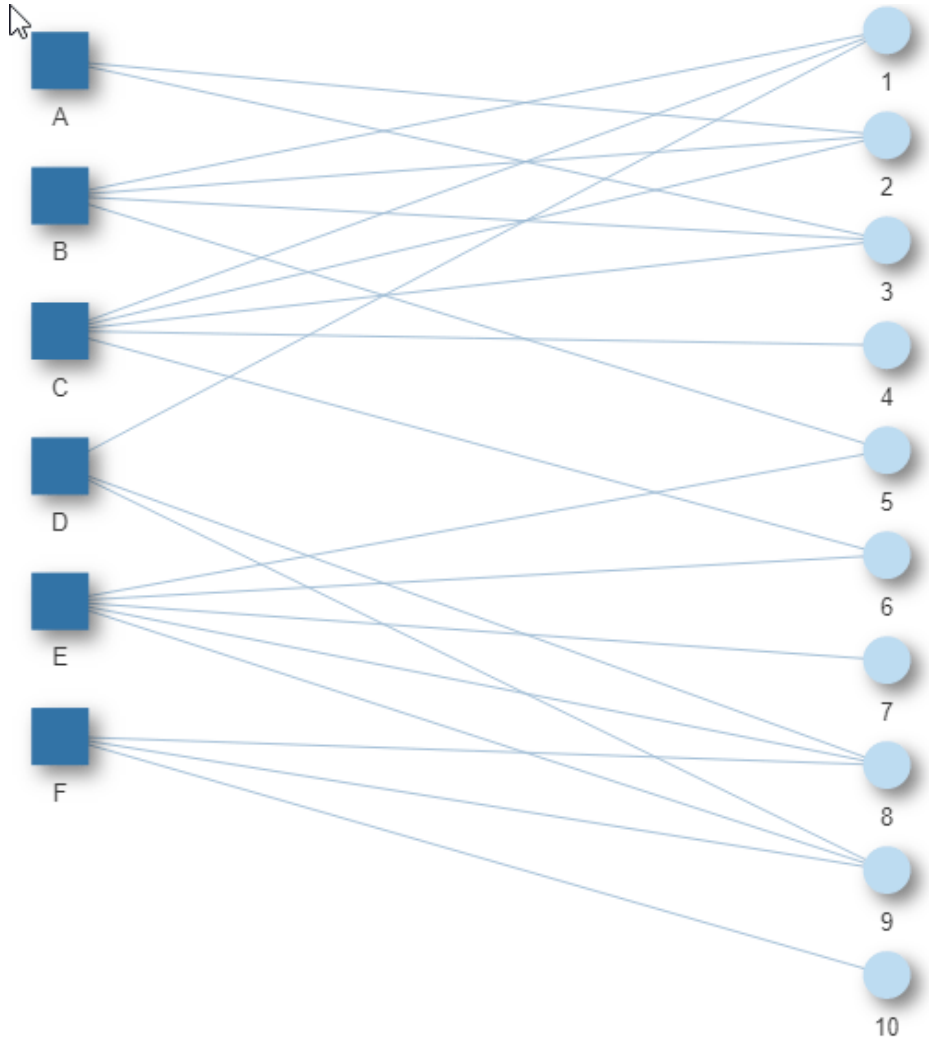
Énoncé du problème

Particularité des bicliques à énumérer :

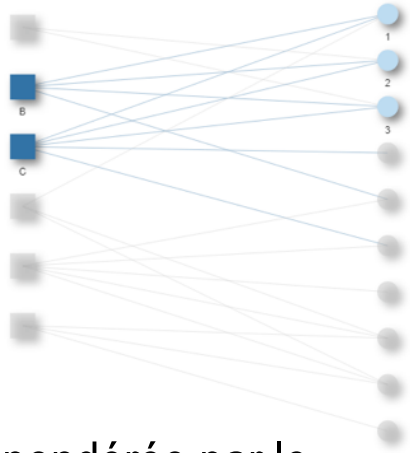
- Au moins deux sommets de type 1 et deux sommets de type 2

Liste des bicliques finales :

Sommets type 1	Sommets type 2
B, C	1, 2, 3
A, B, C	2, 3
D, E, F	8, 9



Énoncé du problème



Calcul de caractéristiques associées à chaque biclique

- Dénombrement : nombre absolu d'entreprises non-conformes dans la biclique
- Proportion : nombre relatif d'entreprises non-conformes dans la biclique
- Somme entreprises : somme des scores d'exposition des entreprises de la biclique
- Somme entreprises pondérée : somme des scores d'exposition des entreprises de la biclique pondérée par le nombre d'entreprises dans la biclique
- Somme ressources : somme des scores d'exposition des ressources de la biclique
- Somme ressources pondérée : somme des scores d'exposition des ressources de la biclique pondérée par le nombre de ressources dans la biclique
- Magnitude : nombre total de sommets de la biclique, le nombre de sommets entreprise de la biclique, le nombre de sommets ressources de la biclique

Énoncé du problème

Calcul d'attributs pour chaque entreprise

- Dénombrement : nombre de bicliques auxquelles appartient l'entreprise
- Moyennes : moyenne de chaque caractéristique de biclique parmi les bicliques auxquelles appartient l'entreprise
- Maximums : valeur maximum de chaque caractéristique de biclique parmi les bicliques auxquelles appartient l'entreprise

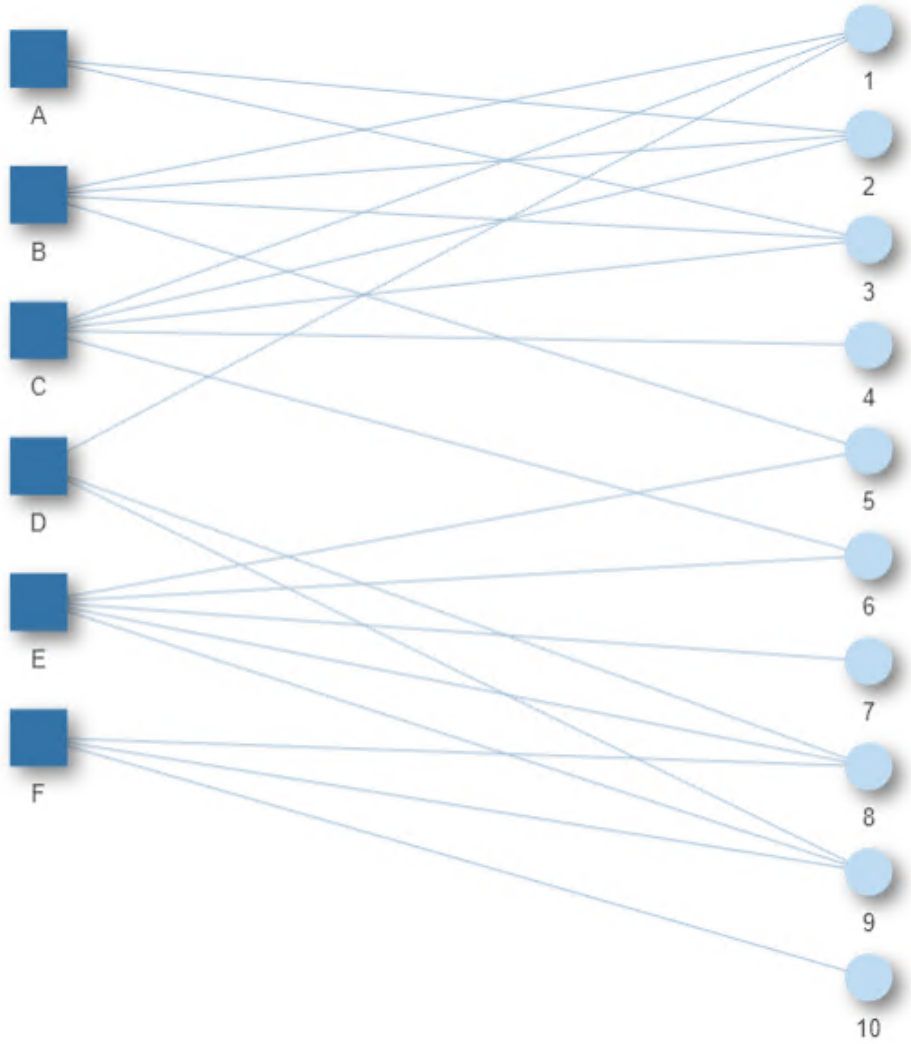
- Moyenne du nombre d'entreprises non-conformes par biclique
- Moyenne de la proportion d'entreprises non-conformes par biclique
- Moyenne de la somme du score d'exposition des entreprises par biclique
- Moyenne de la somme pondérée du score d'exposition des entreprises par biclique
- Moyenne de la somme du score d'exposition des ressources par biclique
- Moyenne de la somme pondérée du score d'exposition des ressources par biclique
- Moyenne de chacune des trois mesures de magnitude des bicliques

- Maximum du nombre d'entreprises non-conformes parmi les bicliques
- Maximum de la proportion d'entreprises non-conformes parmi les bicliques
- Maximum de la somme du score d'exposition des entreprises parmi les bicliques
- Maximum de la somme pondérée du score d'exposition des entreprises parmi les bicliques
- Maximum de la somme du score d'exposition des ressources parmi les bicliques
- Maximum de la somme pondérée du score d'exposition des ressources parmi les bicliques
- Maximum de chacune des trois mesures de magnitude parmi les bicliques

Description de l'algorithme

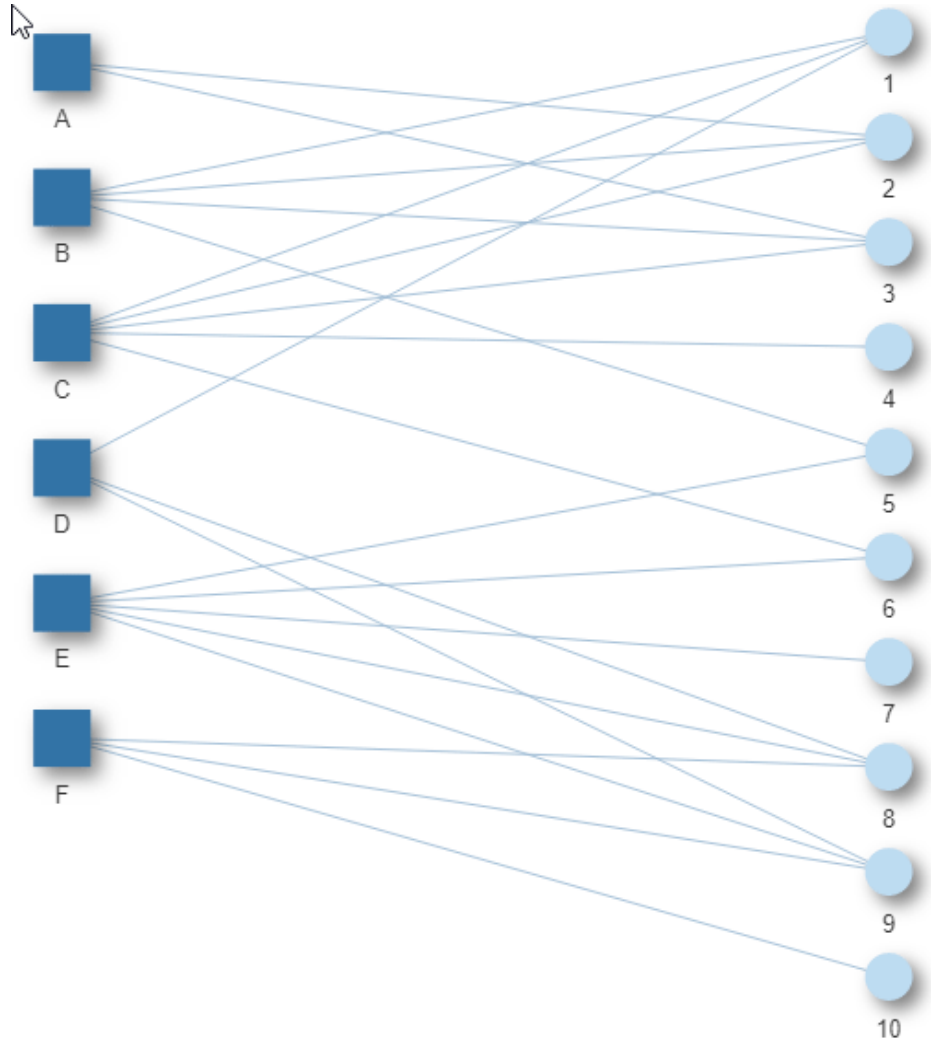
1- Ensemble de bicliques initiales

- Énumérer toutes les paires d'entreprises qui partagent au moins deux ressources



Description de l'algorithme

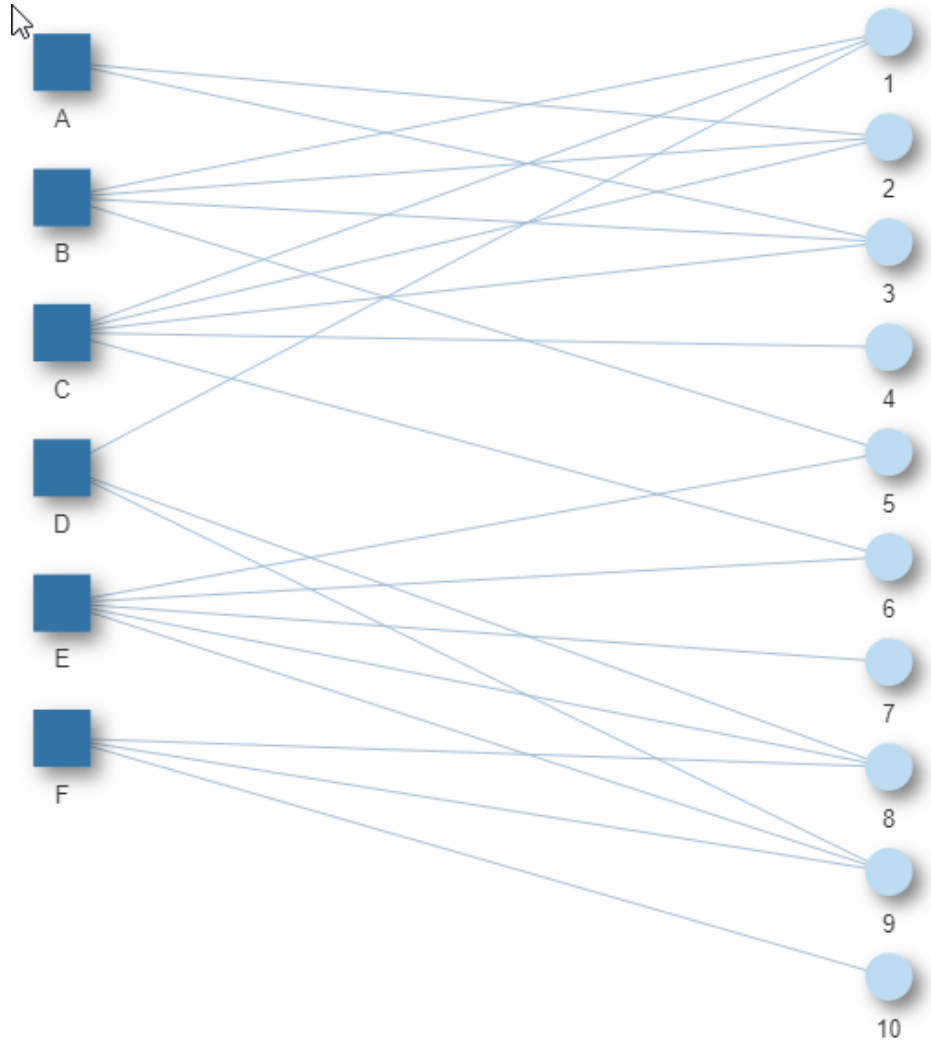
```
enumerer_cliques_init <- function(g) {
  m_inc <- igraph::as_incidence_matrix(g)

  cliques <- list()
  cliques_voisins_communs <- list()

  for (i in seq_len(nrow(m_inc) - 1)) {
    for (j in (i + 1):(nrow(m_inc))) {
      voisins_communs <- m_inc[i, ] + m_inc[j, ]
      voisins_communs[voisins_communs <= 1] <- 0
      voisins_communs[voisins_communs >= 2] <- 1
      if (sum(voisins_communs) >= 2) {
        cliques <-
          append(cliques, list(sort(c(rownames(m_inc)[i], rownames(m_inc)[j]))))
        cliques_voisins_communs <-
          append(cliques_voisins_communs, list(voisins_communs))
      }
    }
  }

  return(list(cliques, cliques_voisins_communs))
}
```



Description de l'algorithme

Liste des bicliques initiales :

Sommets entreprise	Ressources communes
A, B	2, 3
A, C	2, 3
B, C	1, 2, 3
D, E	8, 9
D, F	8, 9
E, F	8, 9



Description de l'algorithme

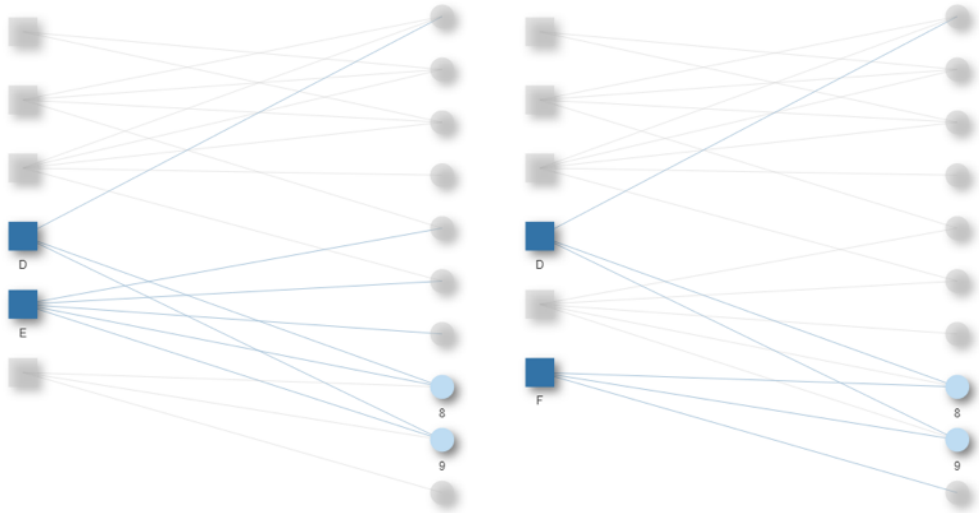
1- Ensemble de bicliques initiales

- Énumérer toutes les paires d'entreprises qui partagent au moins deux ressources

2- Fusion des bicliques

Complètes

```
POUR bc1 dans bicliques FAIRE
  POUR bc2 dans bicliques\bc1 FAIRE
    SI voisins(bc1) == voisins(bc2) ALORS
      bicliques += fusion(bc1, bc2)
      bicliques -= bc1
      bicliques -= bc2
    SINON SI |intersection(voisins(bc1), voisins(bc2))| >= 2 ALORS
      bicliques += fusion(bc1, bc2)
    FIN SI
  FIN POUR
FIN POUR
```



Description de l'algorithme

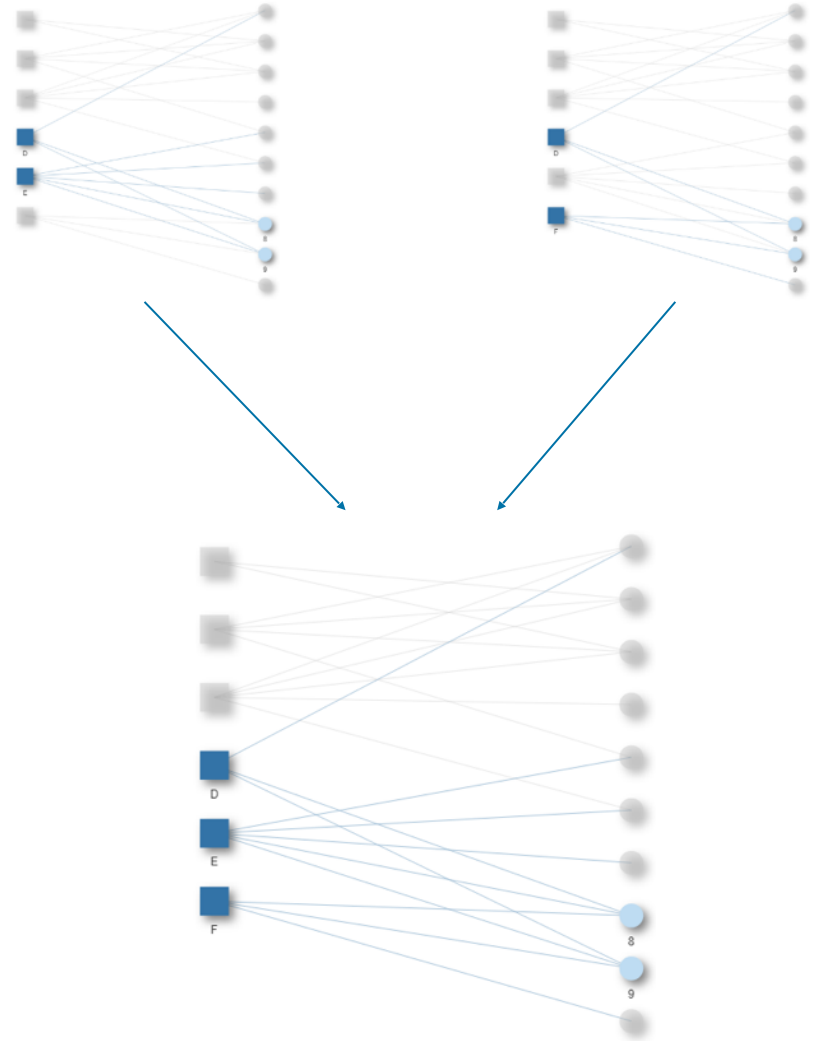
1- Ensemble de bicliques initiales

- Énumérer toutes les paires d'entreprises qui partagent au moins deux ressources

2- Fusion des bicliques

Complètes

```
POUR bc1 dans bicliques FAIRE
  POUR bc2 dans bicliques\bc1 FAIRE
    SI voisins(bc1) == voisins(bc2) ALORS
      bicliques += fusion(bc1, bc2)
      bicliques -= bc1
      bicliques -= bc2
    SINON SI |intersection(voisins(bc1), voisins(bc2))| >= 2 ALORS
      bicliques += fusion(bc1, bc2)
    FIN SI
  FIN POUR
FIN POUR
```



Description de l'algorithme

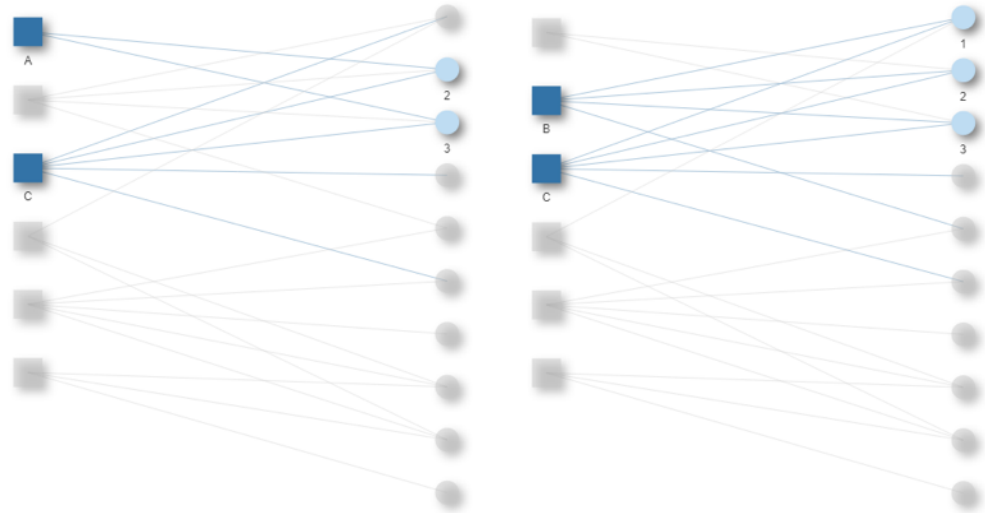
1- Ensemble de bicliques initiales

- Énumérer toutes les paires d'entreprises qui partagent au moins deux ressources

2- Fusion des bicliques

Partielles

```
POUR bc1 dans bicliques FAIRE
  POUR bc2 dans bicliques\bc1 FAIRE
    SI voisins(bc1) == voisins(bc2) ALORS
      bicliques += fusion(bc1, bc2)
      bicliques -= bc1
      bicliques -= bc2
    SINON SI |intersection(voisins(bc1), voisins(bc2))| >= 2 ALORS
      bicliques += fusion(bc1, bc2)
    FIN SI
  FIN POUR
FIN POUR
```



Description de l'algorithme

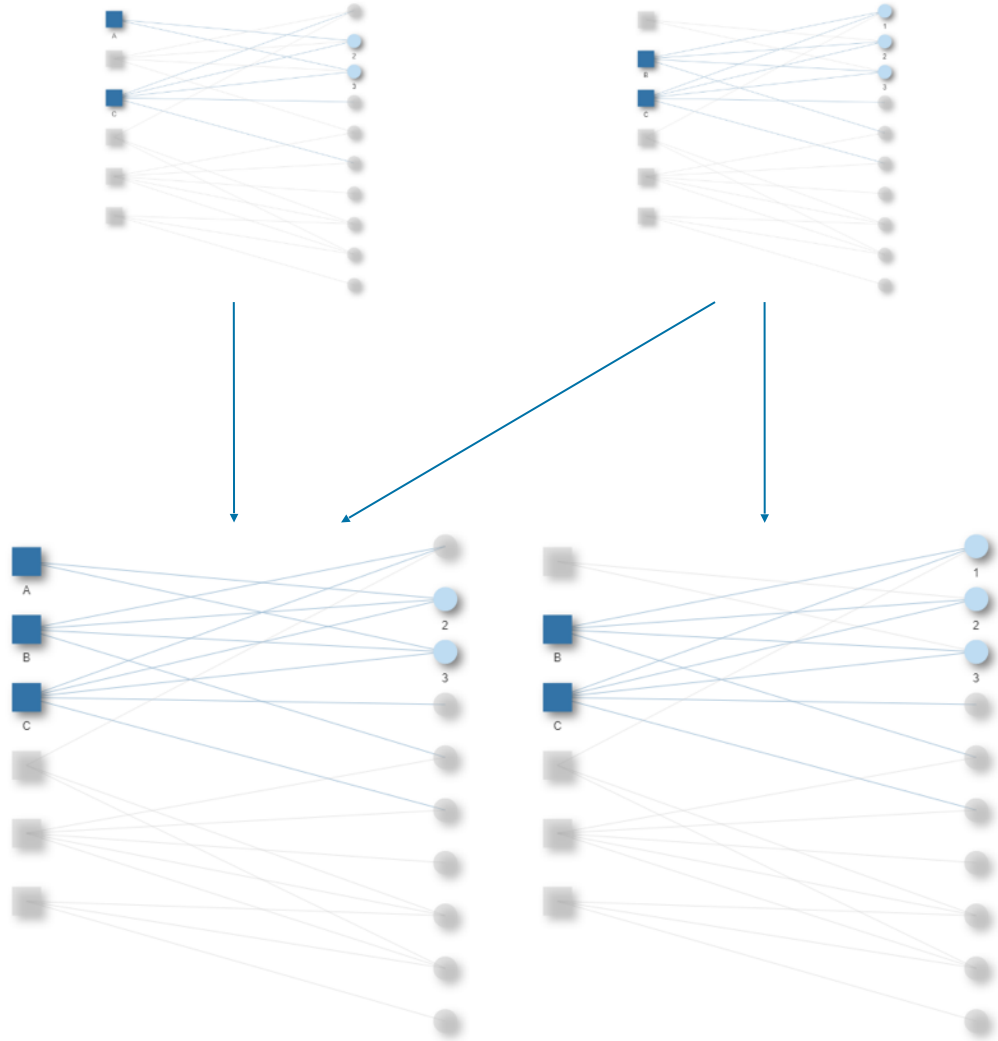
1- Ensemble de bicliques initiales

- Énumérer toutes les paires d'entreprises qui partagent au moins deux ressources

2- Fusion des bicliques

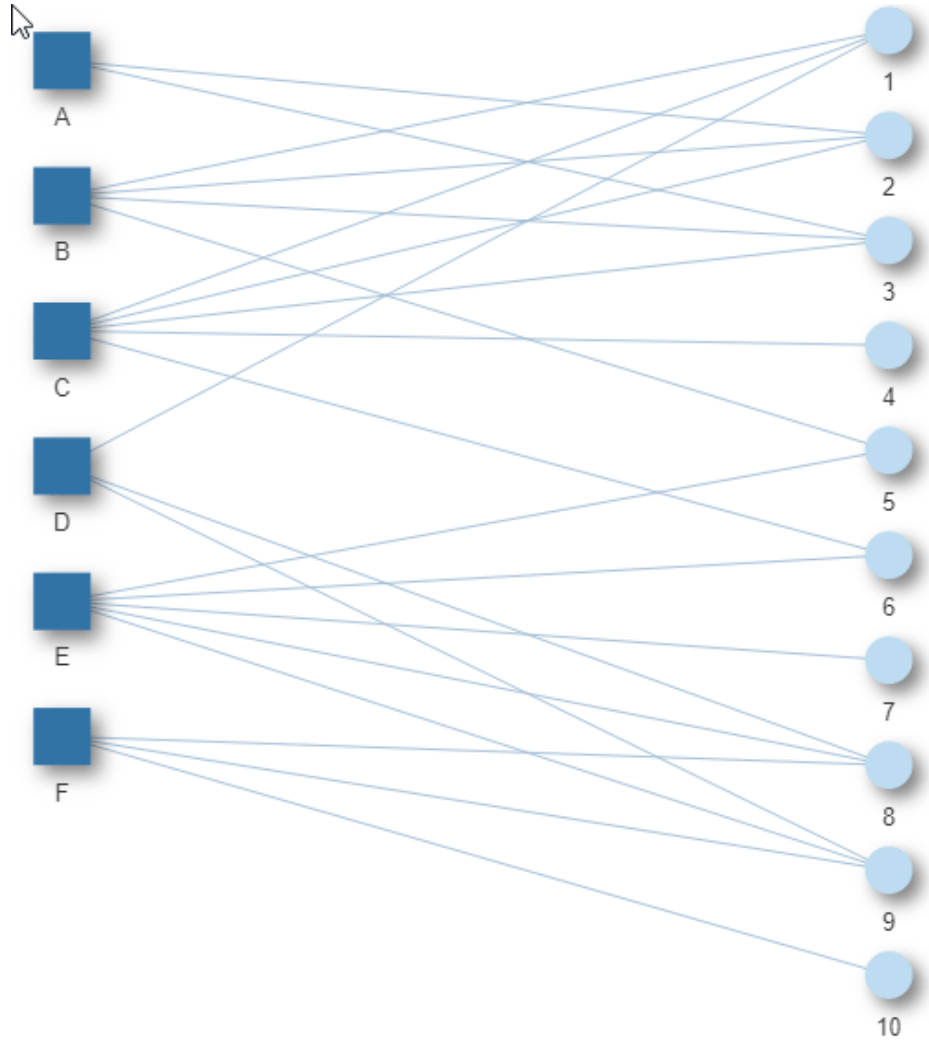
Partielles

```
POUR bc1 dans bicliques FAIRE
  POUR bc2 dans bicliques\bc1 FAIRE
    SI voisins(bc1) == voisins(bc2) ALORS
      bicliques += fusion(bc1, bc2)
      bicliques -= bc1
      bicliques -= bc2
    SINON SI |intersection(voisins(bc1), voisins(bc2))| >= 2 ALORS
      bicliques += fusion(bc1, bc2)
    FIN SI
  FIN POUR
FIN POUR
```



Description de l'algorithme

```
POUR bc1 dans bicliques FAIRE
  POUR bc2 dans bicliques\bc1 FAIRE
    SI voisins(bc1) == voisins(bc2) ALORS
      bicliques += fusion(bc1, bc2)
      bicliques -= bc1
      bicliques -= bc2
    SINON SI |intersection(voisins(bc1), voisins(bc2))| >= 2 ALORS
      bicliques += fusion(bc1, bc2)
    FIN SI
  FIN POUR
FIN POUR
```



Liste des bicliques finales :

Sommets entreprise	Ressources communes
B, C	1, 2, 3
A, B, C	2, 3
D, E, F	8, 9

Attentes

- L'algorithme d'énumération de bicliques décrit dans l'article
  - N'est pas performant
  - Résout un problème d'énumération de bicliques maximales dans un graphe biparti
- Le problème d'énumération de bicliques maximales ne se résout pas en temps polynomial
- Jeux de données de grand volume :
  - Centaines de milliers d'entreprises
  - Millions de ressources

Ce qui est recherché de la participation à l'ARPI

- Algorithme efficace d'énumération de bicliques
- Calcul performant des caractéristiques des bicliques
- Calcul performant des attributs d'entreprise basés sur les bicliques

Astuces algorithmiques
- **Heuristiques**
- **Approche diviser pour régner**
- **Calcul restreint à un sous-ensemble de bicliques**

Stratégies d'augmentation de la capacité de calcul
- **HPC**
- **Parallélisation**

REVENU QUÉBEC

JUSTE.
POUR TOUS.