Énumération de bicliques maximales pour l'aide à la détection d'entreprises RNCF

Karine Dufresne (Revenu Québec)
Hugues-Étienne Moisan-Plante (Revenu Québec)
Mathieu Gervais-Dubé (Polytechnique Montréal)
Nicolas Goulet (UQAM)
Alain Hertz (Polytechnique Montréal)
Odile Marcotte (CRM et UQAM)

26 août 2022

Contexte

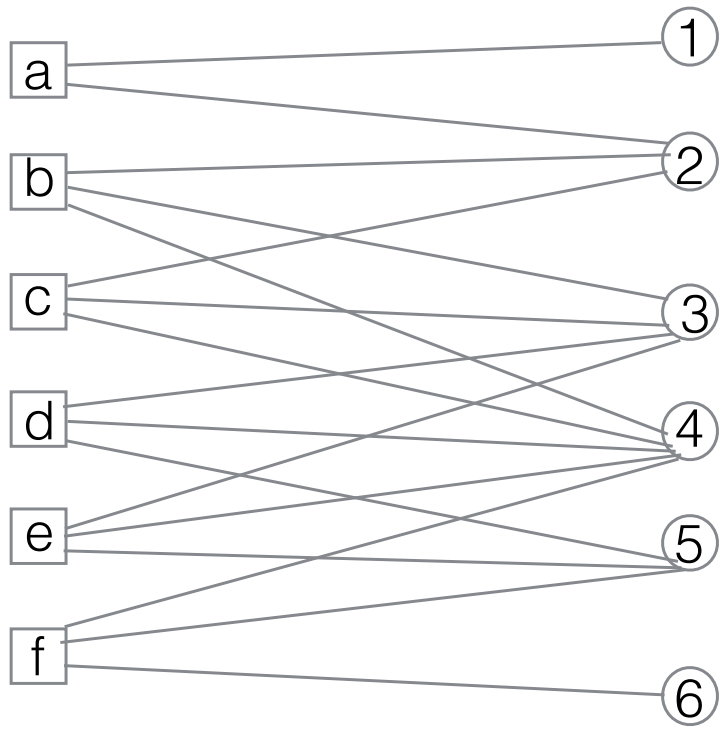
- Revenu Québec s'intéresse aux entreprises présentant un Risque de Non Conformité Fiscale (RNCF).
- L'agence a entrepris une analyse s'inspirant d'une méthode appliquée en Belgique et décrite dans les articles *GOTCHA!* et *Guilt-by-Constellation*.
- La méthode calcule des attributs associés aux entreprises qui peuvent être utilisés dans le cadre d'un algorithme d'apprentissage automatique.

Relations entre les entreprises

- Les attributs d'une entreprise peuvent dépendre des attributs d'autres entreprises, reliées à la première par le biais de leurs ressources communes.
- La méthode doit tenir compte d'un graphe biparti dans lequel les entreprises sont reliées aux ressources qu'elles utilisent.
- Dans ce graphe biparti on peut trouver des structures permettant de calculer des attributs d'entreprises.

Bicliques maximales

Une *biclique* consiste d'un ensemble d'entreprises (A) et d'un ensemble de ressources (B) tels que n'importe quelle entreprise u dans A utilise chaque ressource v dans B ; A doit contenir au moins deux sommets et B aussi. Une biclique est *maximale* si elle n'est pas incluse dans une autre biclique.



Énumération de toutes les bicliques maximales afin de calculer des attributs

- Le graphe biparti fourni par Revenu Québec contient environ 500000 entreprises et 3,3 millions de ressources.
- L'équipe de Revenu Québec a tenté d'utiliser l'algorithme décrit dans l'article *Biclique: an R package for maximal biclique enumeration in bipartite graphs*, de Lu, Phillips et Langston.
- Elle a aussi tenté de mettre en oeuvre l'algorithme de *Guilt-by-Constellation*, qui est un algorithme ascendant.

Réduction du graphe

- L'intuition sous-tendant l'amélioration proposée est que si u est un sommet de degré peu élevé, l'énumération des bicliques maximales contenant u est relativement facile.
- Le graphe fourni par Revenu Québec contient environ 2 millions de sommets de degré 1, qui peuvent être enlevés du graphe ; le processus peut être itéré.
- Si un sommet u est relié à deux sommets seulement (disons v et w), la biclique consistant de u, de v, de w et de tous les sommets adjacents à v ET w sera maximale ; u peut alors être retiré du graphe.

Illustration

Dans notre exemple les sommets a, 1 et 6 peuvent être retirés du graphe.

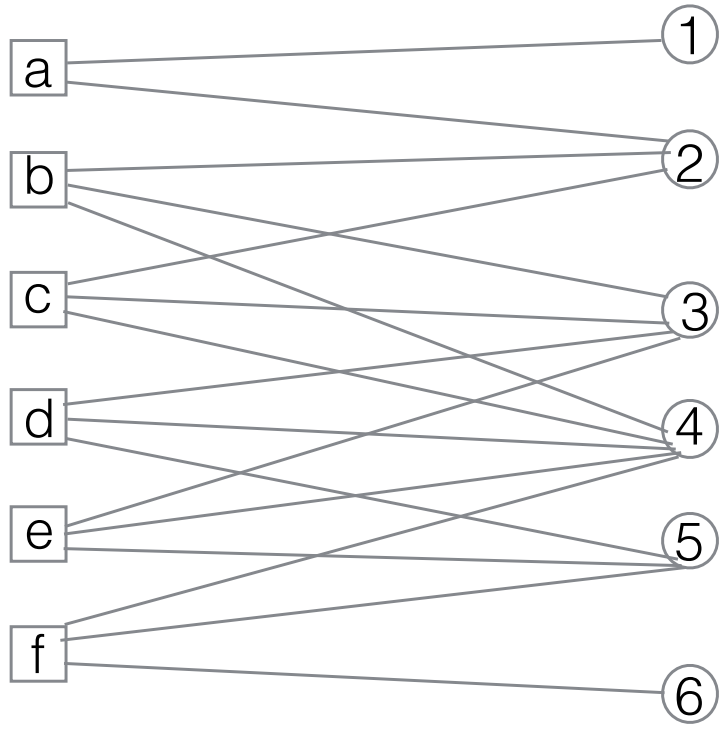
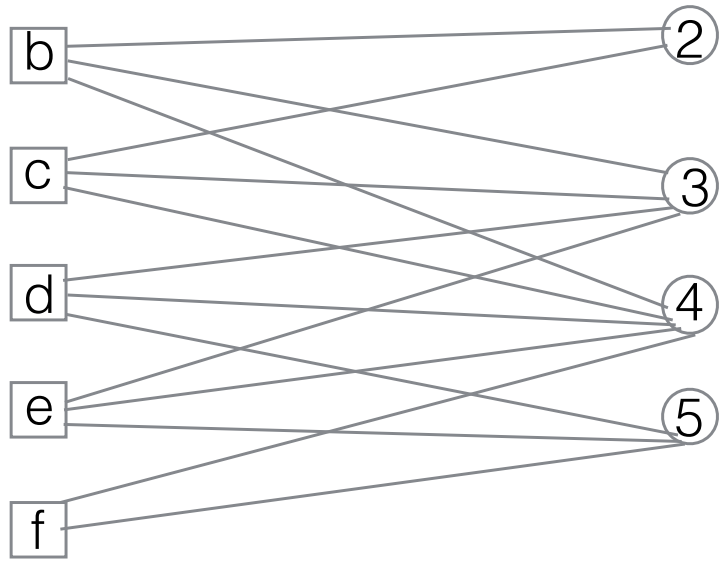


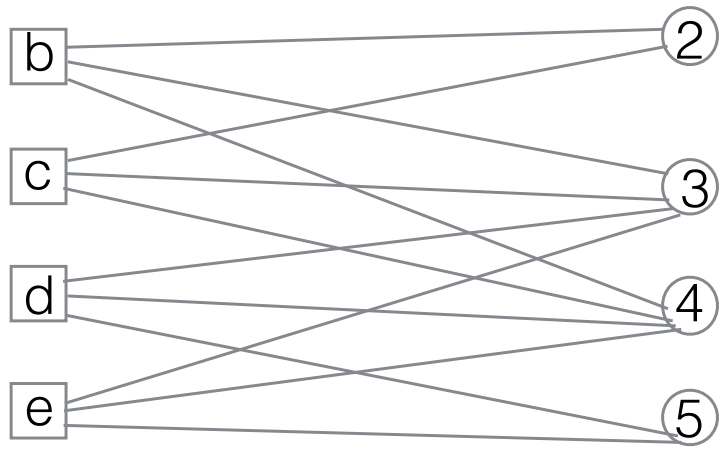
Illustration (suite)

Le sommet f nous fournit alors la biclique maximale consistant des sommets d, e, f, 4 et 5.



Cas des sommets de degré 3

Soit b un sommet de degré 3. Il faut considérer non seulement l'ensemble de tous ses voisins (ce qui donne la biclique $\{$b, c, 2, 3, 4$\}$), mais chacun des sous-ensembles de l'ensemble des voisins contenant au moins deux sommets.



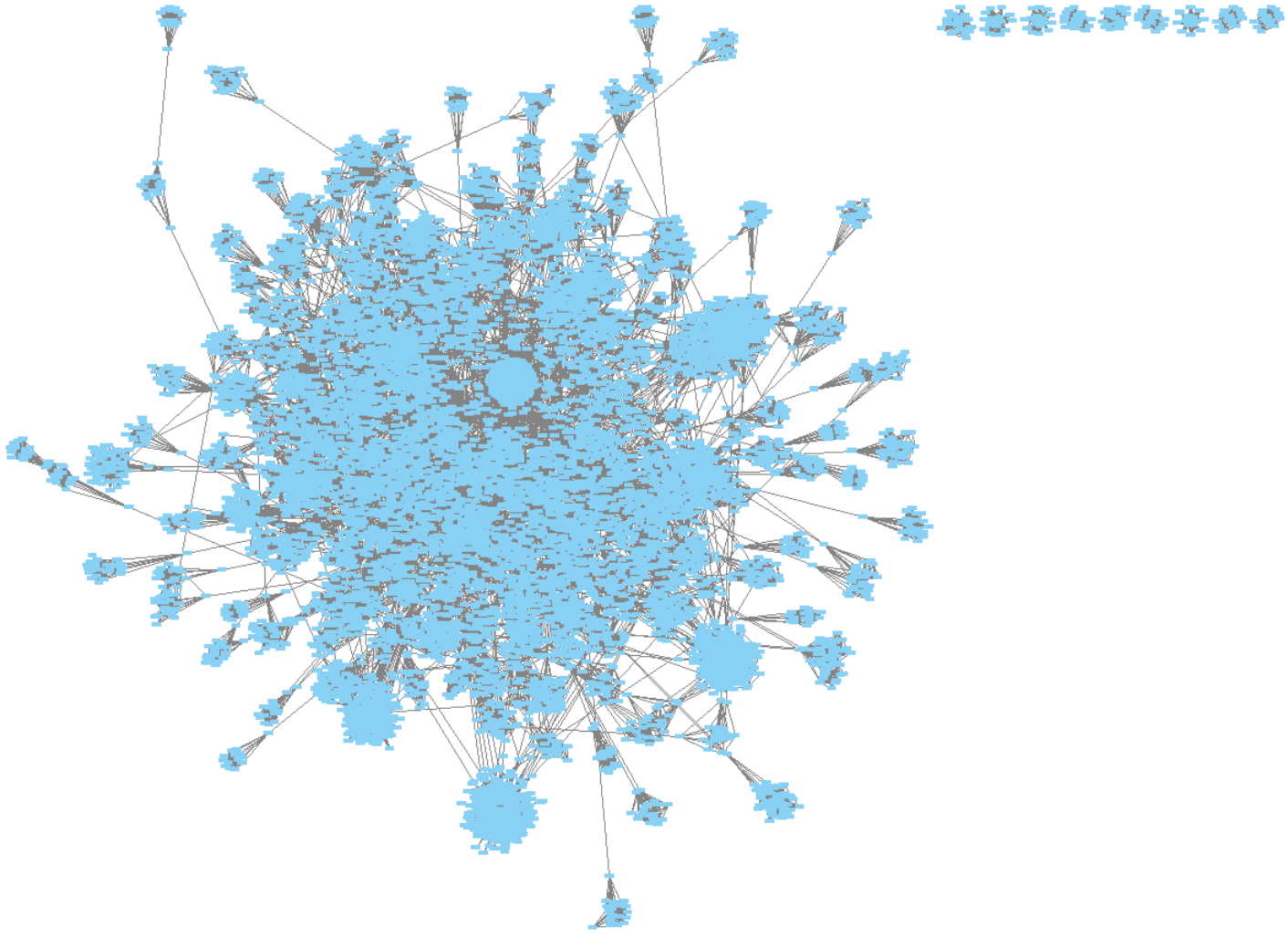
Cas des sommets de degré 3 (suite)

- Le sous-ensemble $\{2, 3\}$ donne la biclique $\{b, c, 2, 3\}$, incluse dans une biclique déjà trouvée.
- Le sous-ensemble $\{2, 4\}$ donne la biclique $\{b, c, 2, 4\}$, incluse dans une biclique déjà trouvée.
- Le sous-ensemble $\{3, 4\}$ donne la biclique $\{b, c, d, e, 3, 4\}$.

Généralisation

- Soit $V = \{v, w, x, y\}$ l'ensemble des voisins du sommet u. V a 11 sous-ensembles contenant au moins deux sommets.
- L'algorithme pourrait donc trouver 11 bicliques maximales lorsqu'il considère le sommet u et 26 bicliques maximales lorsqu'il considère un sommet de degré 5.
- Si on retire les sommets de degré peu élevé itérativement du graphe fourni par Revenu Québec, on constate que le processus trouve très peu de sommets à supprimer à partir du degré 21.

Graphe obtenu par élagage de tous les sommets de degré inférieur ou égal à 6



Approche finale

- On enlève du graphe de Revenu Québec tous les sommets de degré inférieur ou égal à 4 : le graphe ne contient plus que 64000 sommets environ.
- On peut alors utiliser le logiciel de Lu, Phillips et Langston, en vérifiant que les bicliques retournées par ce logiciel n'ont pas déjà été trouvées.
- Il faut s'assurer qu'on n'engendre jamais une biclique qui est un sous-ensemble d'une biclique déjà trouvée, ce qui requiert un examen minutieux des cas pour ne pas faire de travail inutile.

Conclusions

- Les expertises complémentaires des membres de l'équipe ont permis de résoudre un problème a priori très difficile !
- Il faut toujours examiner le graphe sous-jacent au problème afin d'exploiter sa structure dans toute la mesure du possible.