Big Data – Visualisation de graph

GOURRAT Chloé

LAGAILLARDIE Nicolas

Introduction :

Dans ce TP, nous avons étudié deux bases de données sur lesquelles nous avons effectué des modélisations en graph sur le logiciel Gephi.

Récupération des données

Pour sélectionner les données que nous allions utiliser, nous avons cherché les bases publiques en allant sur le site gouvernemental français data.gouv.fr. Nous avons ensuite choisi le thème « Territoires, Transports, Tourisme » où nous avons pris la base de données « Gares ferroviaires de tous types, exploitées ou non » puis le thème « Logement, Développement Durable et Energie » dans lequel nous avons téléchargé la base « Emissions de CO2 et de polluants des véhicules commercialisés en France ».

Traitement des données

Pour pouvoir manipuler les données nous avons dû traiter au préalable les données. Celles-ci provenant d'une organisation francophone, nous avons dû internationaliser les données en convertissant le séparateur des décimal virgule « , » en point « . ».

Nous avons ensuite observé rapidement les données brutes dans un tableur pour vérifier leur « cohérence » en s'assurant par exemple qu'il n'y avait pas trop données manquantes caractérisées par la présence de cases vides ou que les données d'une même colonne correspondait bien à un seul type de données uniforme et commune à toute la colonne.

Manipulation des données

Pour manipuler et visualiser les données, nous avions deux possibilités pour rendre possible leur affichage par Gephi :

- Réduire le nombre d'attributs (colonnes)
- Réduire le nombre d'entrées (lignes)

En effet un trop grand nombre de données interconnectées surchargerait le graph et le rendrait illisible par le trop plein d'information affiché (impossibilité d'à la fois observer les données très éloignées et de distinguer celles qui sont les plus proches). De plus un nombre très important de données conduit aussi à une surexploitation des processus de calcul et donc un ralentissement du logiciel.

La réduction par ligne permet d'avoir un nombre d'éléments moins important et donc chargeable par Gephi, tout en gardant une vue d'ensemble de la base de données. Cela permet de donner une idée

bien que biaisée des différents liens entre les attributs. On a par exemple sélectionné dans le fichier sur les émissions CO2 environ 200 voitures sur les milliers de départ, de même pour les chemins de fer. Pour atténuer le biais dû à la restriction des données à un échantillon, cette sélection s'est réalisée de façon aléatoire.

La réduction par colonne permet de voir au cas par cas les groupements des entrées en fonction de chaque attribut jugé pertinent. On peut ainsi par exemple déterminer les types de voitures généralement construites par chaque entreprises automobiles, si certain types sont plus ou moins représentés ou si certaines entreprises sont plus productives en sélectionnant seulement les colonnes de type de voiture et de nom de constructeur dans la table sur les émissions de CO2.

Exploitation des données

Nous allons dans un premier temps traiter les données sur les lignes ferroviaires.

Pour chaque groupe de données nous effectuons les mêmes premières manipulations en lançant une organisation en « Force Atlas » avec une très forte répulsion pour séparer les données de manière efficace en aérant le graphe et permettre plus de visibilité.

Nous avons commencé par une visualisation générale de la base de données avec un échantillon de 200 entités sélectionnées aléatoirement.

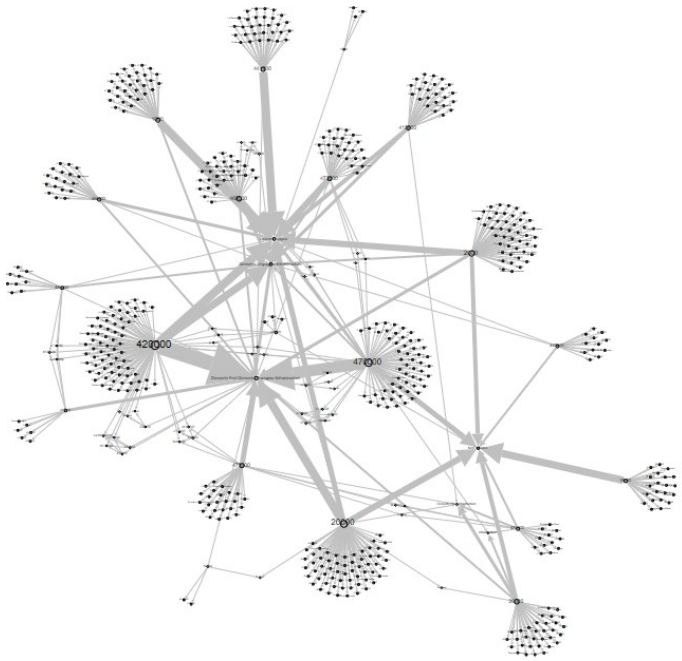


Figure 1 : Vue générale des données de chemins de fer avec un échantillon de 200 entités

On observe sur le graphe trois sortes de liens : les points étant reliés à de nombreux points très proches (ressemblant à des petits bouquets) sont des numéros de lignes ferroviaires associées à toutes les gares que la ligne desserre. Les autres nœuds qui sont reliés aux numéros de ligne indiquent leurs types d'exploitation (ex. desserte de voyageurs ou transport de fret).

On mesure ensuite la modularité de classe dans les statistiques des données. Cette représentation permet de déterminer des communautés qui ont de fortes connexions intra et donc ont une

interprétation intéressante dans le monde réel. Nous avons alors réalisé une partition en couleur des nœuds par cette méthode et avons ainsi obtenu le graphe suivant :

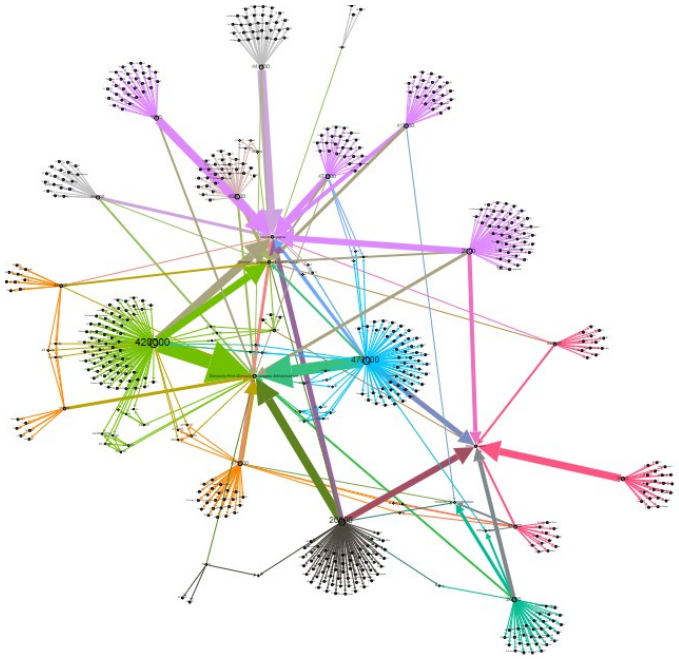
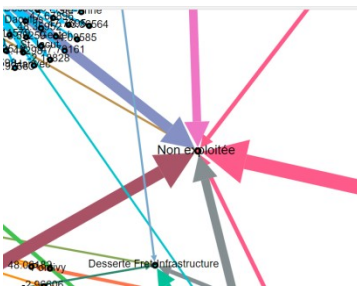
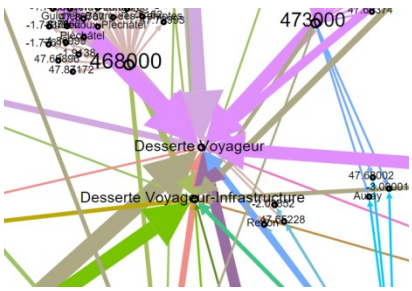


Figure 2 : Partition modulaire de la base sur les gares ferroviaires de France avec un échantillon de 200 lignes de chemin de fer

Analyse de la spatialisation :

La répartition dans l'espace se fait en fonction des affinités des nœuds donc ici en fonction du type d'exploitation réalisée par les lignes. Les lignes situées le plus au sud-est de notre graphe sont les moins exploitées tandis que celles situées au nord desservent le plus souvent des voyageurs et celles plus au centre et au sud-ouest ont l'habitude de transporter des frets.





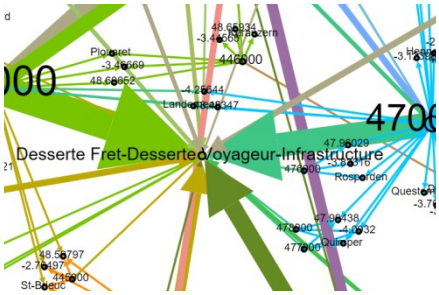


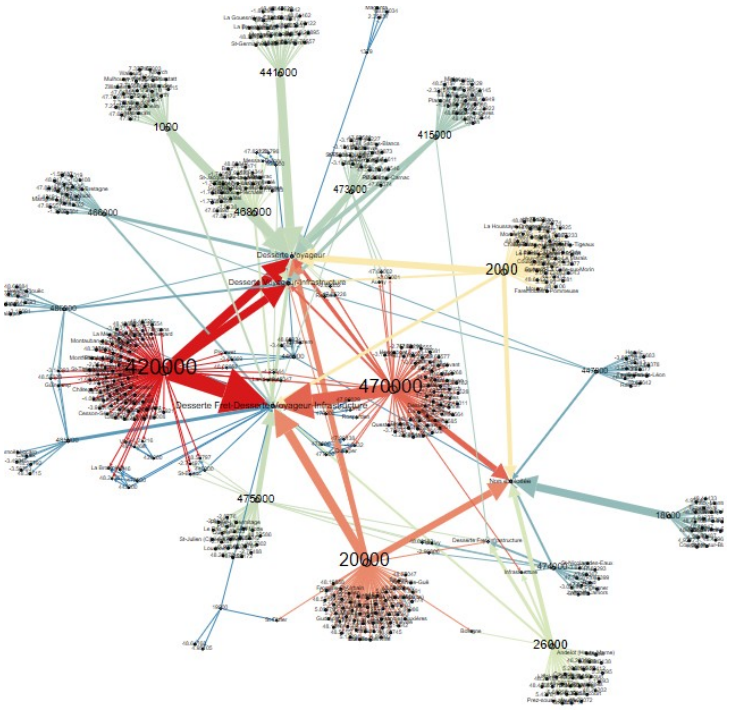
Figure 3 : Zoom parties Sud-est, **Nord et Centre**

Plusieurs gares ne se situent pas dans la périphérie directe d'une ligne. Elles appartiennent à plusieurs lignes différentes donc sont situées entre les deux lignes qui les partagent. On peut ainsi déterminer les gares de correspondance qui sont des gares importantes où plusieurs lignes convergent.

Analyse de la partition :

On observe 5 groupes se former qui correspondent aussi au type d'exploitation principal des lignes.

On réalise aussi une mesure statistique sur le degré sortant pour évaluant la quantité de liens sortant de chaque nœud :



On observe sur ce graphe les lignes servant le plus de gares qui sont en rouge (ex. 420000 représente la ligne Paris-Brest) alors que les lignes plus courtes sont plutôt de couleur bleue (ex. 447000, ligne Morlaix-Roscoff).

On va maintenant analyser les données d'émission de CO2.

On réalise les mêmes évaluations statistiques que dans l'exemple précédent : la modalité et le degré de connexions.

Dan un premier temps on garde tous les attributs puis on se focalisera sur certains pour obtenir de nouveaux graphes.

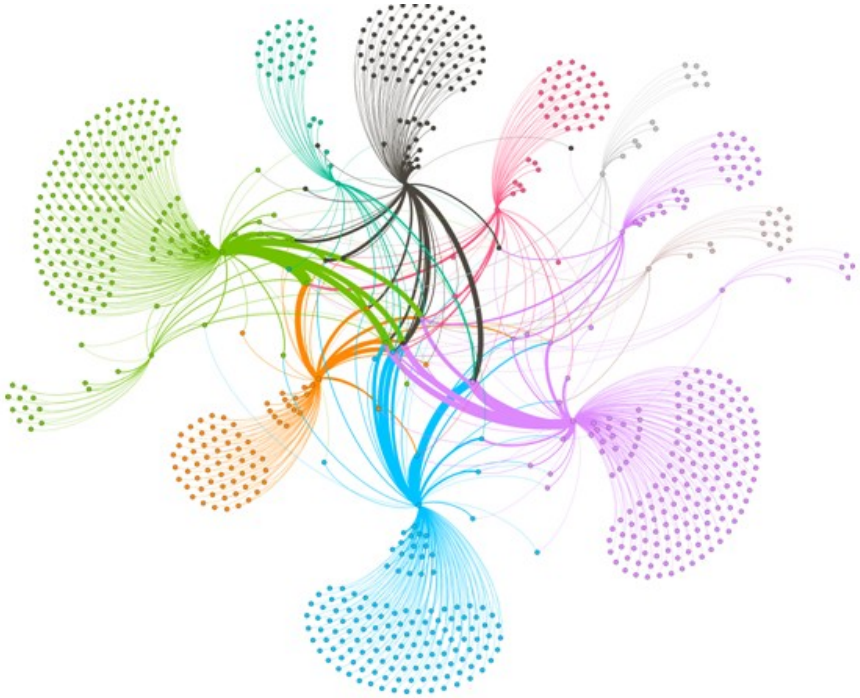


Figure 4 : Graphe modulaire générale des données d'émissions CO2

Les graphes généraux présentent beaucoup trop de liens et d'information pour être lisibles mais on peut tout de même y faire quelques observations. Dans le graphe modulaire, les couleurs permettent de distinguer les différentes marques. Parfois deux marques sont de mêmes couleurs indiquant une relation très forte entre elles avec beaucoup de type de voiture en commun (ex en violet sont représentées les marque BMW, Bartley et Aston Martin qui sont des constructeurs de voitures de luxe et sportives pour les 2 premiers). La répartition spatiale correspond aussi aux tendances de chaque constructeur avec à l'est les marques les plus onéreuses, de luxe ou sportives (ex. BMW, Cadillac), à l'ouest les marques les plus abordables et familiales (ex. Fiat, Dacia) et au nord les marques tout-terrain (ex. Chevrolet).

Les nœuds les plus en périphérie sont les noms de modèles précis. Ils sont isolés car ils sont uniques. Les nœuds aux origines des bouquets sont les marques de constructeurs automobiles ayant construits ces modèles. La taille et l'importance des bouquets peuvent ainsi indiquer la productivité de ces constructeurs en termes de variété de modèles. Cette information peut être quantifiée avec un classement par degré sortant et pourra par la suite être déterminée plus précisément avec un graphe plus spécifique. Les nœuds des bouquets les plus proches du cœur sont les types de modèle qui, plus ils sont proches, plus ils sont représentés par la marque.

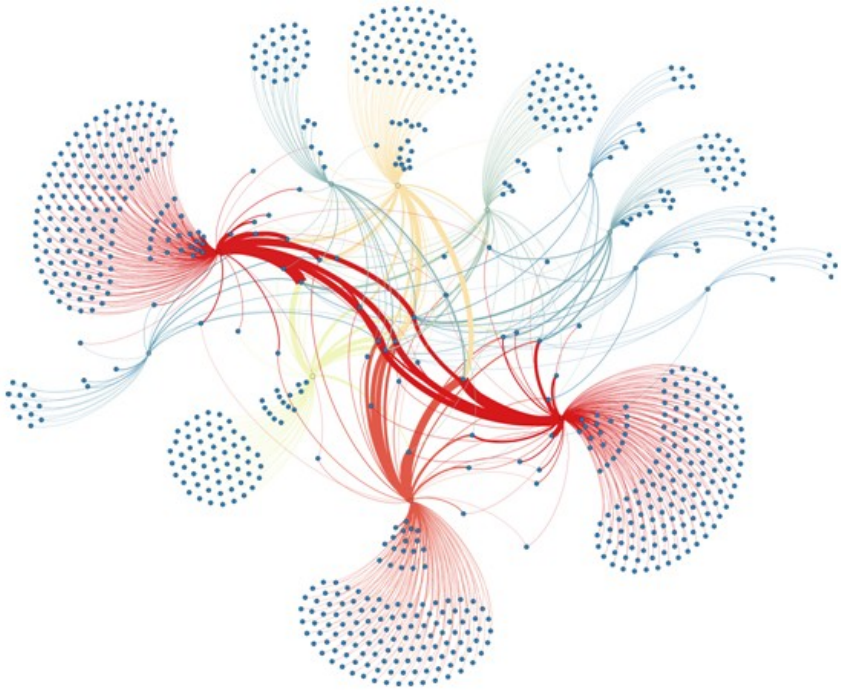


Figure 5 : Graphe général des données d'émissions C02 classées par degré sortant

Dans le graphe classé en degré sortant, les marques en rouge sont les plus productives (ex. Fiat, BMW et Audi) celles en jaune un peu moins (ex. Citroën et Alfa-Romeo) et celles en bleu sont celles ayant le moins de modèles et donc le moins présentes sur le marché automobile car trop spécifique et ciblant une faible partie des consommateurs (ex. Ferrari, Cadillac) ou qui se concentrent sur le renouvellement de pièces plutôt que la création de nouveau modèles (ex. Dacia).

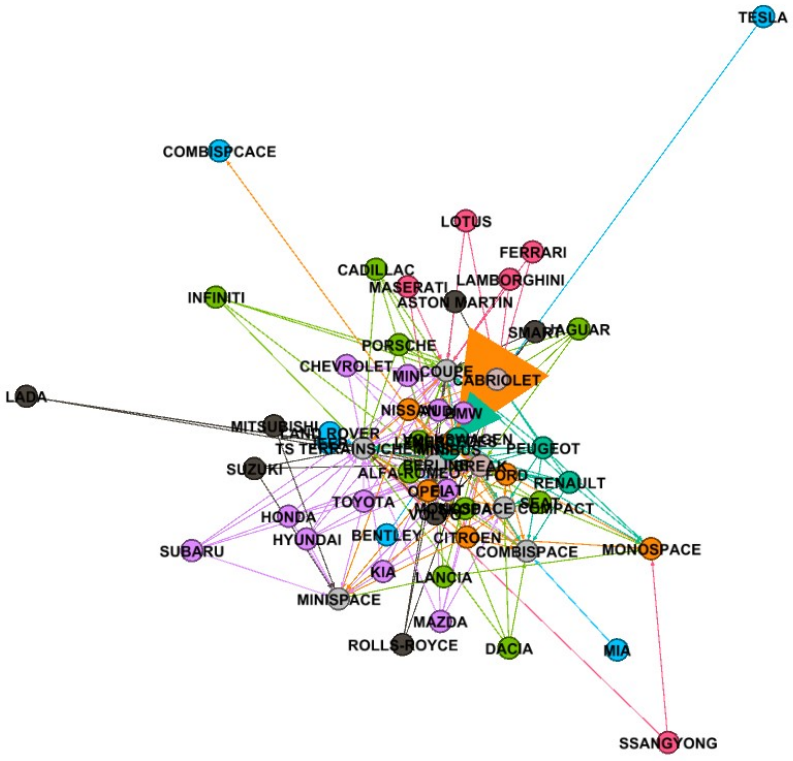


Figure 6 : Graphe général avec les noms des constructeurs et le type de voiture

Dans le graphe précédent, nous pouvons observer les liens entre les constructeurs et les différents types de voiture qu'ils construisent. En périphérie, nous avons les constructeurs qui construisent assez peu de type de voiture, à l'image de Tesla qui ne réalise que des cabriolets, ainsi que les type de voiture construit par assez peu de marques, telle que le combispace.

Nous réalisons ensuite une analyse du degré des nœuds, et voici le graphe que nous renvoi Gephi :

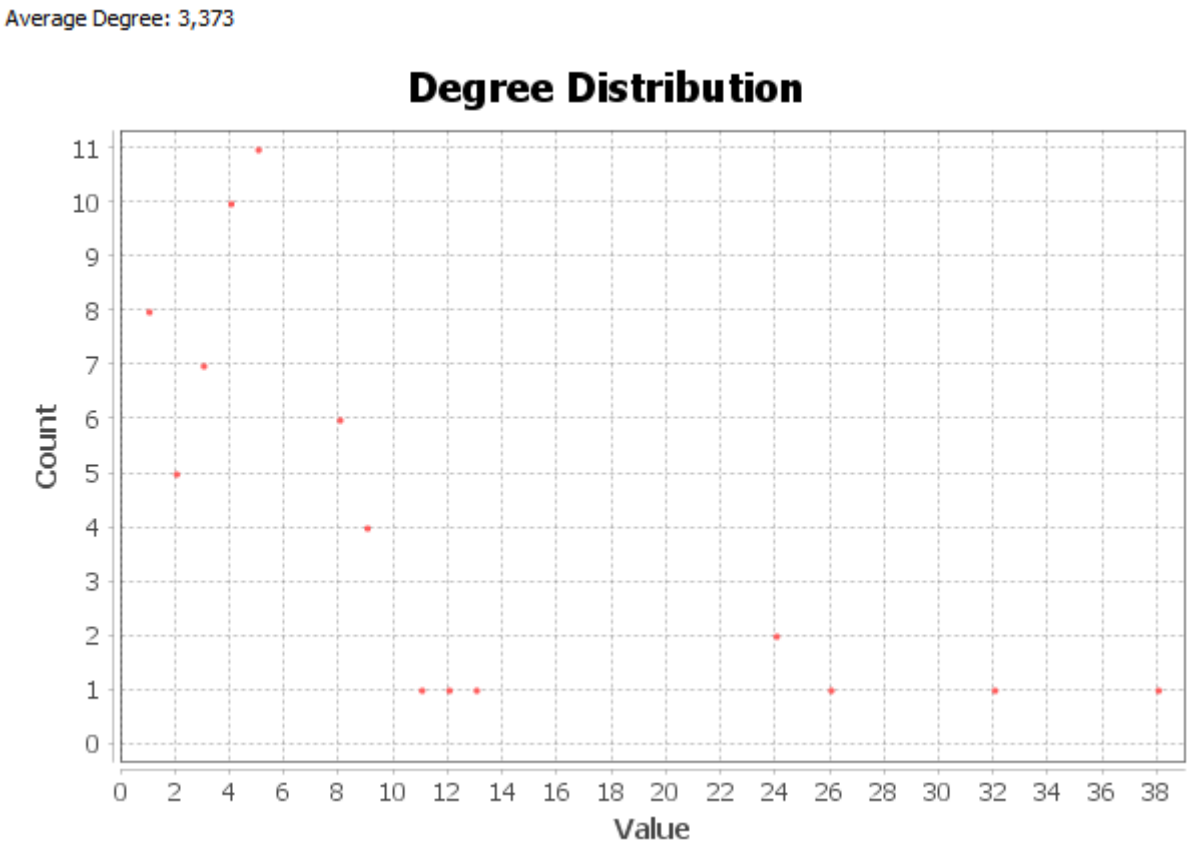


Figure 7 : distribution du degré

Nous observons qu'il existe quelques nœuds avec énormément de liens, par exemple 5 d'entre eux ont 11 liens. D'autre part, un très grand nombre de nœuds possède un unique lien. Si nous cherchons à connaître les nœuds avec un fort degré, nous pouvons zoomer sur le centre du graphe. Il faut ensuite réduire la taille des liens affichés car certains d'entre eux sont au moins 11 fois plus gros que les plus petits. Voici ci-après la figure que nous obtenons après ces manipulations.

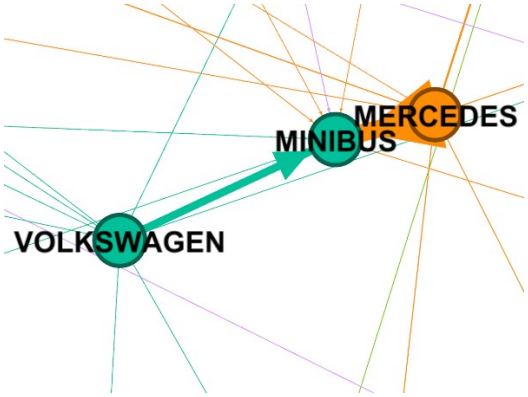


Figure 8 : zoom sur le graphe

Nous constatons que Volkswagen et Mercedes construisent énormément de type de voitures *Minibus*. Après une rapide recherche sur Internet, il s'avère qu'ils sont leaders dans la vente de Minibus, ce qui confirme nos constations.

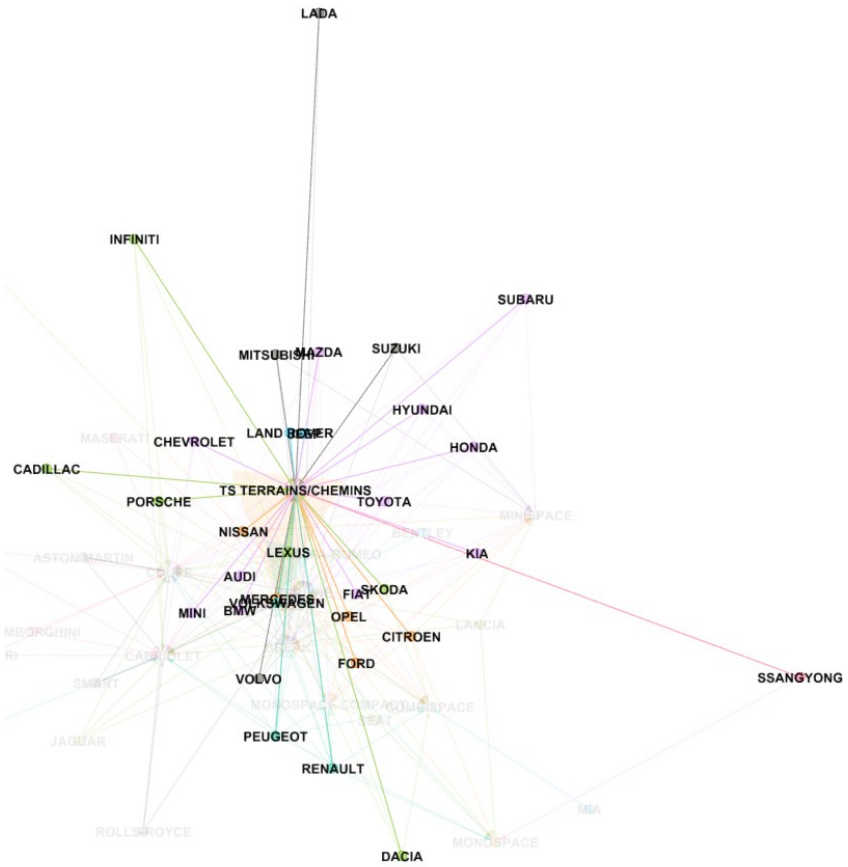


Figure 9 : type de voiture construit par le plus de marques différentes

Voici ci-dessus un autre graphe, qui permet de se rendre compte du type de voiture le plus construit : TS terrains/chemins (tout terrain / tout chemin). Ce constat est inattendu, en effet, les mentalités françaises tendent à privilégier des voitures économes plutôt que des voitures type Tout terrain, beaucoup plus gourmandes. Cependant, ce type de voiture est apprécié par les entités gouvernementales telles que les forces de l'ordre ou de sécurité : leur capacité à aisément accéder aux différents théâtres d'opérations est un véritable plus, malgré leur consommation.

En conclusion

Gephi nous a permis de mettre en lumière des phénomènes très diverses. C'est un outil graphique qui permet à tous de comprendre et lier des éléments qui, au premier abord, semblent totalement décorrélés. Au-delà des deux bases de données que nous avons étudiées, nous pouvons nous intéresser à des domaines très divers, ce qui permettrait de comprendre de nombreux phénomènes.