

Atelier Gephi

Qu'est ce qu'un graphe ?

Les graphes de réseaux s'intéressent à des groupes d'entités, qui peuvent être des individus, mais aussi des mots, des concepts, des villes etc. et surtout aux relations qu'ils entretiennent entre eux.

Une relation peut être à peu près n'importe quoi, les retweets entre utilisateurs, les similarités entre images, les trajets de transports entre des villes etc.

Ils ont gagné en importance avec l'apparition du web, en tant qu'outil permettant le mieux de représenter des réseaux de sites se citant eux mêmes aux travers de liens http.

Or se rendre compte de comment le web se construisait était extrêmement difficile avec une simple visualisation sous forme de liste. Les graphes ont offert un nouvel axe d'analyse.

On a ainsi pu visualiser comment des communautés de sites / forums, se créaient sur internet de façon parfois assez organique.

Cet effet s'est encore accentué avec l'apparition des réseaux sociaux.

Un peu de jargon

Il est composé de:

- nœuds, qui servent à représenter les entités
- liens qui servent à représenter les relations, les liens peuvent:
 - être dirigés, dans ce cas souvent représentés par une flèche
 - non dirigés, représentés par un trait

On parle de **degré** pour indiquer le nombre de liens connectés à un nœud.

On parle de **poids** pour indiquer la force du lien entre deux nœuds.

On parle aussi de cartographie lors de la réalisation d'un graphe légendé sur un domaine.

Les degrés de séparation

S'intéresser aux relations entre les entités, c'est aussi (surtout ?) s'intéresser à la répartition du réseau et à la distance entre les individus.

Bien avant Gephi des chercheurs se sont intéressés à la distance de communication entre les êtres humains.

https://fr.wikipedia.org/wiki/%C3%89tude_du_petit_monde

Cette expérience a mené à une définition des 6 degrés de séparation entre n'importe quelles personnes.

Aujourd'hui les groupes de réseaux sociaux et notamment Facebook, recalculent régulièrement ce chiffre à partir du graphe de leurs utilisateurs.

5 en 2008, 4.5 en 2012, 3.5 en 2016 faisant miroiter l'idée d'un monde toujours plus resserré, plus connecté, là où d'autres études de réseaux montrent au contraire des groupes de plus en plus polarisés. (ce qui n'est pas incompatible en réalité mais tout dépend de l'interprétation que l'on en fait).

Un autre exemple de cela est le monde cinématographique, avec le site l'Oracle de Bacon:

<https://oracleofbacon.org> qui permet de connaître les degrés de séparation entre deux personnes du monde du cinéma.

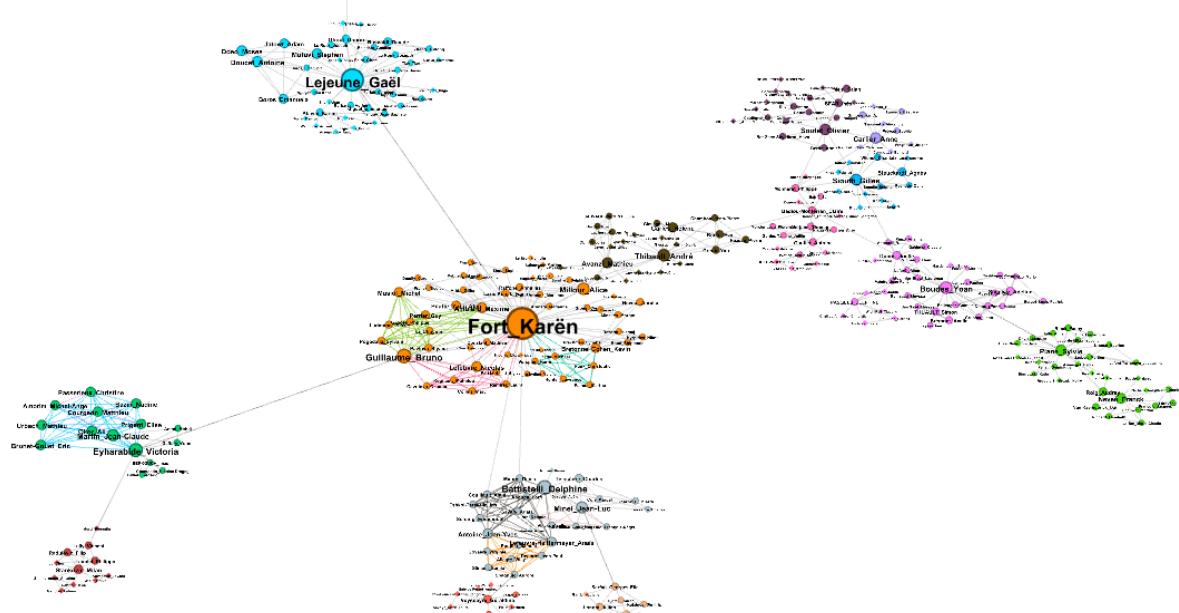


exemple entre Canet et Depp

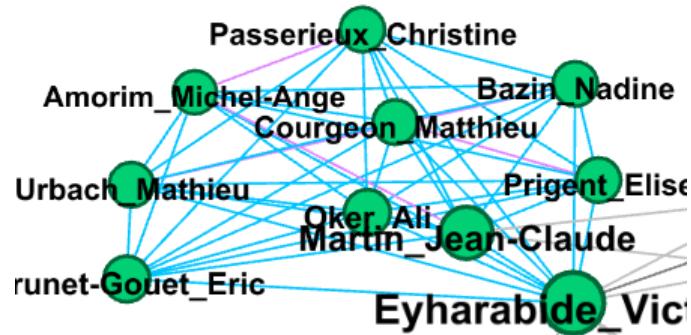
Topologies types

Il existe différents types caractéristiques de réseaux, qui permettent de façon assez rapide et visuelle d'établir des analyses sur le comportements des entités entre elles.

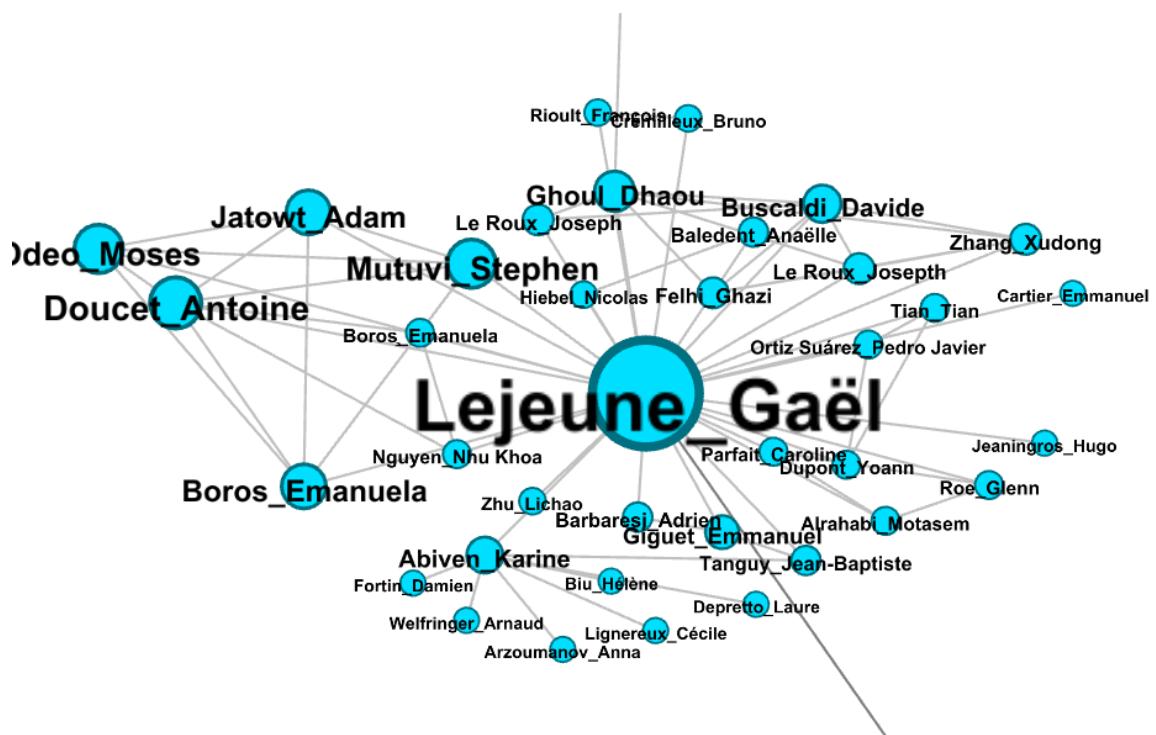
- étalé: réseau peu connecté, regroupé ou non en clusters, les chemins les plus longs sont assez grands (17 en l'occurrence)



- très connexe: réseau très dense mais peu hiérarchisé, les individus ont tous un rôle hiérarchique identique et communiquent entre eux



- très hiérarchique: réseau centré autour "d'influenceurs", c'est ce vers quoi tend le web avec un web de plateforme



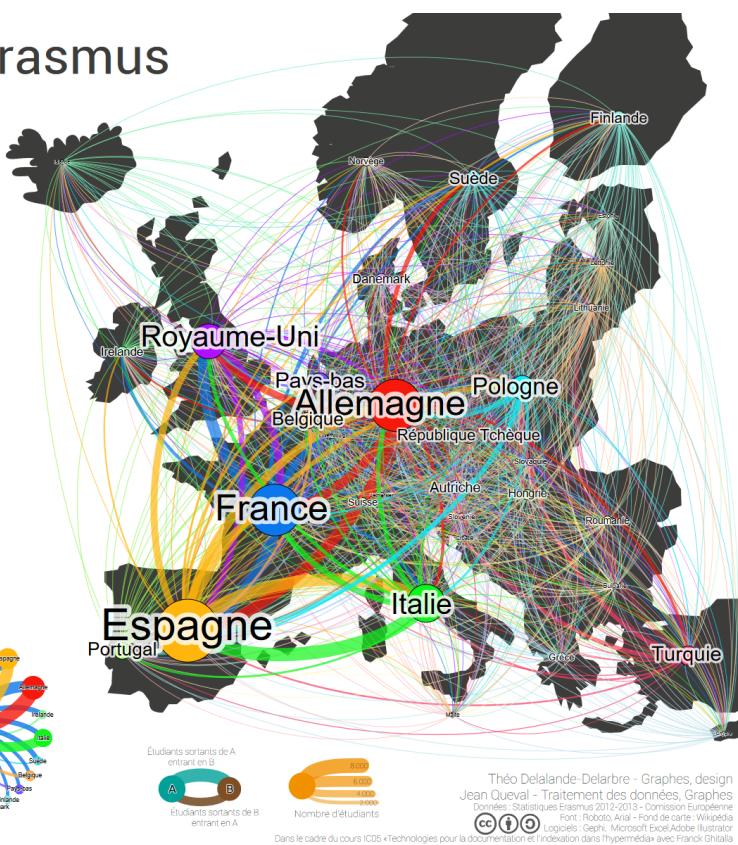
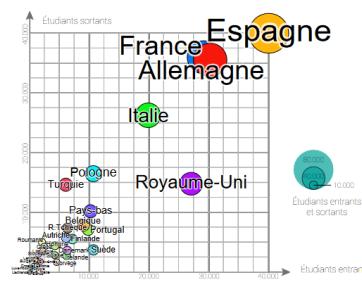
- uniquement des clusters non reliés entre eux, pas vraiment de réseau, uniquement des sous communautés

Exemples d'usages

Graphes des échanges des étudiants Erasmus en 2016, qui est un exemple de graphe contraint

Les Mobilités Erasmus

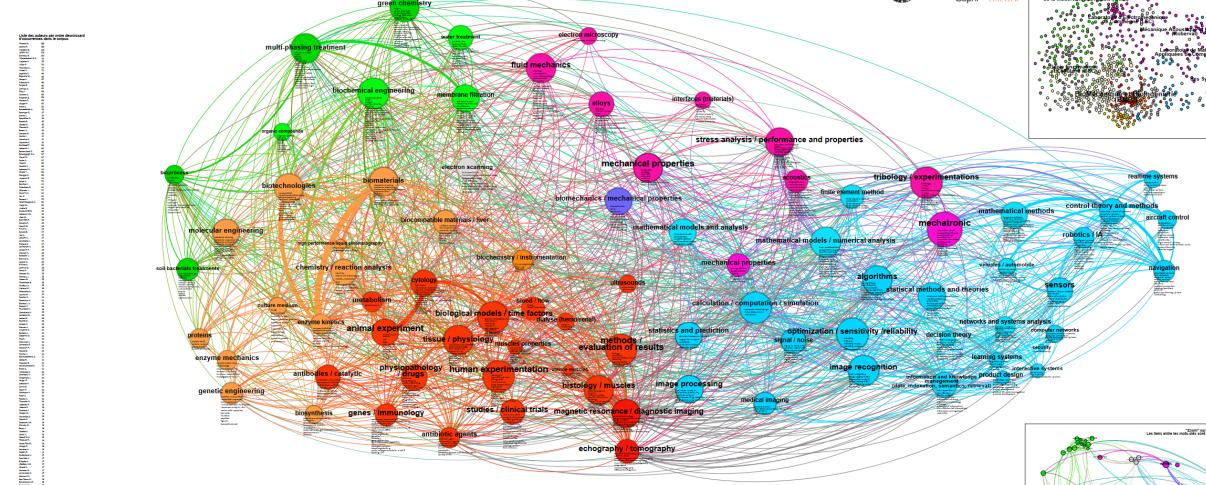
Statistiques d'entrées et sorties



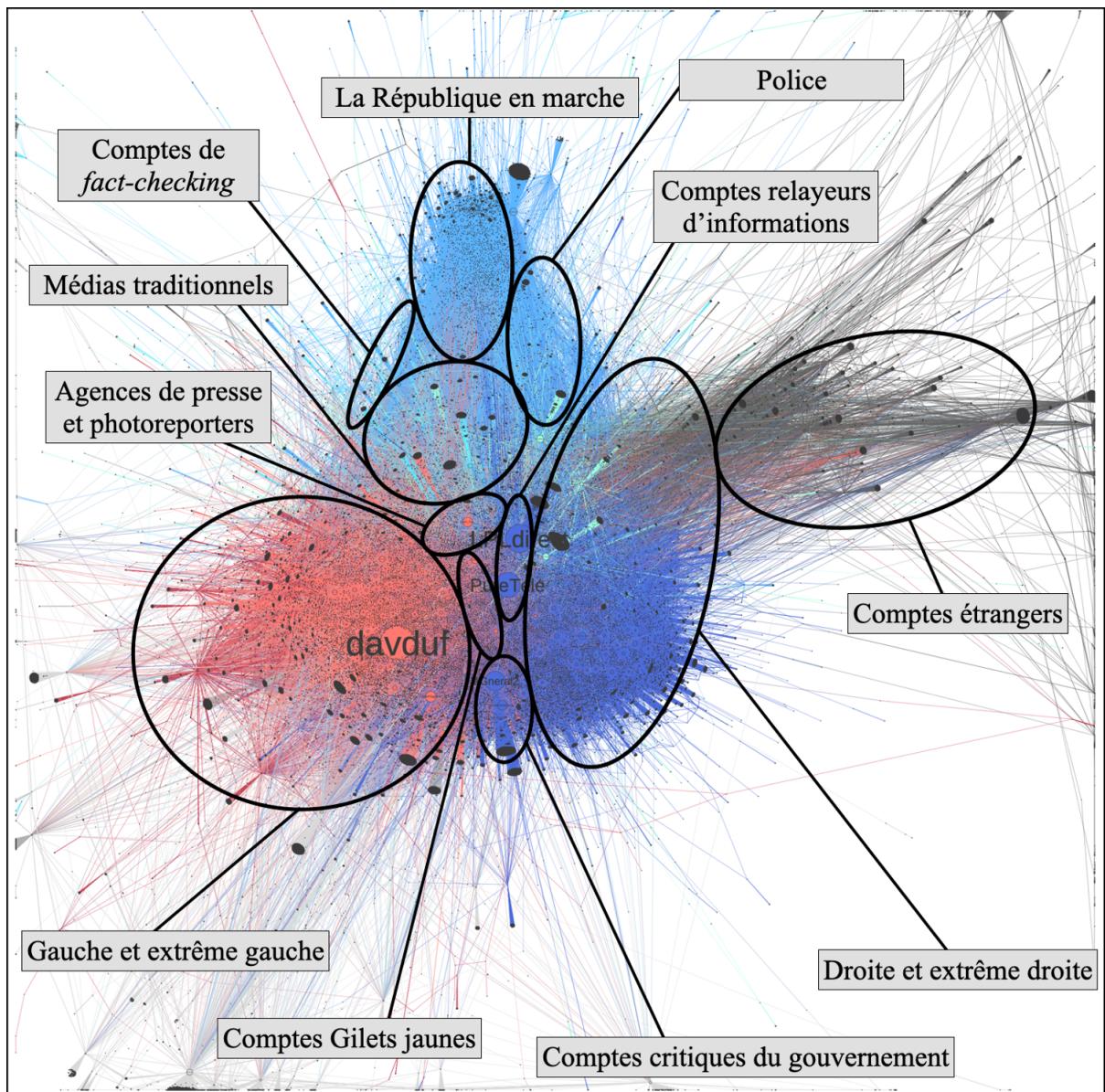
Les cooccurrences de termes dans les recherches d'une université pendant 40 ans:

40 années de publication scientifique à l'Université de Technologie de Compiègne

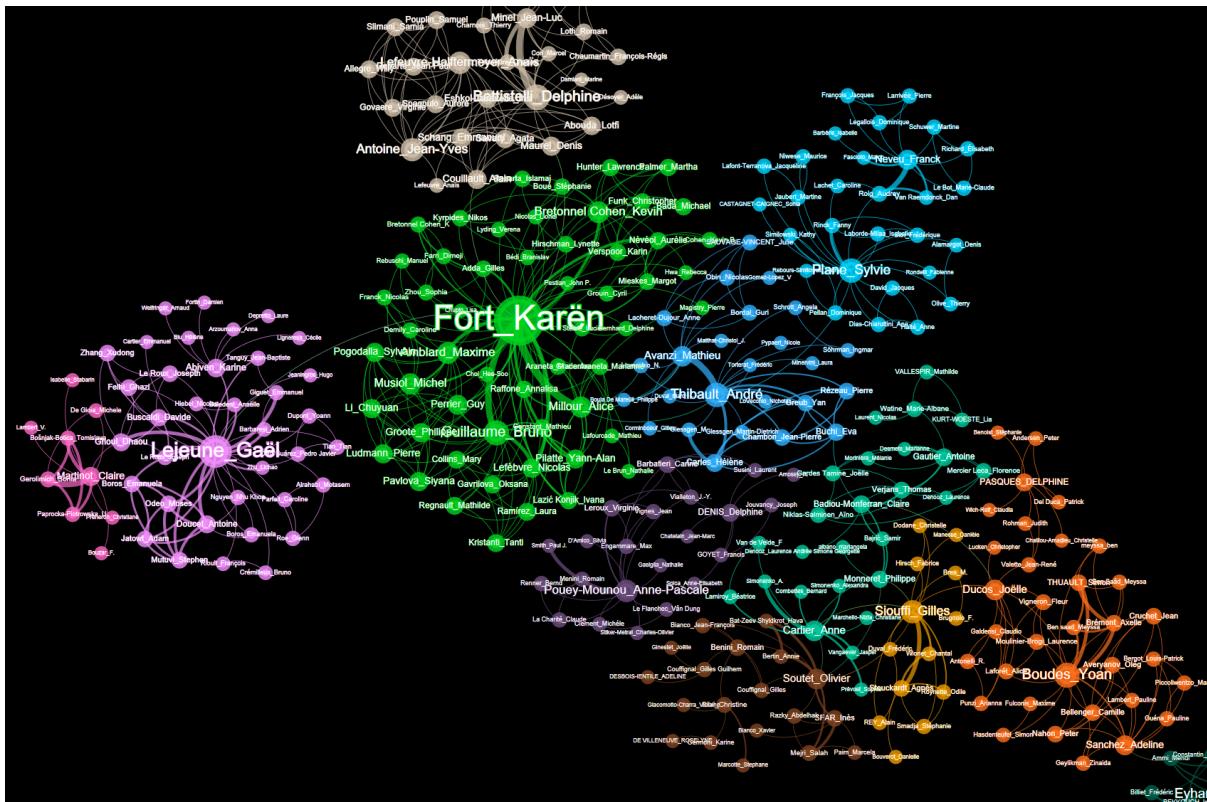
81 thématiques calculées sur plus 5500 notices issues de SCOPUS



Les utilisateurs partageant des images de violences policières sur Twitter:



Le graphe des co-auteurs au sein du laboratoire STIH:



Autres sujets d'application

- biologie, réseaux de molécules
- analyse politique (exemple du travail de linkurious lors des pandora papers: https://fr.wikipedia.org/wiki/Étude_du_petit_monde)
- en urbanisme, sur les transports / flux de personnes
- analyse des réseaux terroristes

Gephi

Gephi est un logiciel **open source** permettant de construire, visualiser et spatialiser des graphes de réseau. Il offre un **très grand nombre** de fonctionnalités, notamment de calculs de statistiques sur le réseau généré, et nous n'aurons clairement pas le temps de tout voir. Nous verrons ici les fonctionnalités de base afin de créer, modifier et exporter un graphe, mais il faut savoir que pour produire des graphes de qualité

Format des données d'entrée

La seule source de données nécessaire est un tableau .csv, .xsl, .xlsx contenant une liste de liens.

Données obligatoires

Les données obligatoires sont les colonnes Source et Target. Ainsi si je veux créer un lien entre deux noeuds A et B mon tableau ressemblera à ceci:

Source	Target
A	B

Ce qui donnera un graph sous la forme:

```
flowchart LR;
A-->B;
```

Je peux ensuite le compléter pour rajouter d'autres liens / d'autres noeuds. Attention à toujours appeler un même noeud de la même façon, sinon les liens seront erronés!

Source	Target
A	B
B	C
A	C
D	C

Ce qui donnera:

```
flowchart LR;
A-->B;
B-->C;
A-->C;
D-->C;
```

Données optionnelles

Dans un tableau d'entrée, on peut rajouter autant de colonnes de données que l'on souhaite, elles n'auront a priori pas d'impact (mais permettront éventuellement de sélectionner certains nœuds directement dans gephi).

Néanmoins certaines colonnes sont réservées à Gephi et permettent d'ajouter des informations supplémentaires:

- Label: permet d'afficher un nom au lien
- Weight: permet de spécifier le poids d'un lien, utile si par exemple on souhaite donner plus d'importance à certaines liens qu'à d'autres

Pour encore plus de données additionnelles il est également possible d'importer un fichier de nœuds en plus du fichier des liens. Il faut néanmoins faire attention à ce que les identifiants des nœuds soient bien les mêmes que les champs Source et Target de la table des liens!



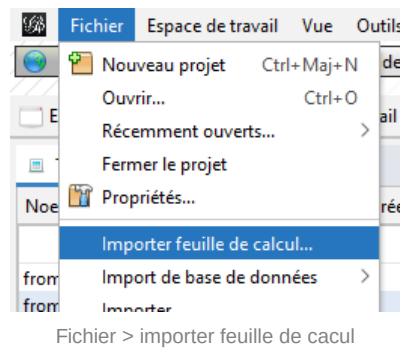
Avoir une table des nœuds permet par exemple de donner des informations tel que le type de nœud, ou d'associer des urls ou des images aux nœuds. Ce qui permettra de rendre ces informations disponibles lors d'un clic sur le nœud lors de certains exports.

Essayez ainsi de créer un tableau avec une dizaine de liens et 5/6 nœuds et de l'enregistrer au format csv.

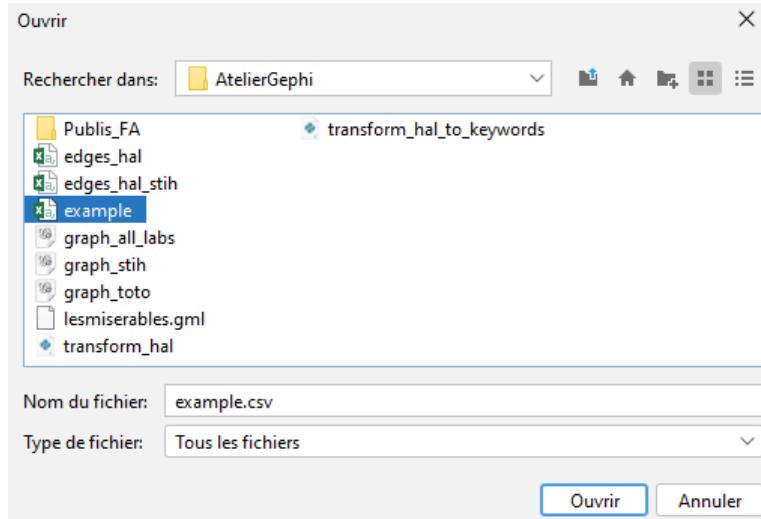
Importer ses données dans Gephi

Importation

Le processus est assez simple:



Fichier > importer feuille de calcul



Choisir le fichier csv / excel à ouvrir

Etapes

1. Options générales du CSV
2. Paramètres d'import

Options générales du CSV (1 sur 2)

Choisissez un fichier CSV à importer :

D:\Alie\Documents\Projets\AtelierGephi\example.csv

Séparateur : Importer en tant que : Encodage :

int-virgule Table des liens UTF-8

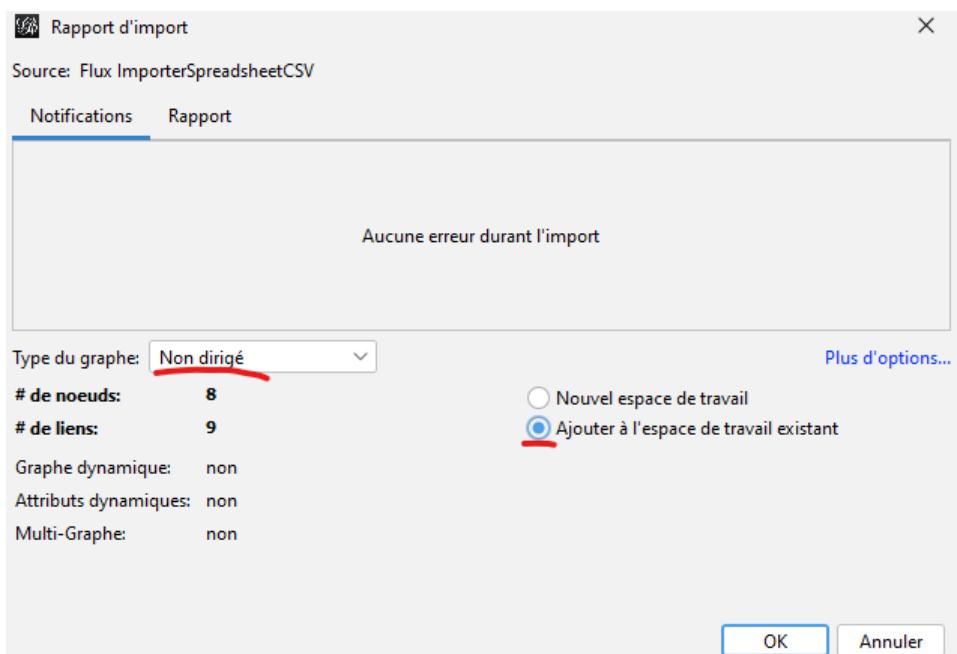
Prévisualisation :

Source	Target
Toto	Tata
Tata	Titit
Titit	Titi
Toto	Tutu
Tete	Titi
Bloup	Tutu

< Précédent Suivant > Terminer Annuler Aide

Bien vérifier que Table des liens est sélectionné et que les colonnes Source et Target sont bien reconnues

A l'étape d'après ne rien sélectionner, puis appuyer sur 'Terminer'



Enfin, sélectionner non dirigé, et ajouter à l'espace de travail existant

Visualisation sous forme de tableau

Une fois les données importées, ces dernières sont automatiquement affichées dans Gephi dans l'espace "Laboratoire de données", pour chaque graphe deux sous tableaux sont créés, une table des liens, et une table des nœuds. On peut switcher de l'une à l'autre comme ceci:

Source	Destination	Type
Toto	Tata	Non dirigé
Tata	Titi	Non dirigé
Titi	Tutu	Non dirigé
Toto	Tutu	Non dirigé
Tete	Titi	Non dirigé
Bloup	Tutu	Non dirigé
Blap	Bloup	Non dirigé
Blap	Bloup	Non dirigé
Blap	Blap	Non dirigé

Première Visualisation

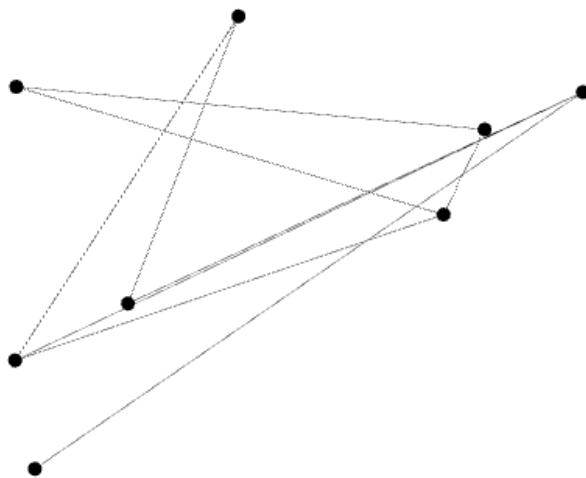
Une fois ces données importées on va pouvoir visualiser leur rendu sous forme de graphe en cliquant sur l'onglet "Vue d'ensemble" qui est le deuxième espace principal de Gephi et permet de visualiser ainsi que de manipuler

le graphe.

Fonctions de base dans la Vue d'ensemble

- Clic droit maintenu pour bouger dans le graphe
- Molette de la souris pour zoomer / dezoomer
- Clic gauche pour sélectionner un nœud

A première vue, le résultat est peu esthétique:



L'idée va ainsi être de **spatialiser** le graphe pour le rendre plus interprétable (même si il n'y a rien à interpréter pour mon graphe d'exemple.) Dans le cas d'un petit nombre de noeuds cela peut même être fait à la main en sélectionnant l'outil "Déplacer" et en bougeant les noeuds pour les réorganiser.



Dans les autres fonctionnalités de base utile, on peut également faire clic droit sur un noeud avec l'outil de sélection (le première icône de la barre des tâches) et supprimer le noeud, ou encore le sélectionner directement dans le tableau des données.

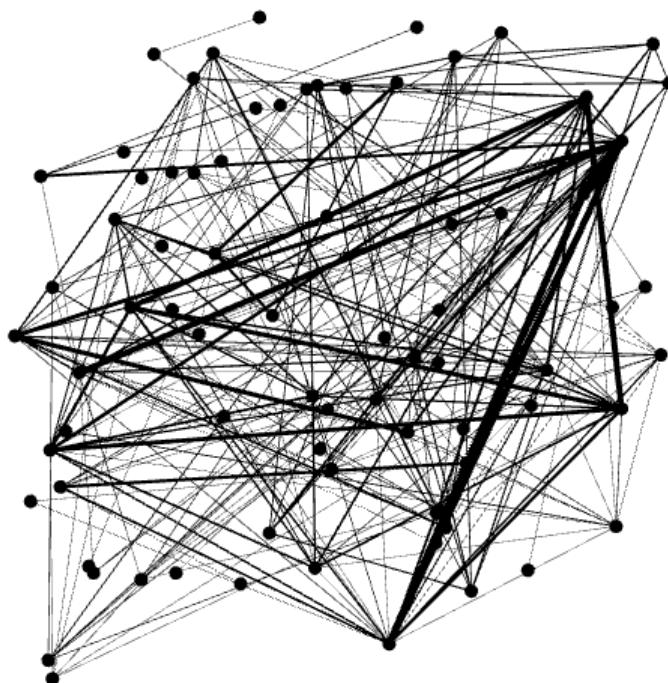
N.B: inversement, il est possible depuis le tableau des données de sélectionner directement un nœud sur le graphe avec clic droit sur la ligne.

Spatialiser ses données

Pour la suite de ce TD, on utilisera le jeu de données misérables. Comme précédemment il faut importer les données, cette fois ci des informations additionnelles ont été stockées au sujet des nœuds du graphe, il y a donc un fichier de nœuds à importer, ainsi qu'un fichier de liens.

Veillez à mettre les deux jeux de données dans le même espace de travail, et d'importer en tant que graphe non orienté.

On retourne à présent sur l'onglet de visualisation et on obtient la vue suivante:



On va utiliser à présent utiliser l'algorithme de spatialisation Force Atlas 2.

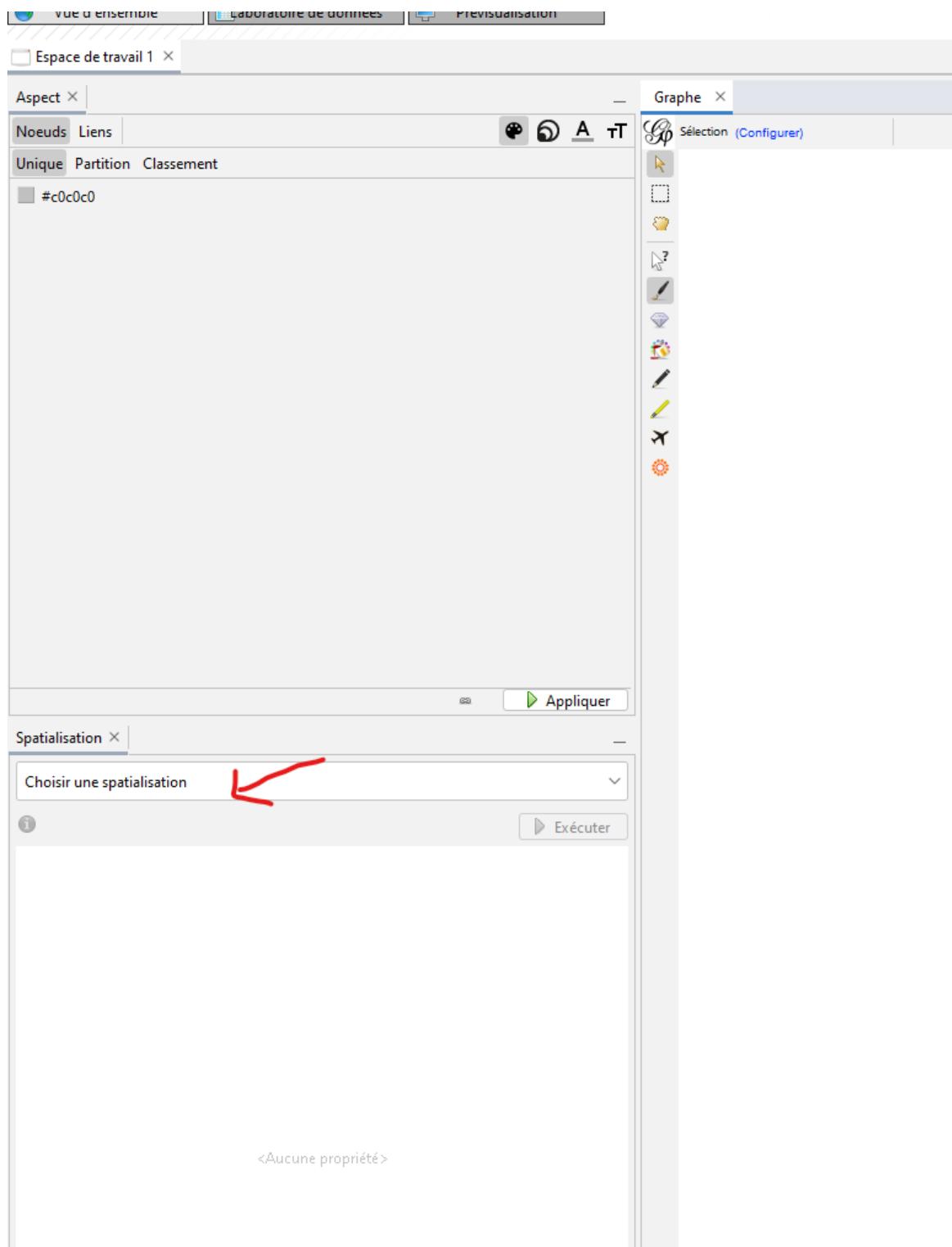
ForceAtlas2 et ses paramètres

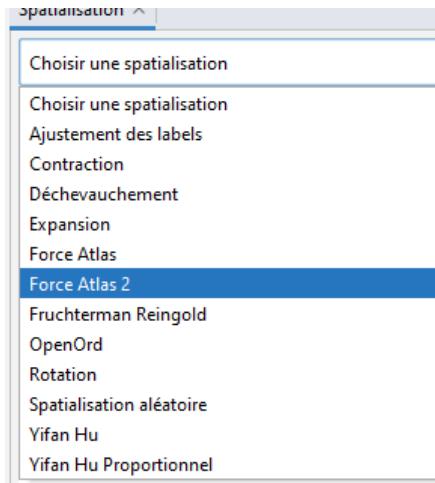
ForceAtlas2 est un algorithme de spatialisation non contraint, c'est à dire qu'aucun nœud n'est figé dans l'espace comme il peut l'être dans le cas de graphes géographiques (cf le graphe erasmus du début).

Son fonctionnement général est le suivant:

- les nœuds reliés entre eux par un lien sont attirés en fonction du poids du lien
- les nœuds non reliés se repoussent

Cela va donner un résultat où les nœuds connectés entre eux vont se regrouper et former plus facilement des communautés.





Sans changer les paramètres, cliquez maintenant sur "Executer".

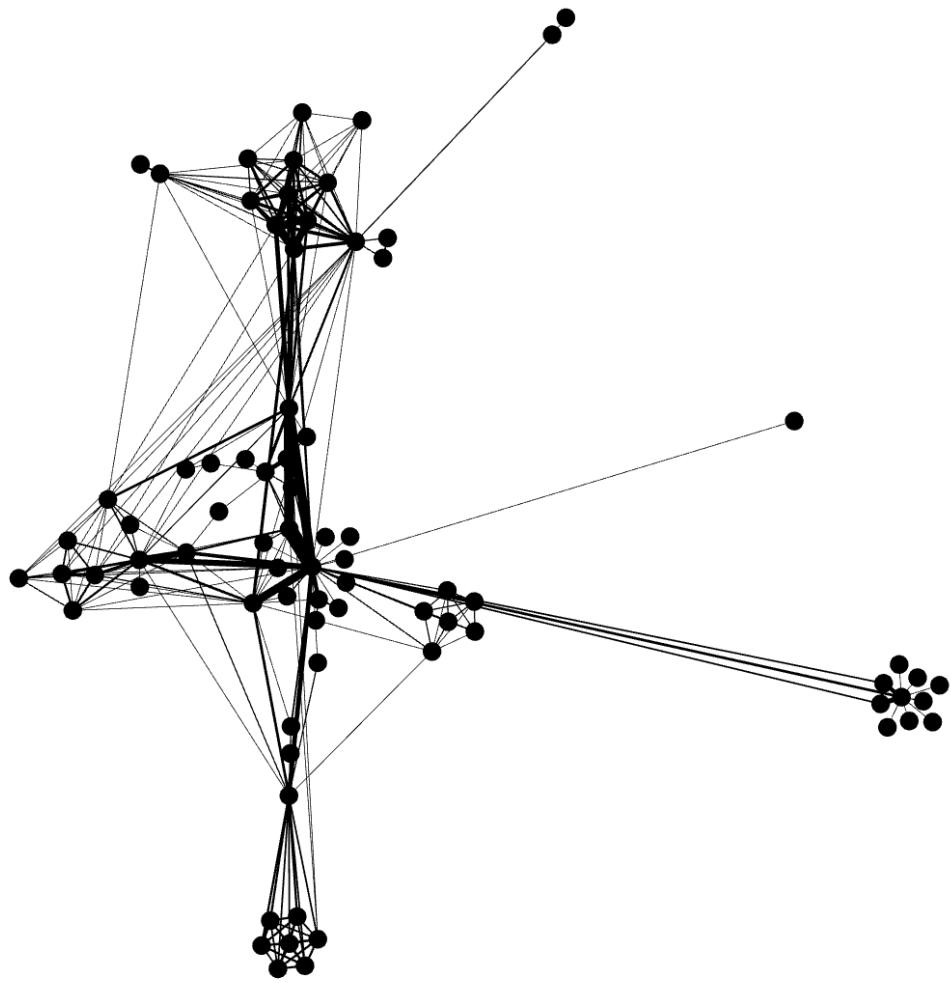
Vous obtenez une première spatialisation.

Maintenant essayez de jouer avec différents paramètres, vous pouvez par exemple passer en mode lin log pour éclater un peu plus le graphe, ou encore jouer avec l'échelle ou dissuader les hubs.



N'activez pas "Empêcher le recouvrement" pendant que la spatialisation est en cours cela ralentit énormément le calcul!

En mode linlog, avec une échelle de 2, en dissuadant les hubs et en appliquant empêcher le recouvrement une fois que le graphe s'est stabilisé on obtient la représentation suivante:

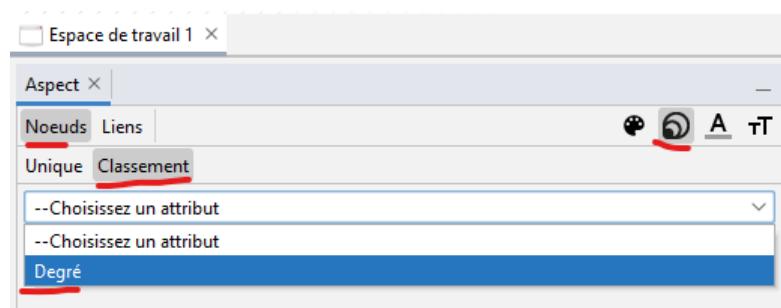


Ce qui semble être une spatialisation intéressante puisque les nœuds sont visiblement regroupés par clusters. Tâchons à présent de le rendre plus lisible.

Customiser son graphe

Changer la taille des nœuds

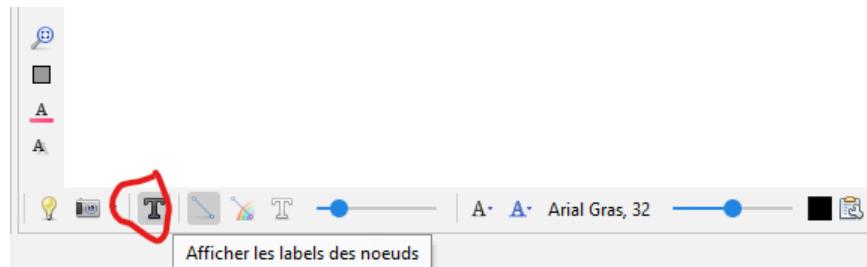
Il est intéressant de représenter la taille des noeuds en fonction de leur importance, cette importance peut être caractérisée par plusieurs critères mais le plus courant reste le degré:



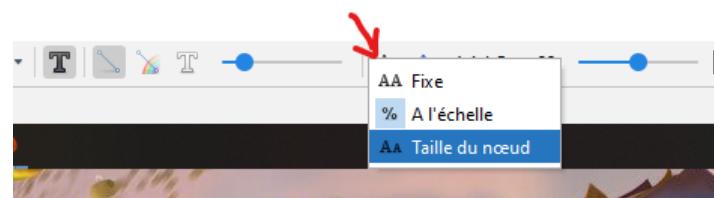
On pourra ensuite choisir les bornes 5 à 30 comme valeurs. Puis cliquez sur Appliquer.

On voit à présent que les noeuds sont un peu l'un sur l'autre à cause du changement de taille, on peut donc réappliquer brièvement ForceAtlas2 avec "Empêcher le chevauchement" pour palier à ce problème.

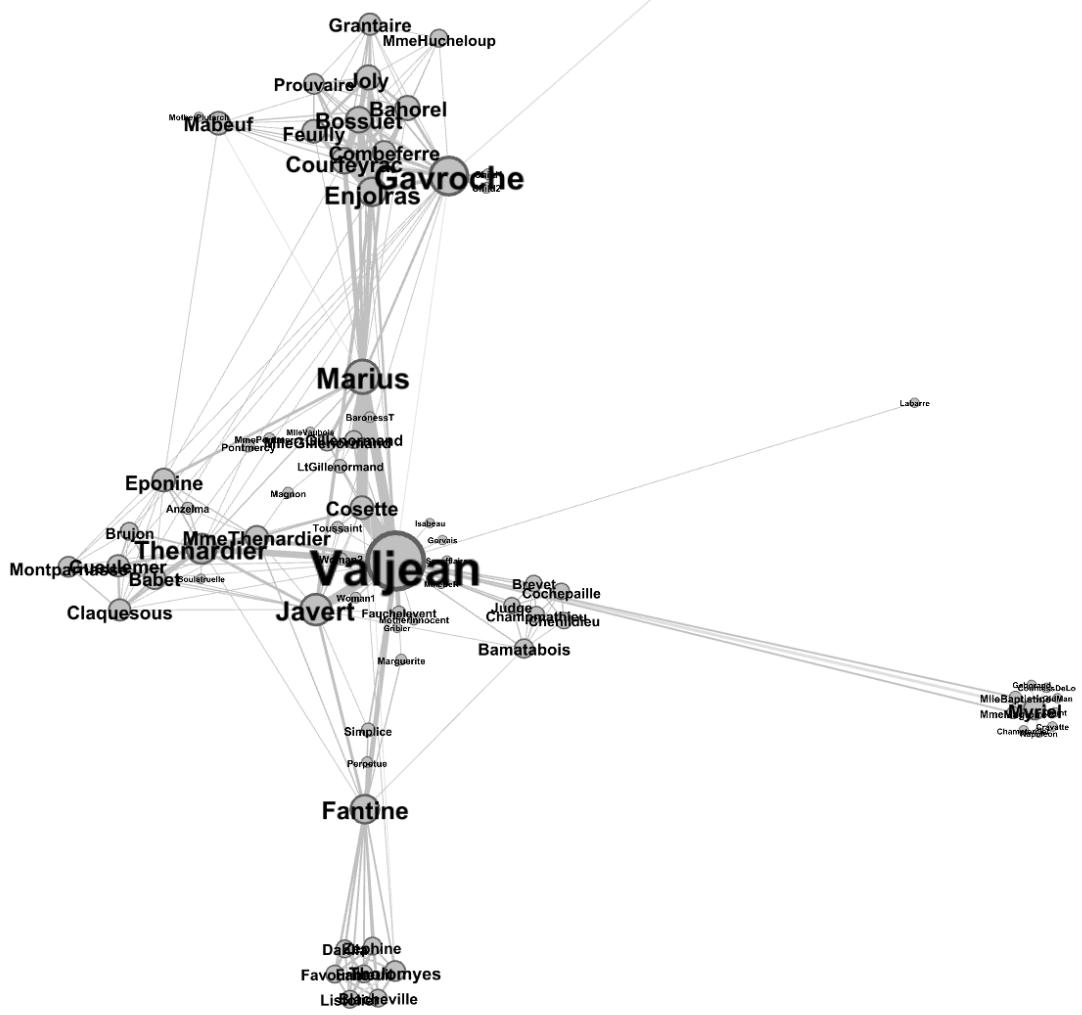
Afficher les labels



Afficher les labels en fonction de la taille du nœud:

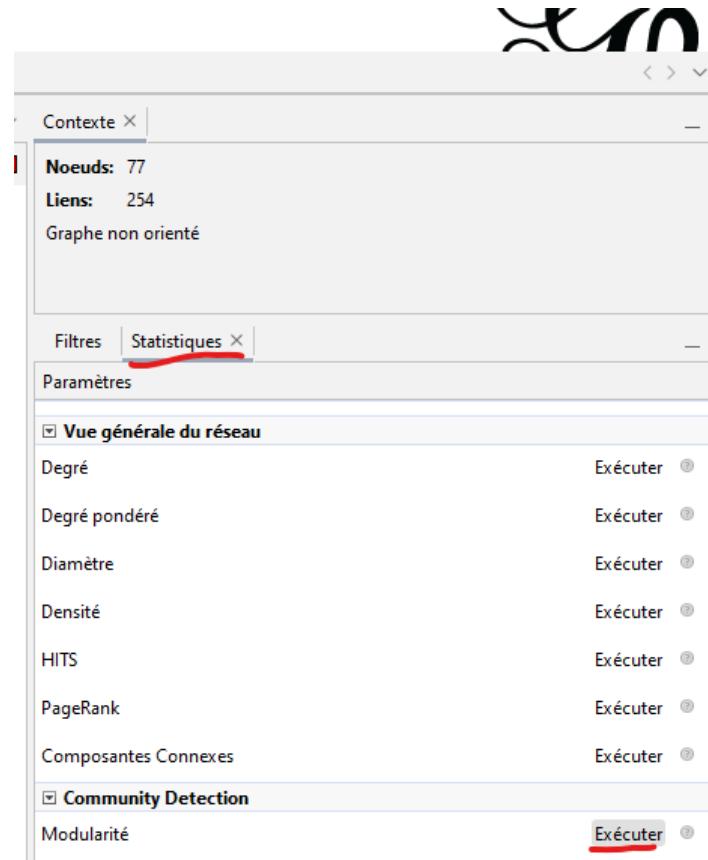


Après le changement de taille des nœuds et des labels le graphe devrait ressembler à cela:



Colorier les nœuds en fonction des classes de modularité

En fonction de la disposition du graphe et des liens entre les nœuds, Gephi peut calculer automatiquement des clusters.



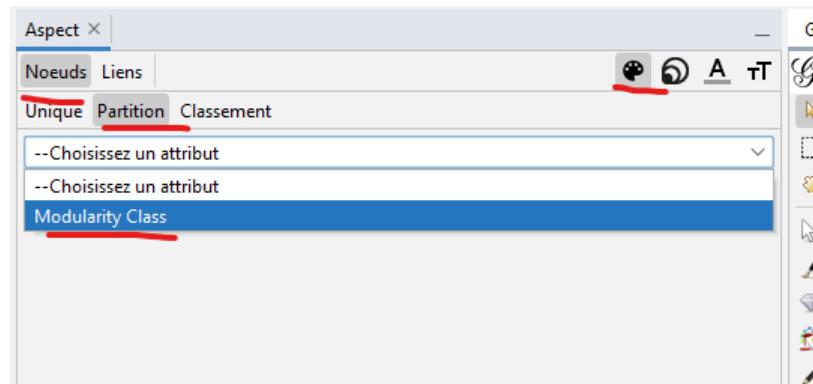
Ne pas changer les paramètres et appuyer sur OK

Dans la fenêtre obtenue:

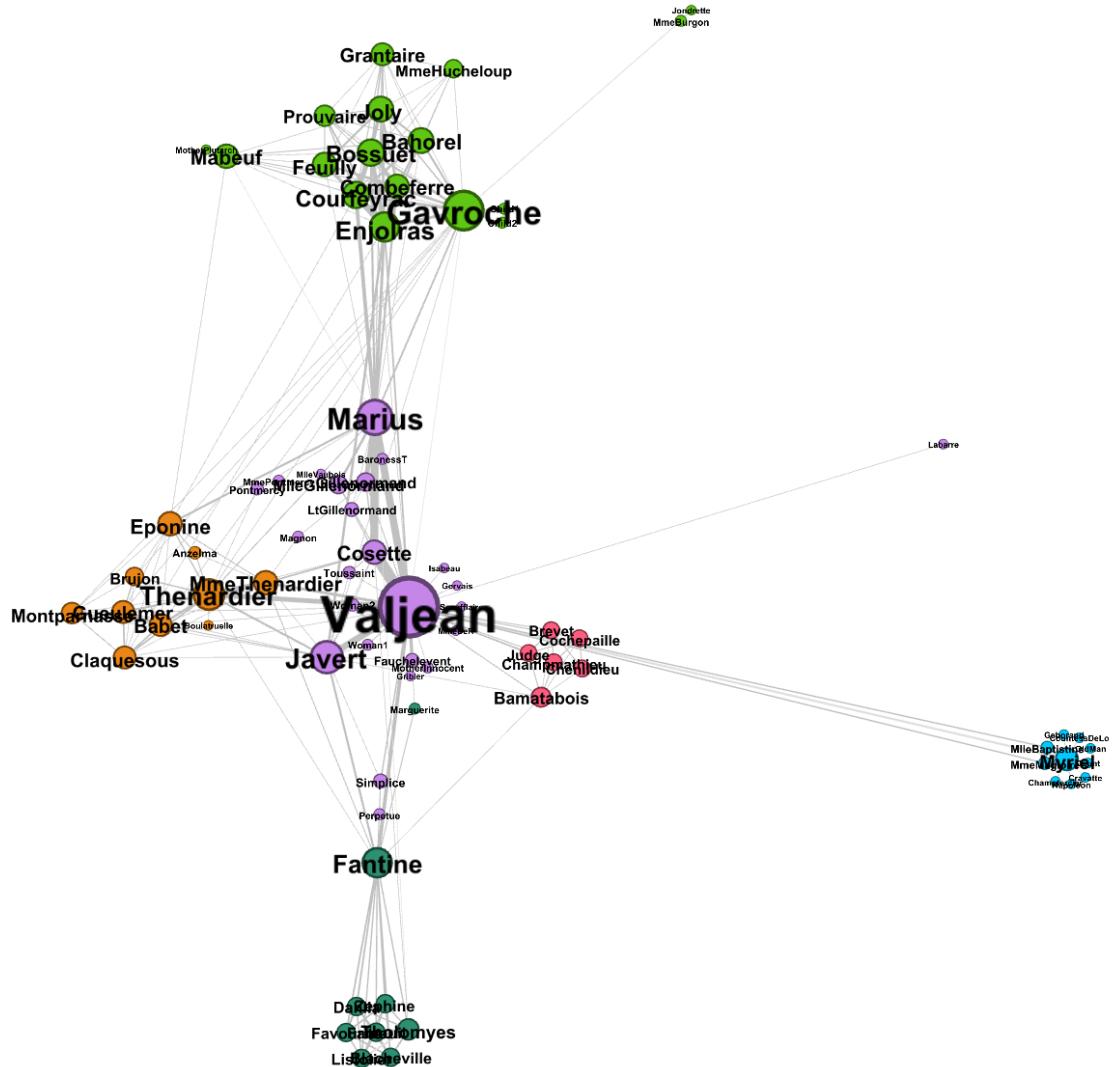
Results:
Modularity: 0,565

Ce score annonce la fiabilité des classes, plus le chiffre est proche de 1 plus les classes sont fiables.

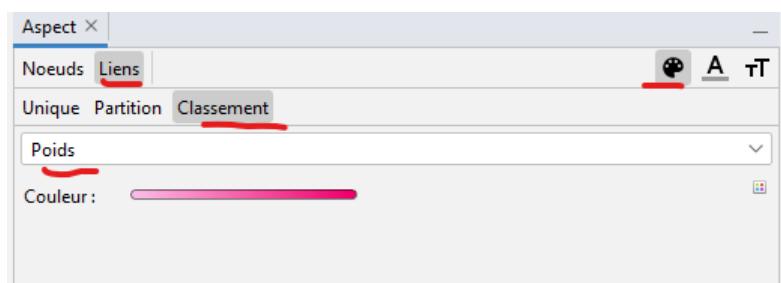
On peut ensuite utiliser ces dernières pour colorier les noeuds.



On appuie sur Appliquer et on obtient:

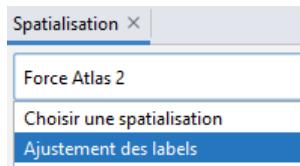


De la même façon que l'on a changé la couleur des nœuds, on peut également changer la couleur des liens, en fonction de leur poids par exemple (le poids est déjà représenté par l'épaisseur des liens, mais rajouter de la couleur permet parfois d'y voir plus clair).



Améliorer l'affichage

Enfin, afin de rendre le graphe plus lisible, on peut également utiliser une spatialisation permettant de réaligner les noeuds pour que les labels ne se chevauchent pas:

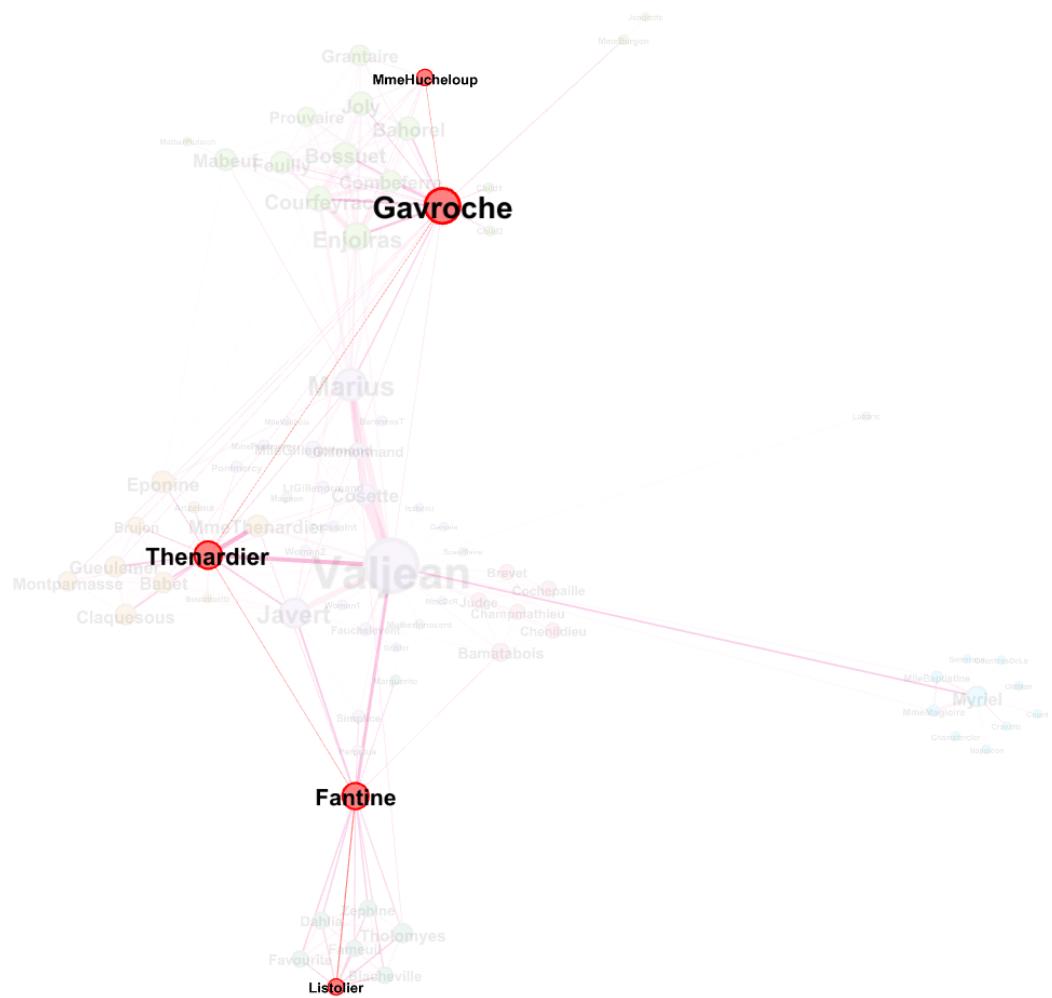


Si cela s'avérait ne pas être suffisant, on peut également utiliser la spatialisation “Déchevauchement”.

Voilà pour la partie d'amélioration visuelle !

On peut également décider de peindre à la main certains nœuds pour les mettre en valeur, peindre leurs voisins, colorer tout le graphe en fonction de la proximité avec un nœud racine ou encore afficher automatiquement le chemin le plus court entre deux points.

Ainsi si je souhaite afficher le chemin le plus court entre Listolier et Mme Hucheloup je peux sélectionner l'outil avec une icône d'avion dans la barre d'outils et sélectionner les deux nœuds qui m'intéressent. On obtient:



Cela s'avère particulièrement pratique en analyse de réseau.

Astuces de spatialisation

Ci dessous quelques astuces pour obtenir de spatialisation plus efficaces et / ou plus rapides:

Commencer par une spatialisation aléatoire

Dans les choix de spatialisation, il est conseillé avant tout ForceAtlas2 de passer par une spatialisation aléatoire afin de ne pas influencer l'algorithme. C'est d'autant plus important lorsqu'on a modifié le graphe (suppressions / ajouts de noeuds / de liens) après qu'une première spatialisation ait été faite. Il vaut mieux tout reprendre à zéro avec un aléatoire.

Ne pas afficher les liens pendant le calcul

En fonction du nombre de noeuds et de liens, ForceAtlas2 peut être très couteux. Il est conseillé de désactiver l'affichage des liens et des labels pendant le calcul:



Filtrer les données

Un autre moyen de rendre le calcul plus facile est de filtrer les noeuds. Si votre graphe contient un grand nombre de noeuds de très bas degré, ils ne seront **peut être** (j'insiste, il se peut qu'ils aient une importance) pas pertinent par rapport à d'autres car peu connectés.

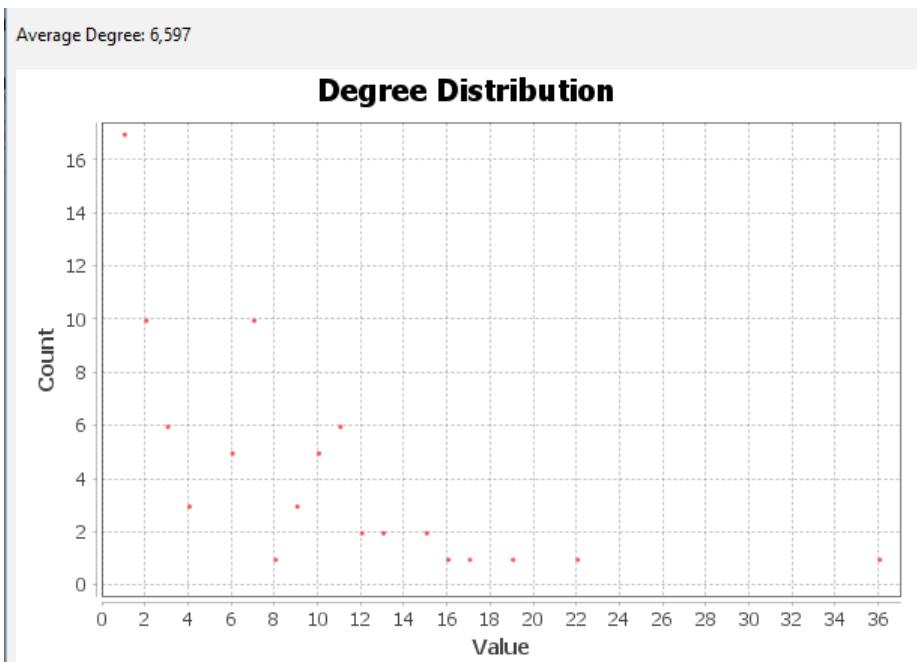
A l'inverse, certains noeuds trop gros font trop d'attraction et empêchent le graphe de se déployer de manière intéressante. (si tous les noeuds sont connectés au même, cela n'a pas grand intérêt).

Il peut donc être parfois pertinent de filtrer ces noeuds en fonction des situations.

Pour ce faire, il est conseillé d'abord de calculer le degré moyen des noeuds du graphe à l'aide de la statistique, "degré":

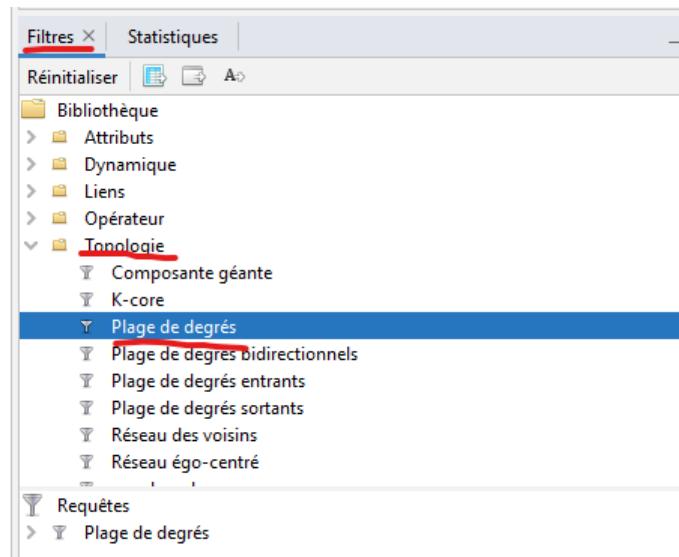


Dans le cas des misérables on voit que la moyenne est de 6.597, idéalement si l'on souhaitait filtrer il faudrait essayer de ne pas monter trop au dessus de 3 (~moyenne / 2).

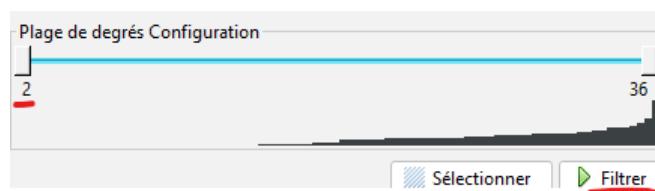


Et encore, on voit qu'un grand nombre de noeuds sont de degrés 1 et 2 on pourrait donc simplement filtrer les noeuds de degré 1.

Pour ce faire on se rend dans l'onglet Filtres > Topologie > Plage de degrés.



Puis sélectionner en bas à droite les plages, puis cliquer sur Filtrer.



On peut également décider de filtrer **après**. C'est à dire utiliser les noeuds de degré 1 pour spatialiser si l'on considère qu'ils doivent avoir une influence sur la répartition du graphe, mais les filtrer post-spatialisation pour ne pas qu'il soit trop chargé.

Figer certains nœuds

Lorsque certains noeuds sont de gros attracteurs et rendent le graphe peu lisible, il est également possible de les déplacer à la main loin les uns des autres, puis de les figer (en faisant clic droit puis Fixer) et de relancer la spatialisation. Cela permet de créer un graphe semi-constraint plus clair.

Quelques statistiques:

Gephi permet de calculer des indicateurs sur le graphe, il en existe beaucoup, plus ou moins complexes (nous avons déjà vu la Modularité) mais on peut également calculer:

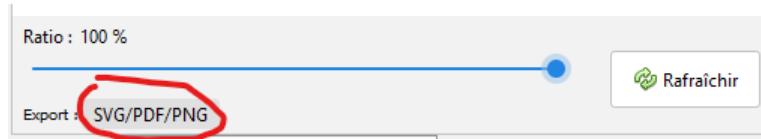
- Le degré moyen des nœuds, et la répartition des degrés
- Le diamètre du graphe, c'est à dire son plus long chemin d'un point à un autre
- Les plus courts chemins moyens, c'est à dire la distance moyenne entre deux points du graphe

Exporter son graph

Une fois tout cela terminé, la dernière étape consiste à exporter son graphe pour le partager et éventuellement le légendier et l'annoter. On utilisera pour cela le troisième espace de Gephi, "Prévisualisation".

Export PDF

Il existe un certain nombre de réglages possibles pour personnaliser l'export, mais les pré-réglages proposés sont déjà relativement efficaces. Vous pouvez les sélectionner un par un et appuyer sur Rafraîchir pour visualiser l'export qui sera réalisé. Il vous suffit ensuite de cliquer sur:



Et choisir le format. Je conseille personnellement SVG puisqu'il permet de zoomer à l'infini sans perte de qualité, et est exportable facilement dans des logiciels de retouche d'image pour annoter le graphe produit.

Export SigmaJS

Il est également possible d'exporter directement le graphe sous forme d'une page html interactive avec option de recherche et de clic sur les liens à l'aide de l'export SigmaJS.

Il faut pour cela installer le plugin dédié dans Outils > Modules Complémentaires > Disponibles et de chercher sigma, d'installer puis de redémarrer Gephi et enfin de sélectionner Fichier > Export > Sigma.

Ressources utiles:

L'atelier cartographie de Franck Ghittalla qui recense plusieurs années de recherche sur les graphes de réseaux et leurs applications <https://ateliercartographie.wordpress.com/>

Son livre associé et disponible gratuitement en ligne et se focalisant particulièrement sur les graphes de réseaux sur le web: <https://books.openedition.org/oep/pdf/15358>

Infranodus

work in progress