

UNIVERSITÉ DE TECHNOLOGIE DE COMPIÈGNE

COMPUTER SCIENCE

DATA MINING

IC05

Étude des données du Pic'Asso

Author:

Alexandre CORTYL
alexcortyl@gmail.com

Author:

Hugo RODDE
rodde.hugo@gmail.com

January 6, 2016



Contents

1	Introduction et contexte	2
1.1	Le foyer Pic'Asso	2
1.2	Données disponibles	2
1.3	Objectifs	3
2	Méthode de travail	4
2.1	Données fournies	4
2.1.1	Extrait du fichier initial	5
2.2	Programme de conversion	5
2.2.1	Modules	5
2.2.2	Classes	6
2.2.3	Base de données	7
3	Étude et résultats	8
3.1	Résultats initiaux	8
3.1.1	Top 10 des plus gros consommateurs en terme de budget	8
3.1.2	Top 10 des plus grosses transactions en terme de prix	9
3.1.3	Top 10 des produits les plus achetés	9
3.1.4	Top 10 des produits générant le plus de revenu	10
3.1.5	Top 10 des plus grosses permanences en terme de fréquentation	10
3.1.6	Top 10 des plus grosses permanences en terme de revenu	11
3.2	Visualisation avec Gephi	12
3.2.1	Brasseurs	12
3.2.2	Consommateurs	19
4	Conclusion	31

1 Introduction et contexte

Dans le cadre de l'UV IC05 "Cartographie des données", nous avons choisi de travailler sur les données relatives au foyer étudiant de notre école : le Pic'Asso.

1.1 Le foyer Pic'Asso

Le Pic'Asso est le foyer étudiant de l'UTC. C'est un espace de rencontre ouvert de 10 heures à 22 heures aux étudiants et au personnel de l'UTC. De nombreux services y sont proposés, avec notamment :

- Vente de boissons fraîches (non alcoolisées) et de boissons chaudes
- Vente de nourriture (viennoiseries le matin, repas préparés par des associations le midi, assiettes le soir, snacks toute la journée, etc.)
- Exploitation de la licence II le soir de 18h30 à 21h30
- Divertissements en tout genre mis à disposition des étudiants (jeux de cartes, jeux de sociétés, billard, baby-foot, jeux-vidéos, boules de pétanque, etc.)
- Une ambiance conviviale réservée à la détente, à la créativité et à la prise d'initiatives

Le foyer étudiant, centre de la vie associative, est un lieu de communication et de recrutement par excellence pour les associations. C'est aussi un endroit d'expression et de réflexion sur l'organisation de la vie « du campus », point de rendez-vous où ont lieu des échanges culturels.

1.2 Données disponibles

Nous avons donc collecté l'ensemble des transactions (ou achats) réalisées tout au long de l'année universitaire 2015 (du 1er septembre 2014 au 30 juin 2015).

Pour respecter la confidentialité des clients du Pic'Asso, ces données ont été anonymisée : dans un premier temps, il est impossible d'associer un nom aux transactions, les acheteurs étant seulement identifiés par un ID.

A cela, nous avons rajouté des données météo obtenues à l'aide de l'API de World Weather Online <https://developer.worldweatheronline.com/> et des informations sur les permanences tenues au foyer sur cette même période.

L'étude et la représentation des données sous forme de graphe est réalisée à l'aide du logiciel gratuit et open-source Gephi : <https://gephi.org/>

1.3 Objectifs

Les objectifs de notre analyse sont :

- Mieux comprendre la consommation au sein de notre foyer
- Identifier des clusters de clients et de produits
- Remarquer des corrélations entre consommations, permanences et conditions météorologique

Le principal bénéficiaire de cette étude est bien entendu l'association en charge de la gestion du foyer, mais chaque étudiant pourrait aussi apprécier de mieux connaître et comprendre ses goûts et ses habitudes de consommation.

2 Méthode de travail

Dans un premier temps, nous présentons comment nous avons traité les données fournies par le Pic'asso, les outils et la méthode utilisée.

En préambule, voici le planning de travail avec l'avancement des différentes tâches réalisées au cours du semestre :

Description	Sept	Oct	Nov	Dec	Jan
Récupérer les données brutes					
Se familiariser avec les données fournies					
Selectionner et nettoyer les données fournies					
Ecrire le programme pour digérer les données dans la BDD					
Se familiariser avec Gephi					
Premiers graphiques avec Gephi					
Ajout des données relatives aux brasseurs et liens entre les bières					
Seconds graphiques avec les données des brasseurs					
Réalisation des visualisations avec Gephi					
Redaction du rapport et du site					

2.1 Données fournies

Le Pic'asso nous a fourni un fichier CSV contenant 253 898 entrées correspondant aux transactions.

Une transaction est définie ainsi : l'achat par un même consommateur d'un produit et d'un seul avec une quantité supérieure ou égale à 1.

Ces données brutes ne sont pas très pratiques en l'état. On distingue mal les noeuds et les liens. Nous avons donc choisi de transformer ces données en fichier SQL. Ce dernier peut alimenter une base de données relationnelle de type MySQL. Il est de même avec les données fournies par l'API de météo.

A terme, nous disposons d'une base de données indexées dans laquelle nous pouvons effectuer des requêtes complexes grâce au langage SQL.

Cette base de donnée présente également l'avantage d'être facilement exportable, modifiable, dupliquable...

Finalement, elle est nettement plus exploitable qu'un seul et unique fichier text de type CSV avec plus de 250 000 lignes et pesant près de 20 Mo.

En effet, SQL permet de créer des Vues. Celles-ci sont en fait des requêtes dont le résultat est sauvegarder dans une table et mis à jour en temps réel. Ainsi, Gephi proposant

d'importer les données depuis deux tables SQL nodes et edges, nous pouvons déterminer autant de Vues que nécessaire pour travailler sur des sous-ensembles du jeu de données complet.

On gagne en modularité et clarté.

2.1.1 Extrait du fichier initial

Pour information, on présente ci-dessous un exemple des données fournies par le Pic'Asso :

Transaction ID	Date	Time	Buyer ID	Product ID	Product Name	Category ID	Category Name	Unit. Price	Quantity	Total Price
578782	30/06/2015	11:29:07	5498	22	Ice Tea Peche	3	Softs	0.7	1.0	0.7
...
274656	03/09/2014	18:29:41	5498	457	Cuvee des Trolls	11	Bieres pression	1.69	1.0	1.69

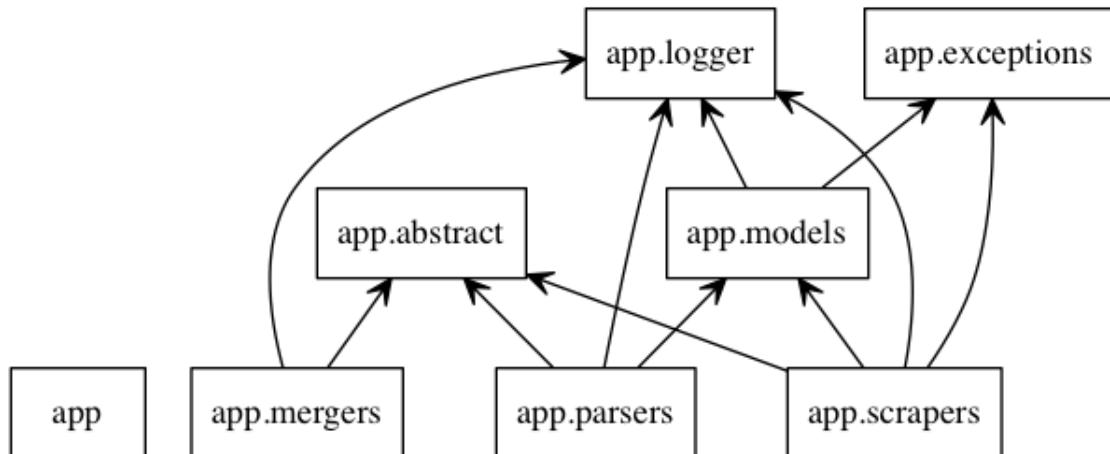
2.2 Programme de conversion

Nous avons donc développé un programme en Python 2.7 afin de lire le fichier CSV et générer les requêtes SQL correspondantes.

Ce programme est très modulaire, il utilise les bonnes pratiques de la programmation orientée objet afin de pouvoir réutiliser au maximum les procédures écrites.

Il est disponible à l'adresse suivante : <https://gitlab.utc.fr/roddehug/ic05-picasso>

2.2.1 Modules



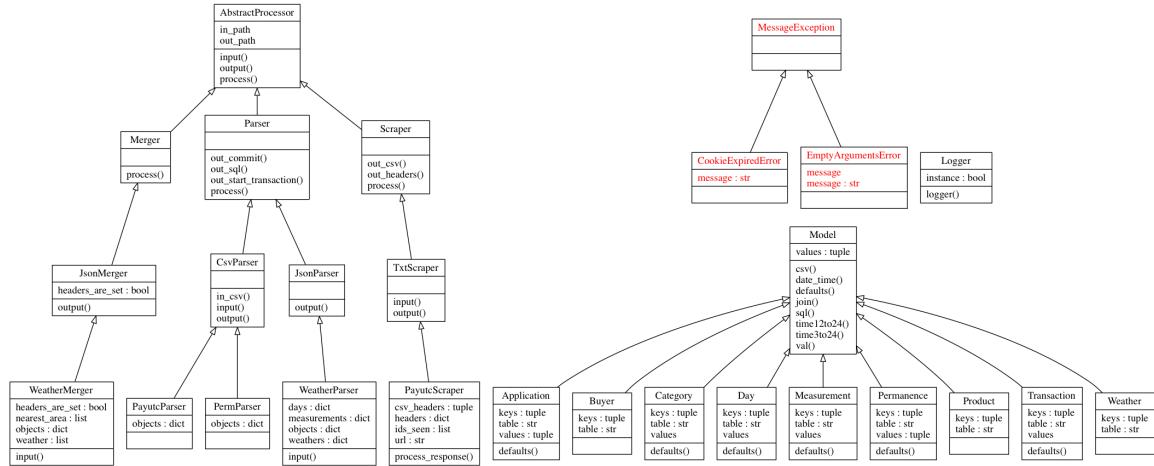
Le module principal est le module app.

Nous avons découpé le programme en fonctionnalités ou modules:

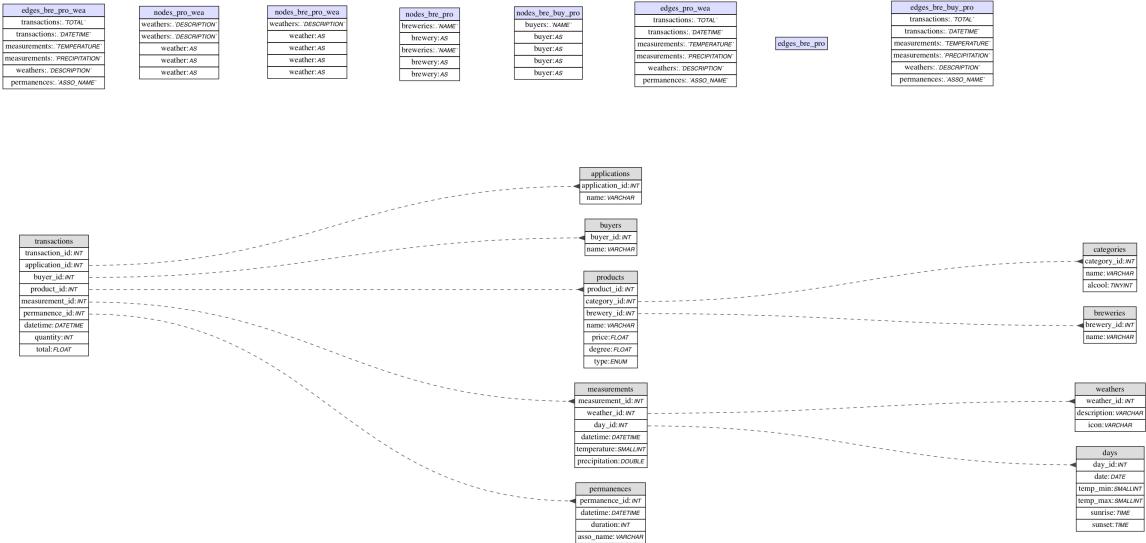
- **exceptions** qui rajoute des nouvelles exceptions particulières à nos processus de transformation.

- **logger** qui permet de paramétriser le module de logging afin de pouvoir debugger l'application.
- **models** qui contient tous les modèles de notre base de données afin de générer les requêtes SQL.
- **abstract** qui permet d'ajouter un niveau d'abstraction afin de factoriser le code.
- **mergers** qui permet de fusionner plusieurs fichiers en un seul à la sortie (pour les données météo contenues dans plusieurs JSON).
- **parsers** qui permet de parcourir un fichier et de créer les objets correspondants en langage SQL (pour les données CSV).
- **scrapers** qui permet d'effectuer des requêtes HTTP vers un site afin de récupérer du contenu et l'enregistre dans un fichier CSV.

2.2.2 Classes



2.2.3 Base de données



3 Étude et résultats

Dans cette partie, nous ferons figurer les visualisations les plus parlantes et les représentations les plus intéressantes de notre étude.

3.1 Résultats initiaux

Ci-dessous, nous répertorions quelques résultats statistiques permettant de mieux comprendre les données sur lesquels nous travaillons.

3.1.1 Top 10 des plus gros consommateurs en terme de budget

buyer_id	quantity	sum	from	to
6541	2381	2135.45	2014-09-05 19:39:14	2015-06-28 19:26:25
7332	1562	1329.87	2014-09-08 18:54:07	2015-06-25 20:38:51
4816	1524	1301.14	2014-09-03 18:32:45	2015-06-19 19:47:18
4838	963	1263.58	2014-09-08 18:47:58	2015-06-27 18:33:28
4923	1197	1138.63	2014-09-08 10:00:34	2015-06-24 21:20:29
5908	978	1033.53	2014-09-09 16:48:33	2015-06-25 10:40:34
5463	840	1033.20	2014-09-09 14:27:59	2015-06-25 21:06:05
5498	843	1003.45	2014-09-03 18:29:41	2015-06-30 11:29:07
1147	2212	942.33	2014-10-07 19:48:19	2015-06-28 19:26:07
4932	859	864.55	2014-09-08 10:47:23	2015-06-25 19:27:30

3.1.2 Top 10 des plus grosses transactions en terme de prix

transaction_id	buyer_id	name	asso_name	datetime	quantity	total
386417	168	Menu Perm Plep		2014-12-18 18:28:05	40	60
431233	5041	Kinder Maxi	La Finepic	2015-03-12 13:14:40	32	3.2
407535	5103	Barbar Blonde	Picasso	2015-02-17 21:11:53	24	46.8
454541	5103	Grand Cru St Feuillien	B3M	2015-03-26 21:33:39	24	46.8
456406	7283	Rochefort 8	P15 du Turfu	2015-03-27 20:24:39	24	46.8
465356	5432	Chimay Bleue	semaine des poles	2015-04-02 21:18:55	24	43.2
467353	7356	Rochefort 8	semaine des poles	2015-04-03 21:19:38	24	46.8
317759	6442	Duvel	Ski UTC	2014-10-07 20:20:58	24	39.6
494753	6008	Duvel		2015-04-25 19:40:49	24	39.6
524440	4931	Duvel	Charcut	2015-05-20 21:29:44	24	39.6

3.1.3 Top 10 des produits les plus achetés

product_id	name	quantity
458	Delirium Tremens	17073
457	Cuvee des Trolls	15733
466	Tripel Karmeliet	13333
85	Cafe	12983
1395	Peche Mel Bush	11016
1401	Gauloise Rouge	9705
1397	Duvel	7538
3565	Jo Colina	7337
1394	Carolus Triple	7129
1403	Barbar Blonde	6531

3.1.4 Top 10 des produits générant le plus de revenu

product_id	name	revenue
458	Delirium Tremens	30788.73
457	Cuvee des Trolls	26616.97
466	Tripel Karmeliet	24024.30
1395	Pech Mel Bush	18735.73
1401	Gauloise Rouge	17984.45
1394	Carolus Triple	12473.67
1397	Duvel	12428.13
3796	Biere	12266.40
1399	Chimay Bleue	10938.63
1403	Barbar Blonde	10457.24

3.1.5 Top 10 des plus grosses permanences en terme de fréquentation

permanence	date	nb_client
ESTU PARKING	2015-05-13	1487
Picasso	2015-02-16	1018
Vieux CompiBitches	2014-11-10	950
Rugby UTC	2015-02-26	841
Picemons	2014-12-22	811
Imaginarium Festival	2015-04-21	804
Les Schtroumpfs	2015-03-13	798
Comedie Musicale	2015-03-20	795
Picasso	2015-02-19	791
Picasso	2015-02-18	782

3.1.6 Top 10 des plus grosses permanences en terme de revenu

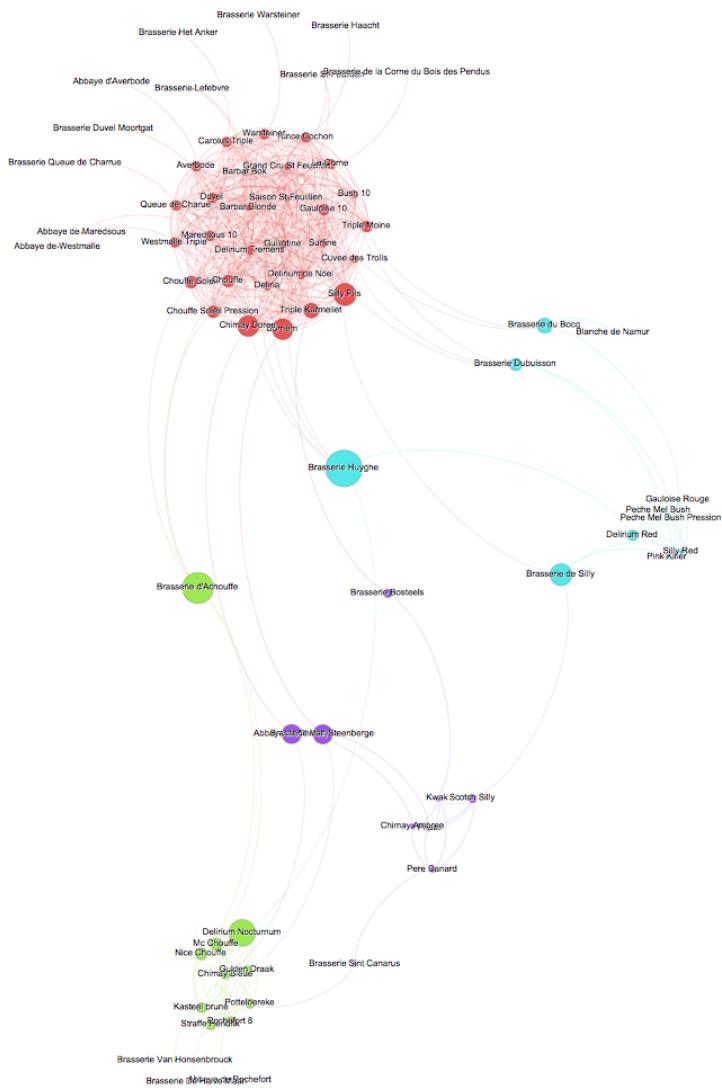
permanence	date	revenue
ESTU PARKING	2015-05-13	13930.35
Vieux CompiBitches	2014-11-10	5468.38
Picasso	2015-02-16	5192.17
	2014-09-08	4820.93
Picemons	2014-12-22	4119.22
Rugby UTC	2015-02-26	3987.62
Picasso	2015-02-20	3976.90
P15 du Turfu	2015-03-27	3968.16
	2015-01-15	3898.39
SDF	2015-05-22	3828.47

3.2 Visualisation avec Gephi

Ci-dessous, nous faisons figurer les cartographies les plus intéressantes, obtenues sur Gephi.

3.2.1 Brasseurs

Le premier type de visualisation concerne les données relatives aux brasseurs.



Paramètres de visualisation :

- **Disposition** : Force Atlas 2
- **Colorisation** : Modularity Classes
- **Taille des noeuds** : Betweenness Centrality

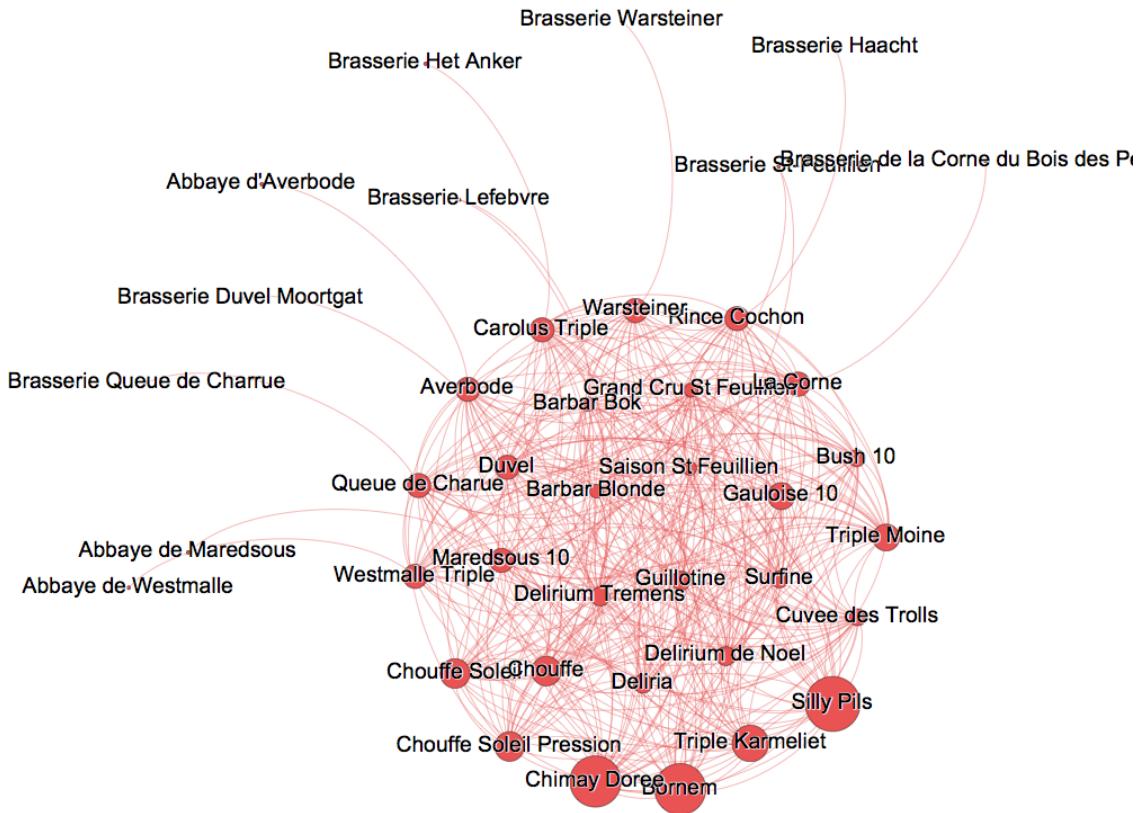
Méta-données obtenues:

- **Network Diameter** : 7
- **Radius** : 4
- **Average distance** : 2,75
- **Modularity** : 0,266
- **Number of classes** : 4

Nous sommes alors en mesure de distinguer plusieurs catégories de bières, avec les brasseries associées. Nous détaillons par la suite la décomposition correspondante, qui permet d'identifier les différents types de bières vendus au Pic'Asso.

Il est tout particulièrement satisfaisant de constater que la spatialisation Force Atlas 2 a été en mesure d'organiser les bières (et brasseries associées) en 4 catégories logiques en terme de type de bière. C'est en effet les 4 catégories principales de bière en vente au Pic'Asso !

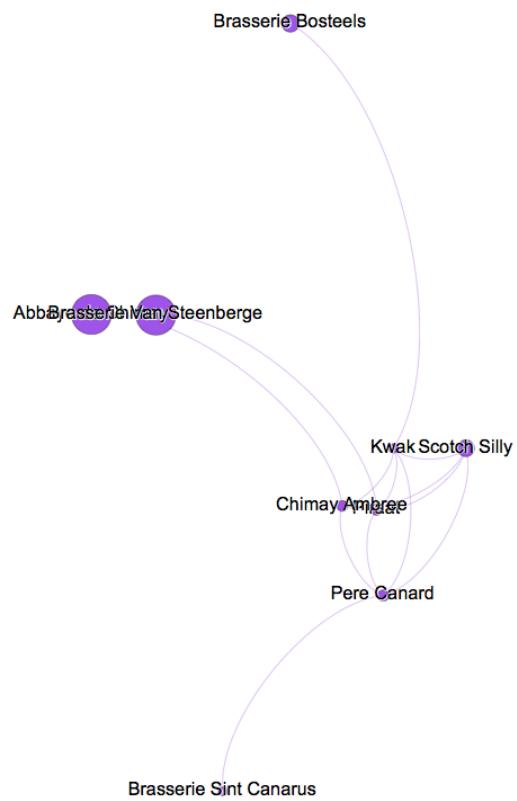
Bières blondes



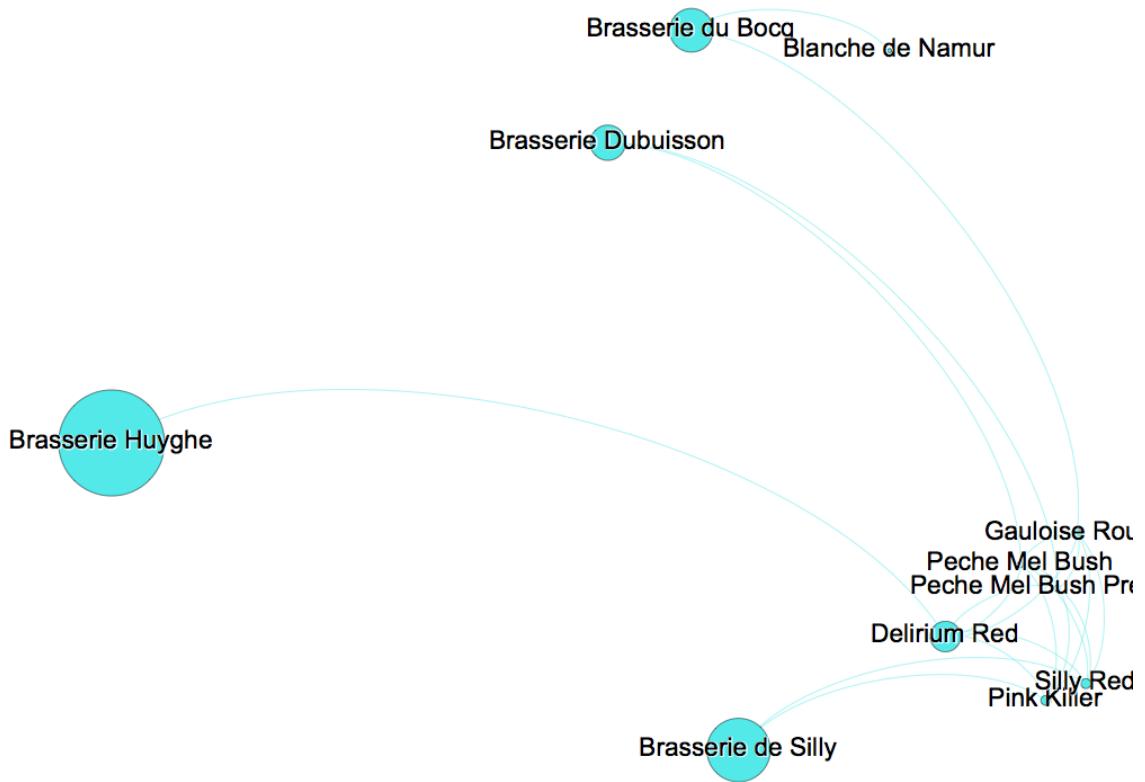
Bières brunes



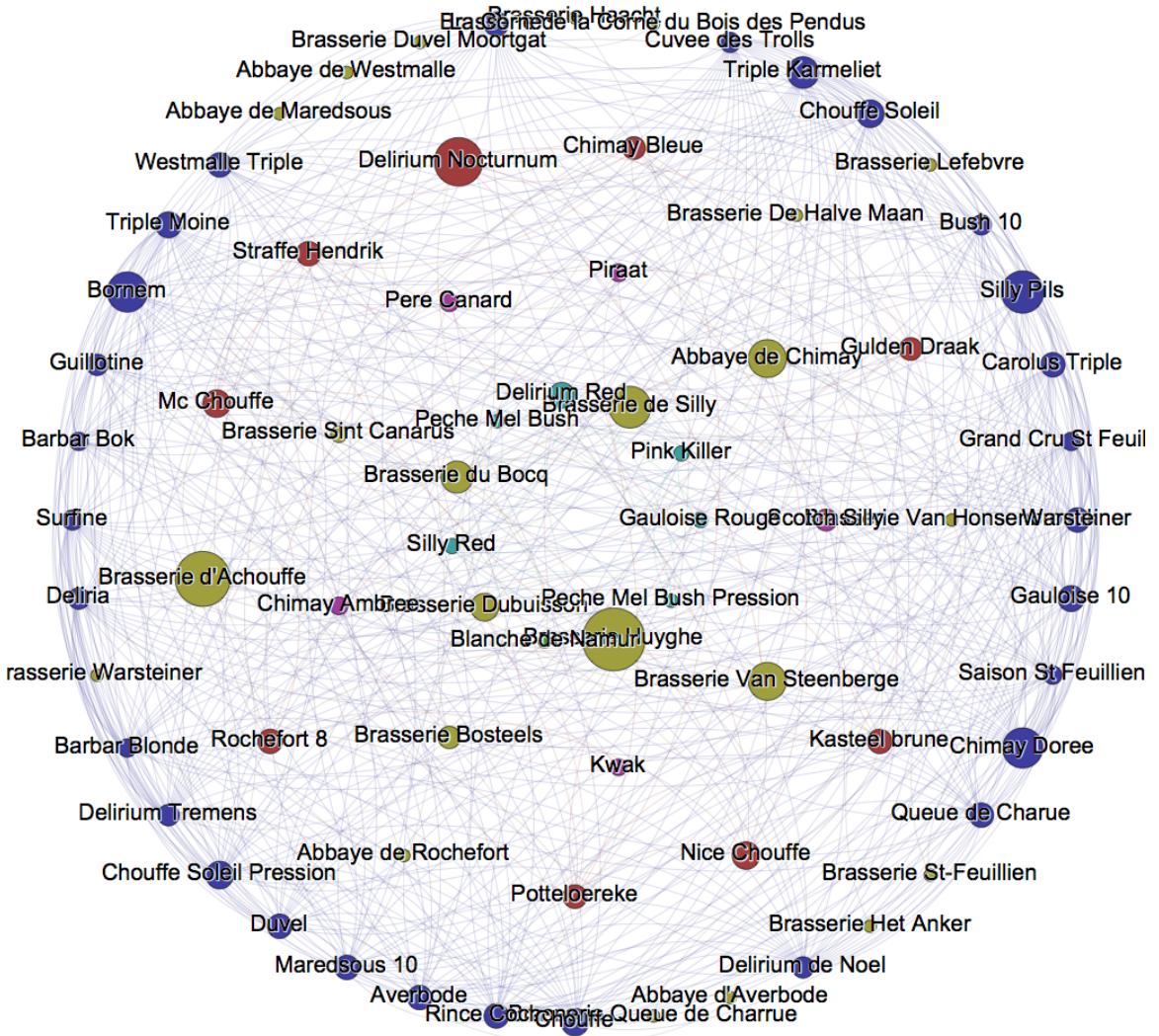
Bières ambrées



Bières fruitées



Une autre visualisation, utilisant cette fois-ci la spatialisation "Layered" est ajoutée ci-dessous :



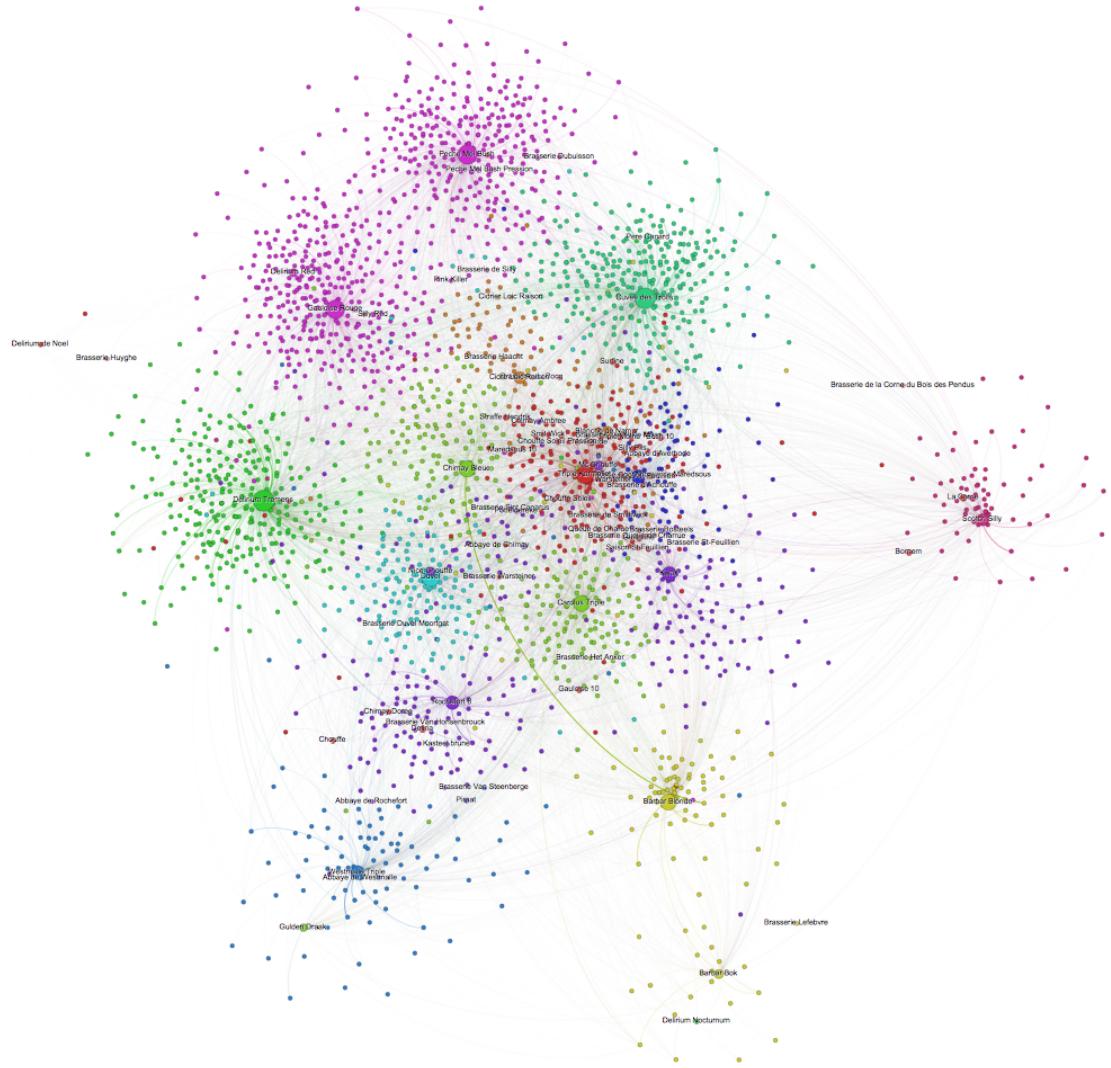
Paramètres de visualisation :

- **Disposition :** Layered Layout with Modularity
- **Colorisation :** Beer type
- **Taille des noeuds :** Betweeness Centrality
- **Couches :** Modularity Class

Nous sommes alors en mesure de constater que les bières de type ‘Blanches’ sont intégrées dans la classe ‘Bières fruitées’.

3.2.2 Consommateurs

Le second angle d'étude se concentre sur le point de vue consommateurs.



Paramètres de visualisation :

- *Disposition* : OpenOrd
 - *Colorisation* : Modularity Classes
 - *Taille des noeuds* : Degree

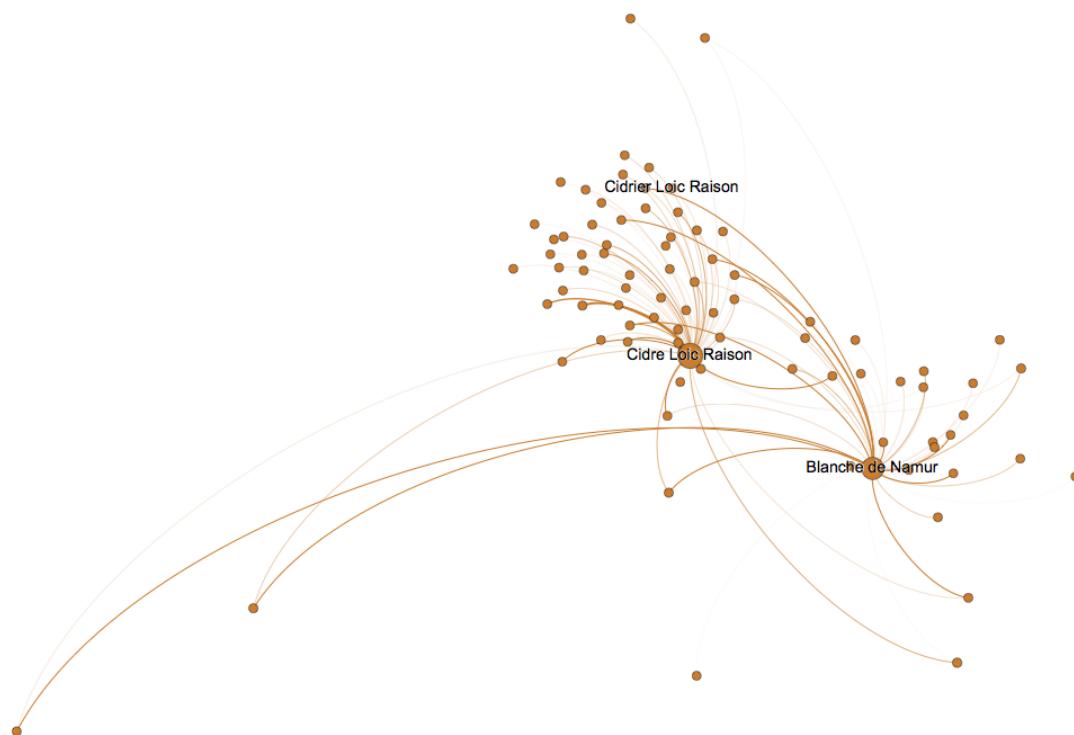
Méta-données obtenues:

- *Network Diameter* : 4
 - *Radius* : 3
 - *Average distance* : 2,18
 - *Modularity* : 0,396
 - *Number of classes* : 11

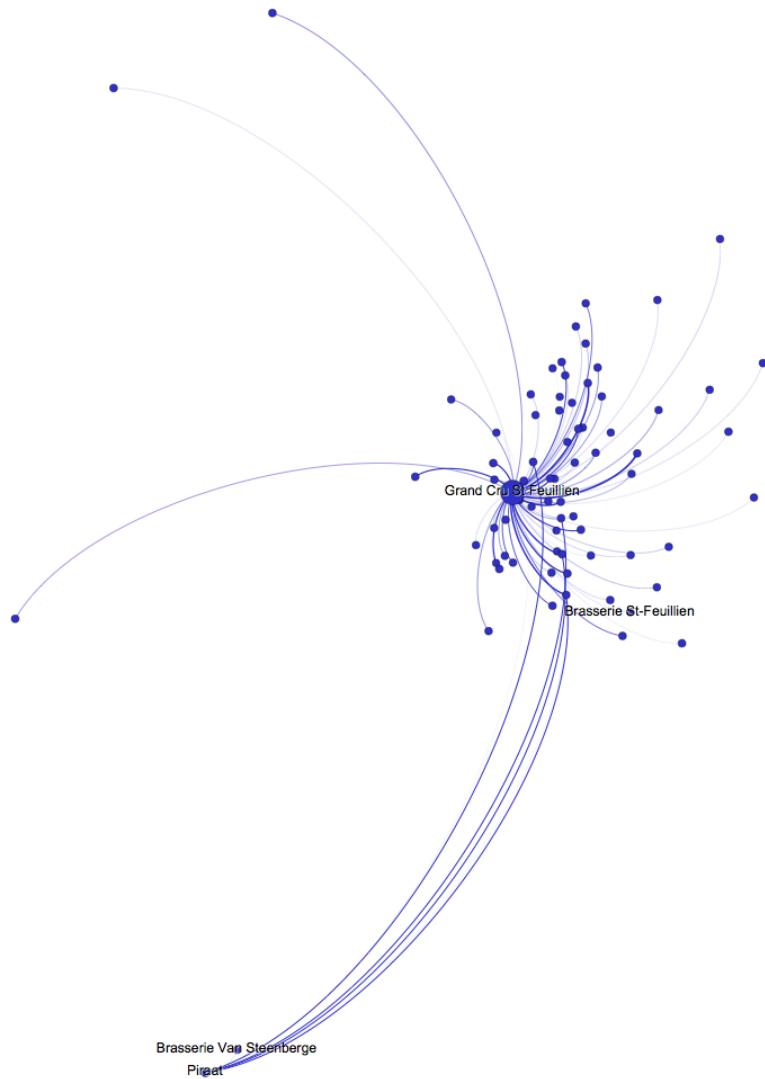
On est alors capable de distinguer plusieurs catégories de bières, avec les consommateurs associées. Cette clusterisation en 11 groupes permet d'identifier différents profils de consommateurs, suivant leurs goûts et leurs habitudes de consommations.

Ci-dessous la décomposition correspondante :

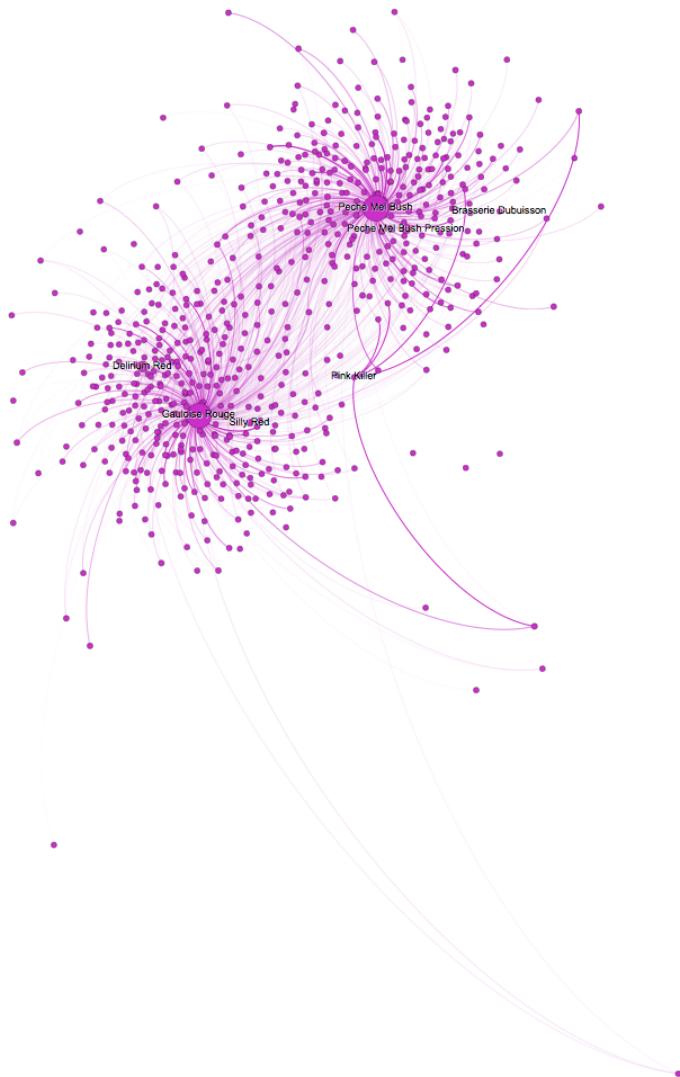
Cidres et bières blanches



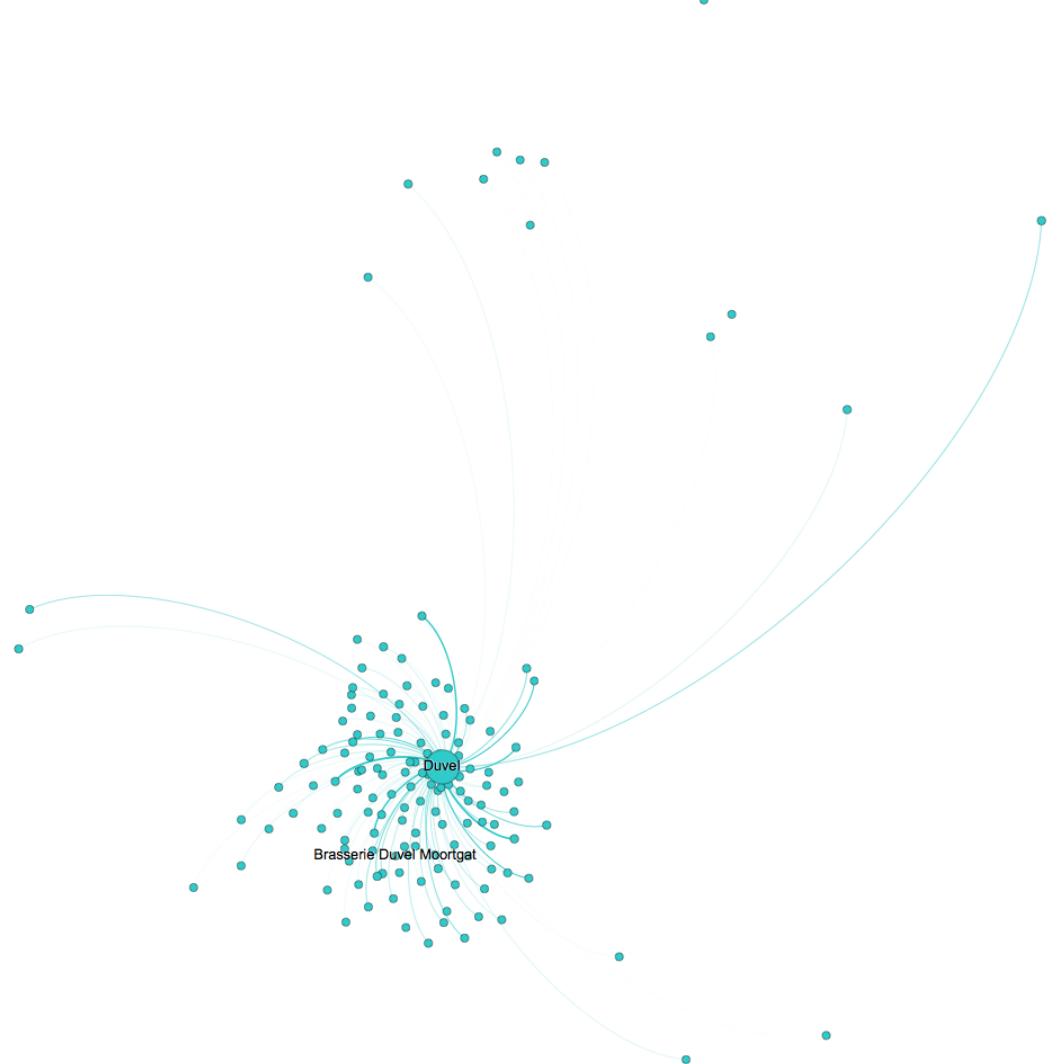
Bière St-Feuillien Grand Cru



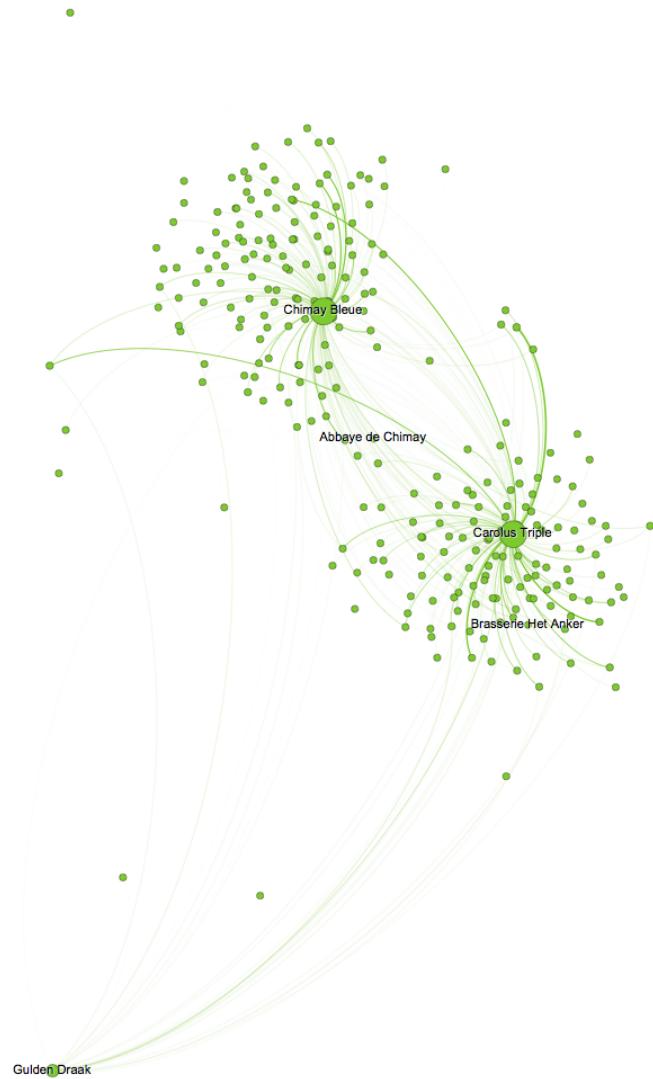
Bières fruitées



Bière Duvel



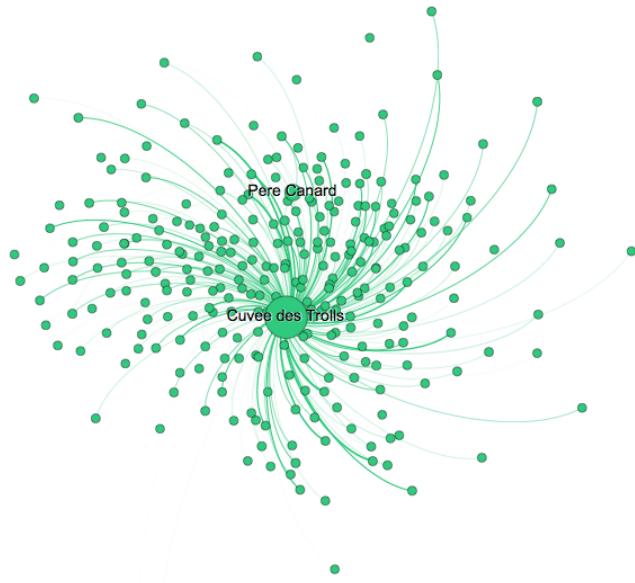
Bières blondes fortes



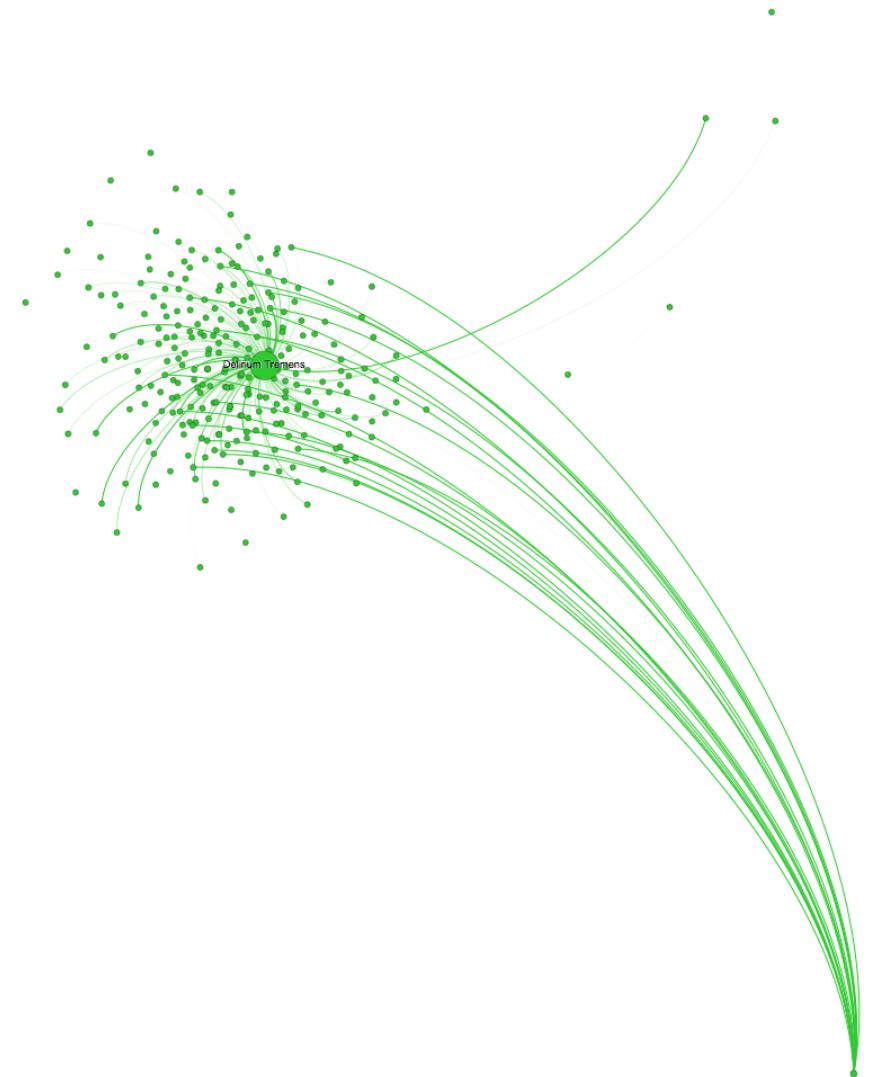
Bière Scotch Silly



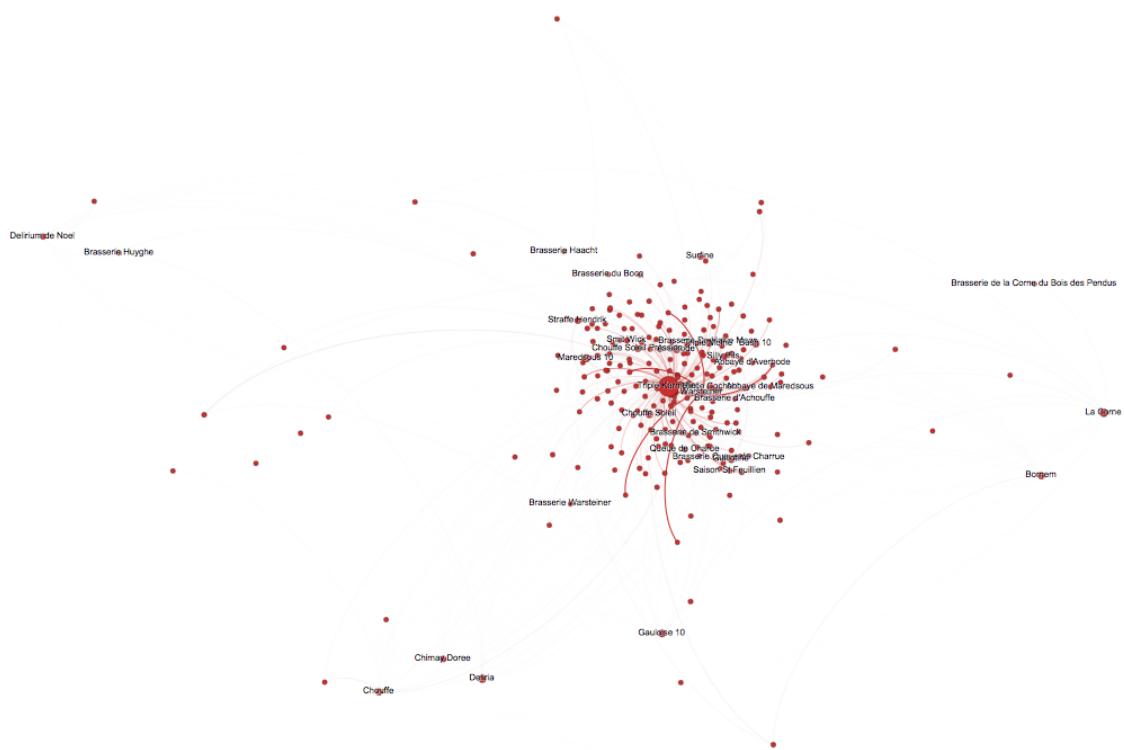
Bière Cuvée des Trolls



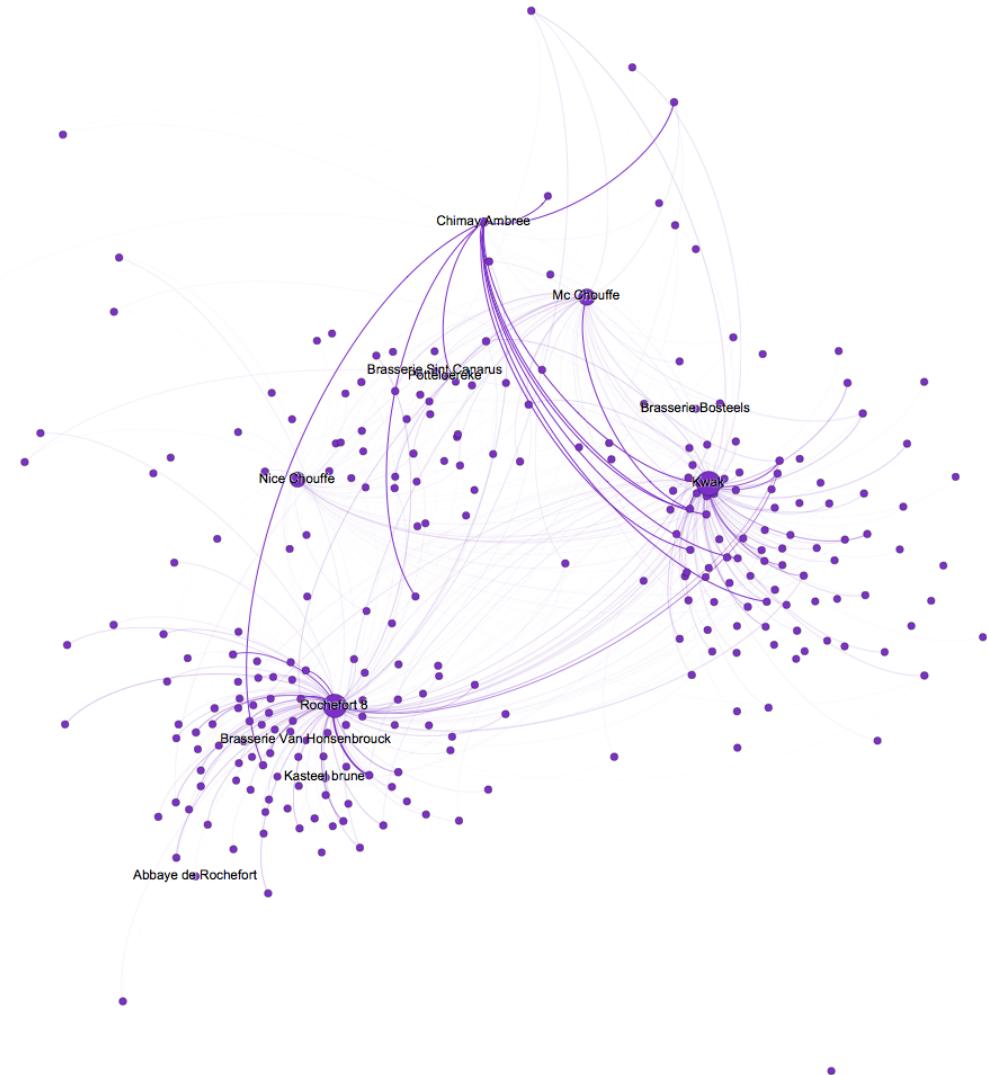
Bière Délirium Tremens



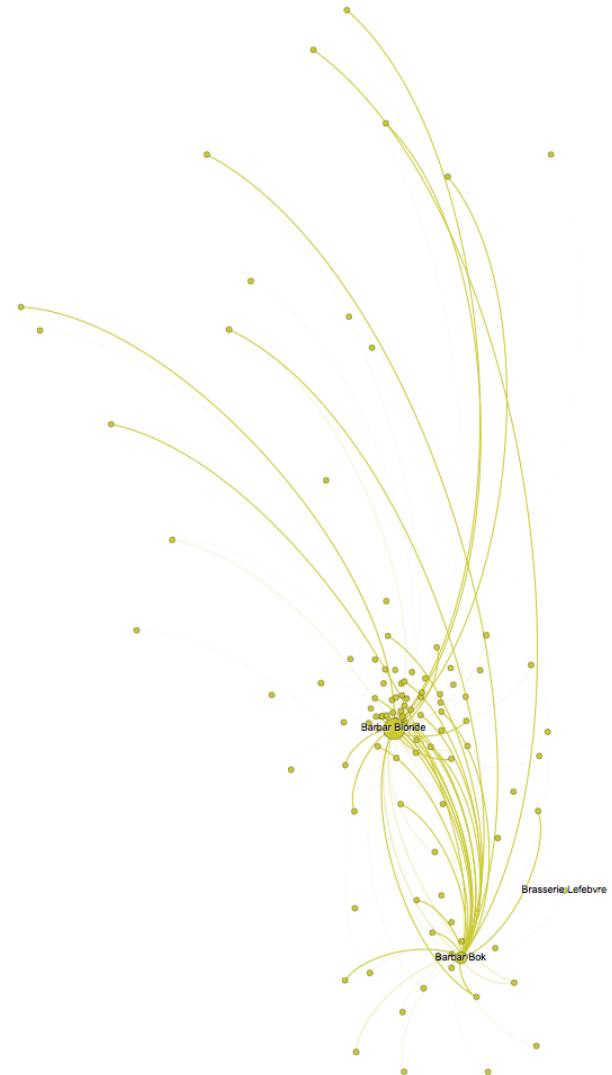
Bières blondes



Bières brunes



Bières Barbar



4 Conclusion

Ce projet nous aura permis de prendre en main l'outil Gephi, aujourd'hui incontournable dans le monde de la représentation des données et en cartographie.

Cependant ce dernier est encore en stade de développement et s'avère peu stable pour un usage en entreprise.

Nous avons apprécié les multiples plugins proposés par la communauté. Ces derniers nous ont permis d'exporter nos graphs en format web avec l'export SigmaJS.

Notre étude aura permis de catégoriser les différentes bières vendues au Pic'Asso, mais aura surtout rendu possible la mise en évidence de clusters de consommateurs suivant leurs habitudes de consommations.

Les diverses visualisations produites permettent à la fois de confirmer certaines informations connues, mais ont aussi rendu possible l'identification de nouvelles choses. En cela, le projet a atteint ses objectifs et est une réussite !

Mais encore beaucoup reste à faire, en commençant notamment par :

- Étendre l'étude des consommations aux produits autre que la bière
- Exploiter des données relatives à la météo
- Approfondir l'identification de cluster clients

Nous espérons avoir ouvert la voie qui permettra à d'autres étudiants de poursuivre une analyse plus poussée des données de consommation au Pic'Asso.

Si vous souhaitez obtenir de plus amples informations sur le travail réalisé, n'hésitez pas à prendre contact avec nous (emails disponibles en première page de ce rapport).

En particulier, notre travail est disponible sur Gitlab et peut être partagé sur demande.