

Royaume du Maroc
UNIVERSITÉ MOHAMED V - RABAT

ECOLE NATIONALE SUPÉRIEURE D'INFORMATIQUE
ET
D'ANALISE DES SYSTÈMES



Rapport de projet du Analyse des réseaux sociaux- SNA

Réduction de dimensionnalité sur EuroRoad

Filière : Génie de la Data (GD)

Réalisé par :

AIT YOUB Abdelmoughit
CHERKAOUI Nabil

Encadrant :

Mr. JIBOUNI Ayoub

Année universitaire : 2023 - 2024

Introduction

En raison de ses infrastructures routières complexes et interconnectées, l'Europe est un continent en constante évolution qui reflète sa diversité culturelle, économique et géographique unique. Bien que la cartographie des routes européennes soit très riche en informations, elle présente un défi important car elle nécessite des solutions innovantes pour extraire des connaissances importantes à partir de données massives et complexes. C'est pour cela qu'il devient important de considérer l'approche visant à réduire la dimensionnalité originale des graphes des réseaux routiers européens afin de faciliter leurs analyses, visualisations et pour une interprétation plus efficace des données.

Ce rapport traite du problème de la représentation des réseaux routiers européens en dimension réduite en utilisant trois techniques de réduction de dimensionnalité : l'Analyse en composantes principales (PCA en anglais), l'embedding localement linéaire (LLE) et l'Uniform Manifold Approximation and Projection (UMAP). Ces techniques offrent diverses méthodes pour réduire la complexité inhérente à ces données tout en maintenant les structures et les relations essentielles.

Notre recherche vise à obtenir une compréhension approfondie des méthodes employées pour réduire la dimensionnalité des réseaux routiers européens. Nous chercherons également à déterminer quelle méthode offre les meilleurs résultats en analysant les résultats obtenus avec PCA, LLE et UMAP. Nous examinerons les principes fondamentaux et les avantages spécifiques de chaque technique de réduction de dimensionnalité dans ce rapport. Au cours d'une étude de cas pratique, nous appliquerons ces méthodes à un graphe représentant le réseau routier européen afin de souligner leurs capacités à révéler des données importantes et à rendre la visualisation de données complexes plus simple entre autres.

Chapitre 1

Généralités & Objectifs

1.1 Contexte du réseau routier européen

Le réseau routier européen est un vaste réseau de routes et d'autoroutes qui relie les pays, les villes et les régions du continent européen. Son histoire remonte à des milliers d'années, remontant à l'Antiquité.

1.1.1 Historique

Le réseau routier européen a des racines historiques profondes. À l'époque romaine, les Romains ont construit un réseau routier impressionnant pour faciliter le commerce, les déplacements militaires et la gouvernance de leur empire. De nombreuses routes romaines ont survécu et ont été utilisées comme base pour le développement ultérieur du réseau.

Au Moyen Âge, les routes médiévales ont émergé pour relier les villes et les régions. Ces routes étaient souvent utilisées pour le commerce et les pèlerinages, contribuant ainsi à l'essor des villes et des centres économiques.

1.1.2 Caractéristiques

Le réseau routier européen se caractérise par sa diversité. Il comprend des autoroutes modernes à plusieurs voies, des routes nationales, des routes secondaires et des voies locales. Les autoroutes européennes sont un élément clé du réseau, favorisant la mobilité à grande vitesse à travers le continent.

Les routes européennes sont marquées par une signalisation standardisée, facilitant la navigation pour les conducteurs et les voyageurs. Les infrastructures du réseau incluent des ponts majestueux, des tunnels impressionnantes et des échangeurs complexes.

1.1.3 Coopération internationale

La coopération internationale est essentielle pour maintenir et améliorer le réseau routier européen. L'Union européenne (UE) joue un rôle central dans la coordination des

1.2.3 Détection de structures

En réduisant la dimension, on peut mettre en évidence les structures sous-jacentes dans les données. Cela peut aider à identifier des clusters, des communautés, des motifs ou des relations significatives entre les noeuds du graphe, ce qui est vital dans plusieurs domaines notamment l'étude des réseaux de transport.

1.2.4 Gains d'espace et de temps

Dans des domaines comme la recherche opérationnelle et la planification de réseaux, la réduction de la dimensionnalité peut permettre une résolution plus efficace des problèmes complexes en réduisant le nombre de variables à considérer.

Chapitre 2

Contexte du Projet

Dans ce chapitre, nous examinons diverses approches et techniques proposées pour prédirer les liens au sein des réseaux sociaux, en nous basant sur différentes sources d'inspiration. Nous expliquons les principes sous-jacents de ces approches en introduisant quelques concepts essentiels liés à ces techniques, afin de donner une vue d'ensemble du domaine de la prédition des liens. Pour structurer cet état de l'art, nous commençons par présenter le domaine de l'analyse des réseaux sociaux, en définissant certains concepts et propriétés fondamentaux relatifs aux réseaux sociaux. Ensuite, nous abordons plus en détail notre revue de l'état de l'art, en classant les différentes approches de prédition des liens que nous sommes en mesure de présenter en fonction de catégories principales que nous décrirons brièvement.

2.1 Analyse des réseaux sociaux

2.1.1 Définition

L'analyse des réseaux sociaux est définie comme étant l'étude des entités sociales (les personnes dans les organisations qu'on appelle acteurs) ainsi que leurs interactions et leurs relations .

Ces interactions et relations peuvent être représentées par un graphe, dans lequel chaque noeud représente un acteur et chaque lien est une relation. Nous pouvons étudier les propriétés de la structure et son rôle ainsi que la position et le prestige de chaque acteur social.

Nous pouvons rechercher aussi les différents types de sous-graphes comme par exemple les communautés formées par des groupes d'acteurs ayant des intérêts communs, en isolant le groupe d'individus ayant une densité élevée.

Les réseaux sociaux peut être aussi une source permettant l'élaboration de recommandations : trouver un expert dans un domaine donné, suggérer des produits à vendre, proposer un ami, etc. Cette élaboration peut être fondée sur des algorithmes d'exploration de chemins, d'analyse de degrés...

- **Chemin** : Un chemin est une séquence d'arêtes qui relie deux nœuds.
- **Chemin orienté** : Il s'agit d'une séquence d'arêtes reliant deux nœuds tout en respectant la direction spécifique de chaque arête.
- **Géodésique** : Une géodésique est l'un des plus courts chemins possibles entre deux nœuds donnés dans un graphe.
- **Diamètre** : Le diamètre d'un graphe est la plus grande distance géodésique entre deux noeuds du graphe.
- **Graphe complet** : Un graphe est considéré comme complet lorsqu'il existe une arête reliant chaque paire de nœuds, ce qui signifie que chaque noeud est directement connecté à tous les autres nœuds du graphe.

2.2 Description du jeu de données

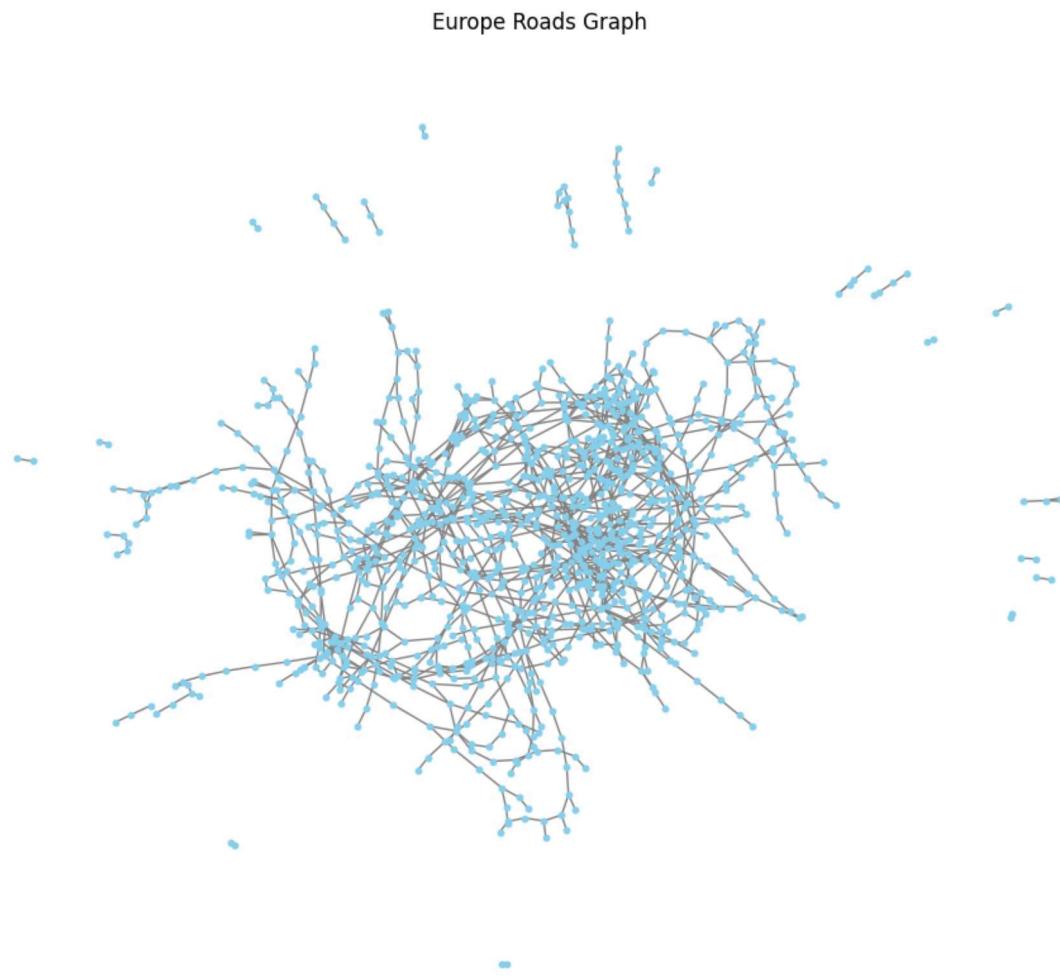


FIG. 2.1 : Représentation graphique du graphe europa roads

Chapitre 2. Contexte du Projet

- Il y a un total de 96 triangles dans le graphique, ce qui suggère une certaine structure de clustering, bien que le nombre moyen de triangles soit faible.
- Le coefficient de clustering moyen est de 0,0167316, ce qui indique un faible niveau de connectivité locale entre les noeuds.



FIG. 2.2 : La carte réelle des routes en europe

Le graphique Euro-Road semble représenter un réseau routier européen avec des caractéristiques de connectivité et de clustering. Cependant, il est plutôt sparse, ce qui suggère que certaines régions peuvent être moins bien reliées que d'autres. Le coefficient d'assortativité positif indique une tendance à la formation de regroupements de noeuds similaires, ce qui pourrait correspondre à des réseaux routiers régionaux ou nationaux distincts.

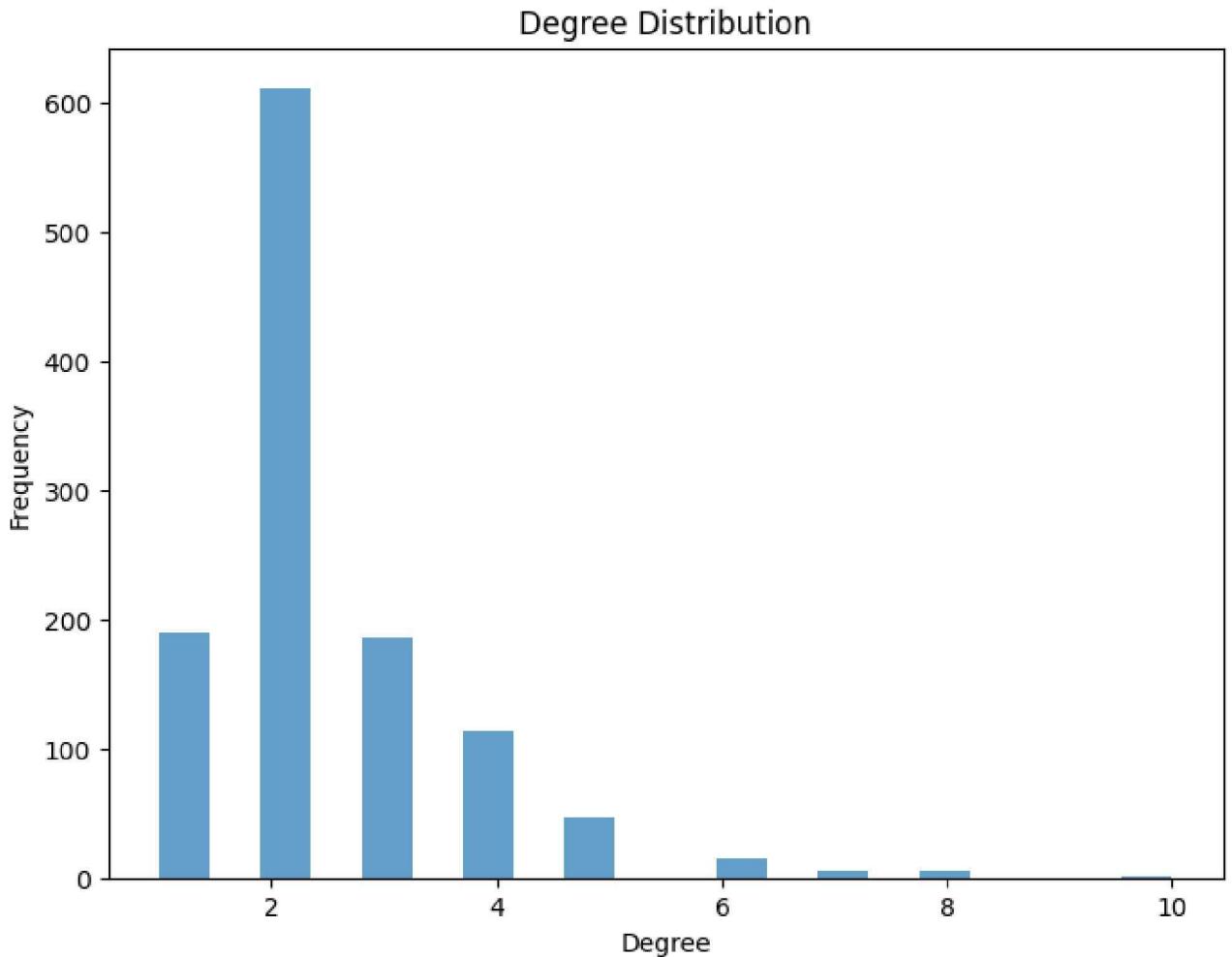


FIG. 2.4 : La distribution des degrés dans le graphe

2.3 Prédiction des liens

Les réseaux sont de plus en plus utilisés pour modéliser des systèmes complexes composés d’éléments en interaction, tels que : les réseaux sociaux, les réseaux biologiques qui ont été décrit dans les sections précédentes.

Différentes études ont montré qu’il est possible de prédire de nouvelles relations entre les éléments présents dans la topologie d’un réseau. Cette thématique qui consiste à chercher de nouvelles relations dans les réseaux est appelée prédiction de lien. Elle vise à prédire le comportement de lien, c'est-à-dire si une relation entre deux éléments dans un réseau peut être créée ou si une relation entre eux est manquante en basant sur les relations actuellement observées. Beaucoup d'études et de recherches ce sont orientés vers ce domaine compte tenu de son champ d'application.

Pour cette raison plusieurs méthodes ont été conçues et appliquées pour rechercher et de prédire des liens dans différents types de réseaux.

Chapitre 3

Implémentation et Expérimentations

3.1 Analyse en Composantes Principales (PCA)

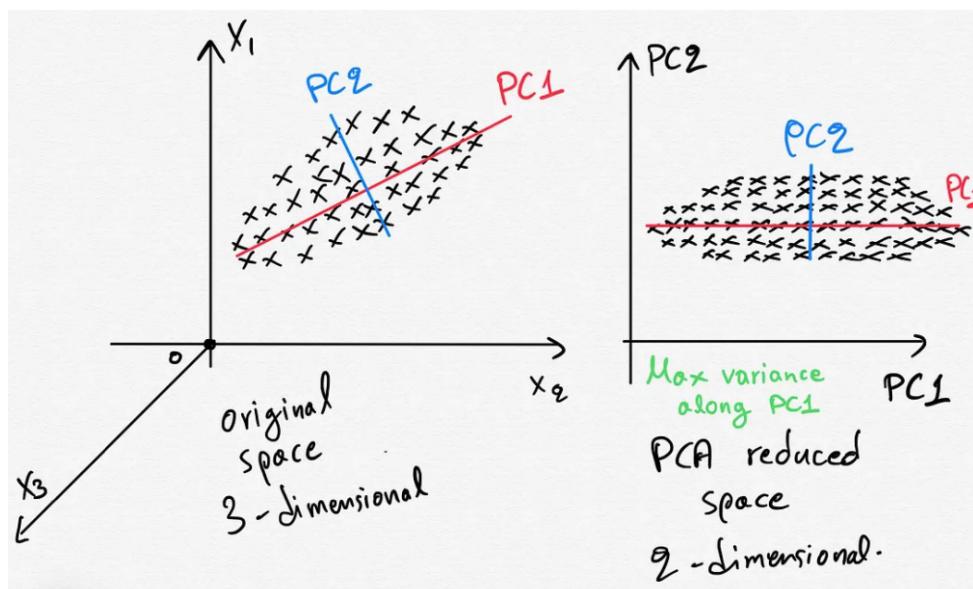


FIG. 3.1 : PCA algorithm

L'Analyse en Composantes Principales (PCA) est une technique puissante de réduction de dimensionnalité qui vise à extraire des informations pertinentes à partir de données multidimensionnelles tout en réduisant leur dimension. Cette méthode est couramment utilisée en analyse de réseaux sociaux pour simplifier la représentation des données tout en préservant leur structure essentielle.

3.1.1 Objectif de la PCA

L'objectif principal de la PCA est de transformer un ensemble de données multidimensionnelles en un nouvel espace de dimension réduite, tout en préservant au maximum l'information contenue dans les données. Cela permet de simplifier la complexité des données tout en maintenant leur structure fondamentale.

composantes principales expliquent la majeure partie de la variance des données, ce qui permet de réduire la dimension tout en préservant l'information cruciale.

En conclusion, la PCA est un outil essentiel en analyse de réseaux sociaux pour la réduction de dimensionnalité et la simplification des données tout en maintenant leur structure fondamentale.

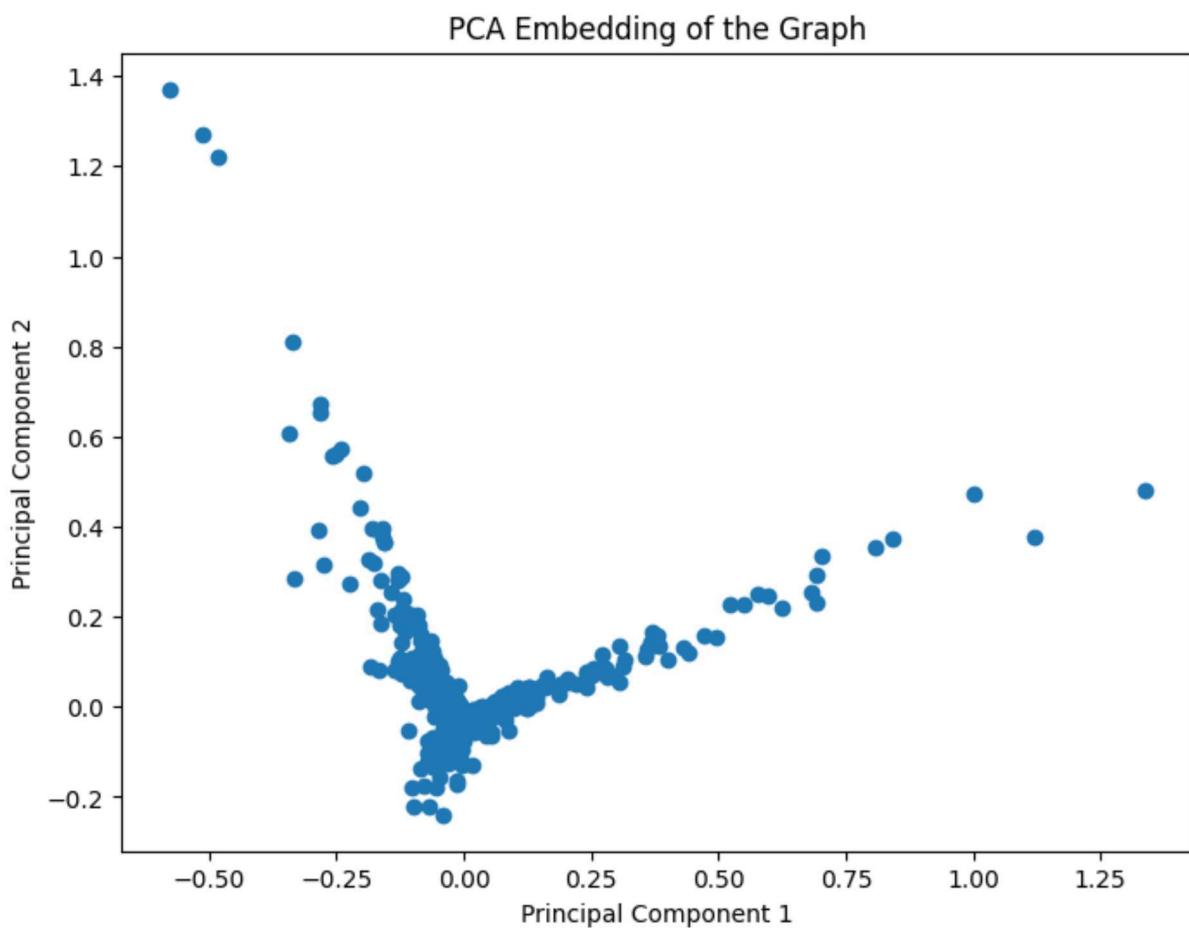


FIG. 3.2 : Representation du graphe en 2D après application de l'ACP

3D Scatter Plot of Embedding

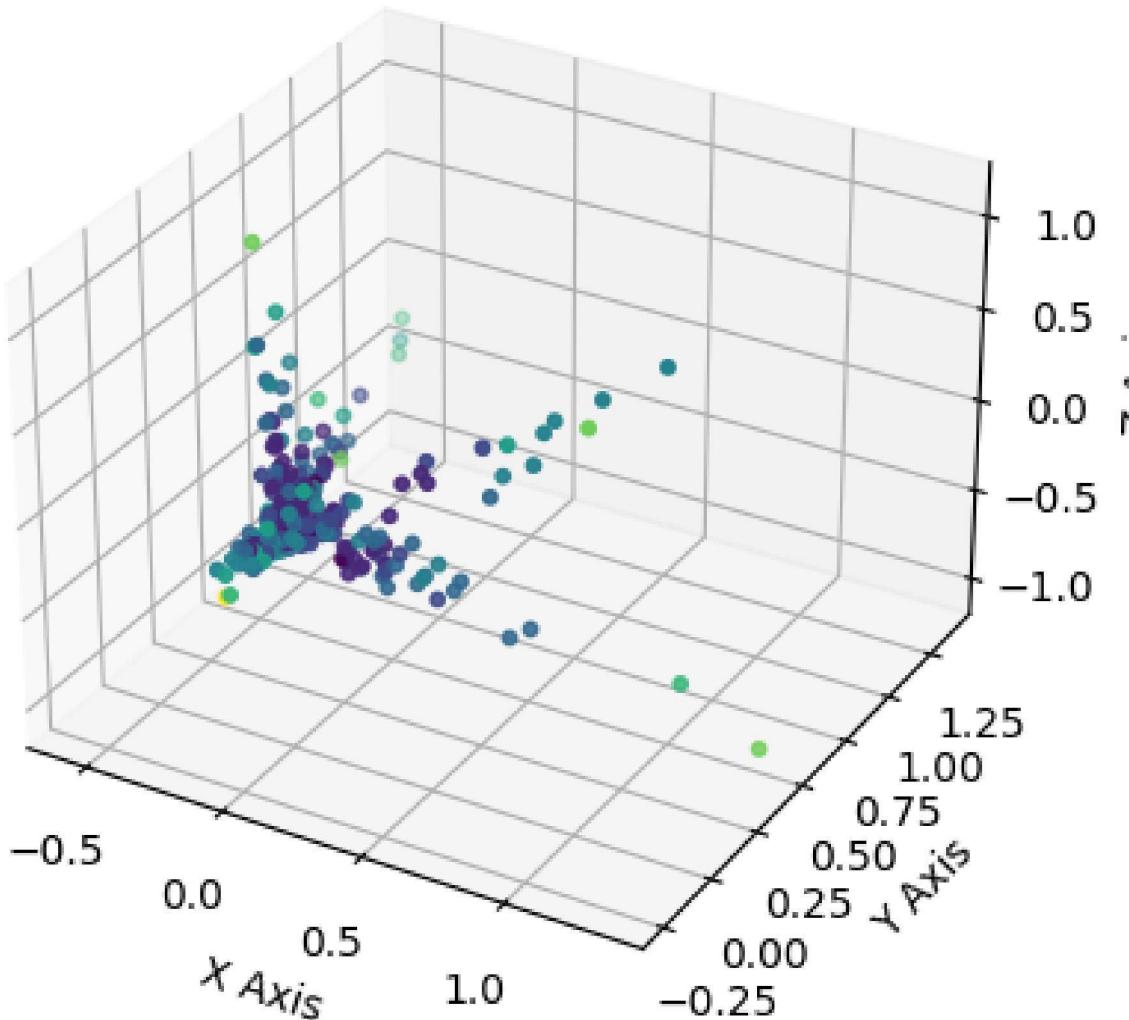


FIG. 3.3 : Représentation du graphe en 3D après application de l'ACP

3.2 Locally Linear Embedding (LLE)

Locally Linear Embedding (LLE) est une méthode de réduction de la dimensionnalité non linéaire proposée par Sam T. Roweis et Lawrence K. Saul en 2000 dans leur article intitulé « Réduction de la Dimensionnalité Non Linéaire par Embedding Localement Linéaire ». L'algorithme Locally Linear Embedding (LLE) est utilisé pour conserver les relations locales entre les données dans un espace de dimension inférieure. Il est largement utilisé pour la visualisation de données et la prédition de liens importants dans les réseaux de graphe.

After applying the LLE Algorithm :

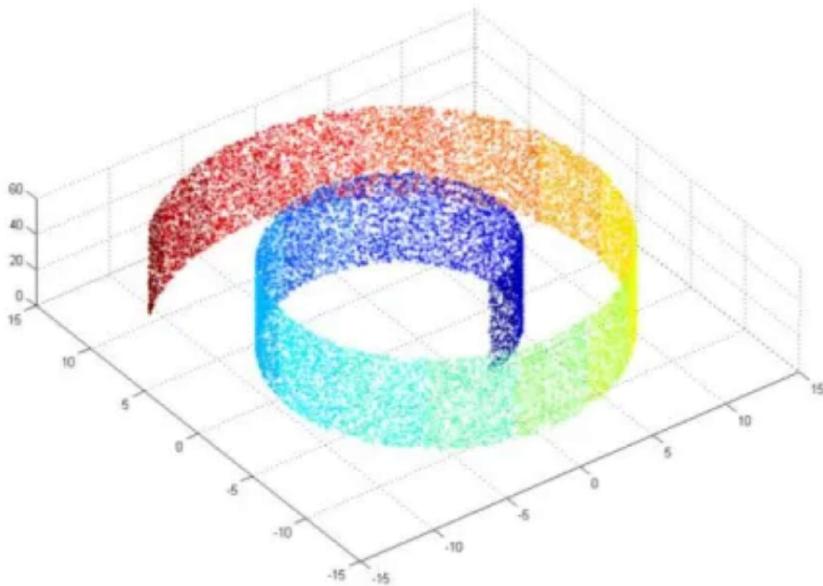


FIG. 3.4 : Swiss Roll features in 3 dimensions

3.2.1 Fonctionnement de LLE

LLE fonctionne en trois étapes principales :

3.2.1.1 Étape 1 : Sélection des Voisins

Pour chaque point de données, LLE identifie ses voisins les plus proches dans l'espace de dimension supérieure à l'aide d'une mesure de distance appropriée. Ces voisins forment le voisinage local du point.

3.2.1.2 Étape 2 : Reconstruction Linéaire Locale

Une fois les voisins identifiés, LLE approxime chaque point de données comme une combinaison linéaire de ses voisins locaux. Cette approximation est réalisée en minimisant l'erreur de reconstruction sous contrainte. Les poids de la combinaison linéaire sont ajustés pour obtenir la meilleure reconstruction possible.

Les formules mathématiques pour cette étape sont les suivantes :

$$\min_{w_{ij}} \sum_j \|x_i - \sum_j w_{ij}x_j\|^2$$

sous contrainte :

$$\sum_j w_{ij} = 1$$

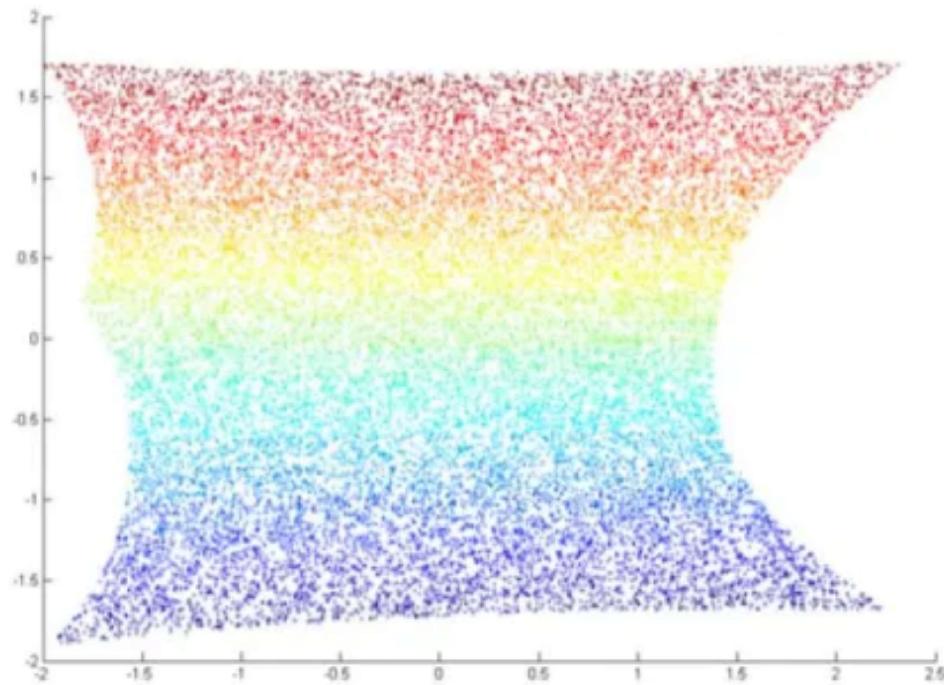


FIG. 3.5 : Swiss Roll features using LLE

3.2.1.3 Étape 3 : Construction de la Carte Globale

Une fois que les poids de reconstruction ont été calculés, une carte globale est construite en plaçant chaque point de données dans un espace de dimension inférieure de manière à préserver autant que possible les relations linéaires locales.

La formule mathématique pour cette étape est la suivante :

$$\min_{y_i} \sum_i \|y_i - \sum_j w_{ij}y_j\|^2$$

3.2.2 Prédiction de Liens Importants

LLE est également utilisé pour la prédiction de liens importants dans les réseaux de graphe. Après avoir réduit la dimension des données à l'aide de LLE, vous pouvez calculer les distances entre les points réduits. Les paires de points qui sont proches les unes des autres dans l'espace réduit peuvent indiquer des liens importants ou des relations similaires dans le graphe d'origine.

LLE offre ainsi une approche puissante pour explorer et analyser les données de manière efficace, en préservant les structures locales et en permettant la détection de liens significatifs dans les réseaux de graphe.

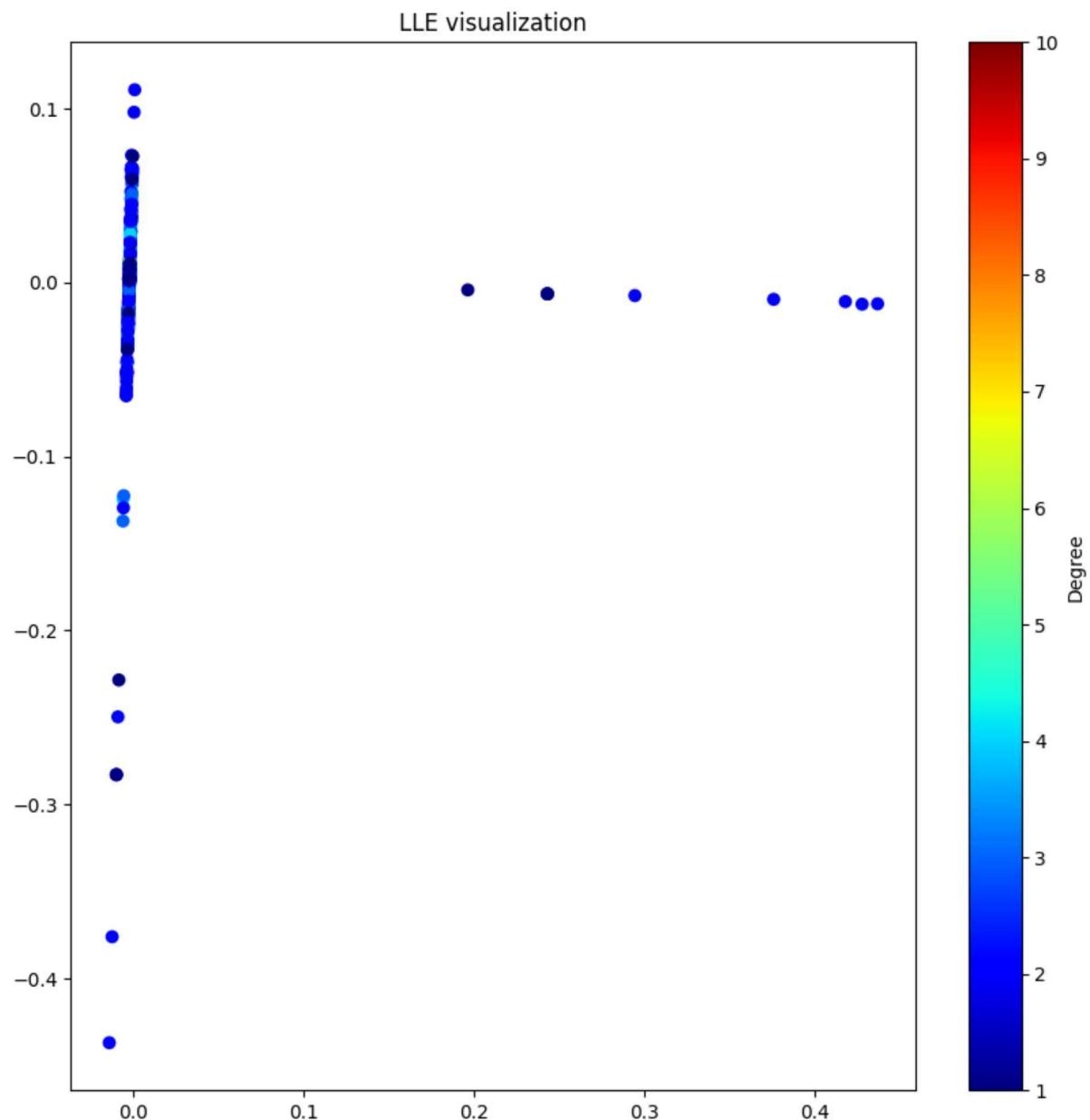


FIG. 3.6 : Representation du graphe en 2D après application de LLE

3D Scatter Plot of Embedding

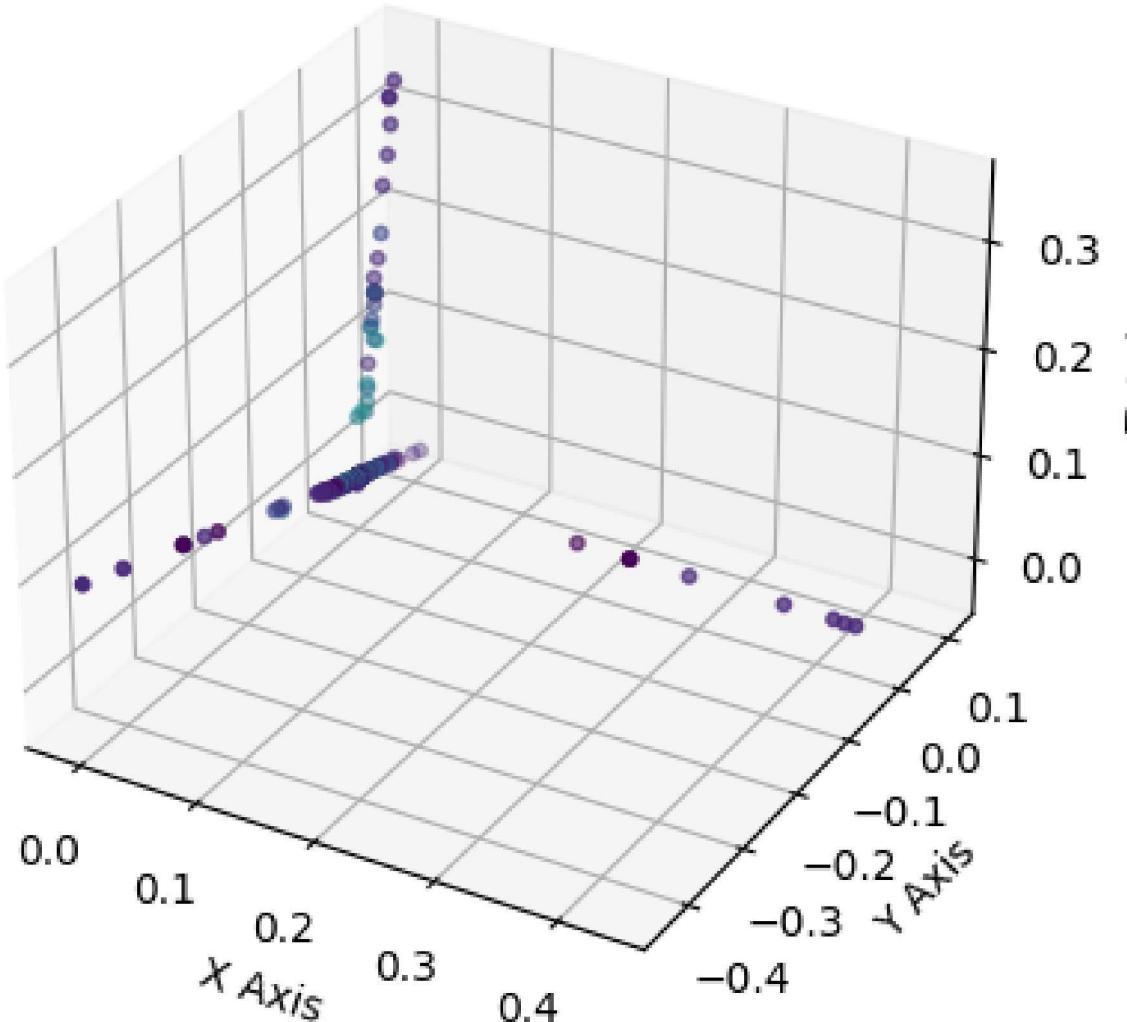


FIG. 3.7 : Representation du graphe en 3D après application de LLE

3.3 Uniform Manifold Approximation & Projection (UMAP)

UMAP est un algorithme de réduction non-linéaire de dimension dont le principe général est similaire à celui de LLE : chercher un jeu de données dans un espace de dimension réduite qui présente les mêmes structures locales au voisinage de chaque point. Cependant, UMAP généralise la notion de voisinage déterministe de LLE à une notion probabiliste. On cherchera non plus à conserver les voisins d'une observation, mais plutôt à conserver les probabilités que d'autres observations lui soient voisines. Nous allons chercher une représentation des données en plus faible dimension qui présente la même topologie que le nuage des observations dans l'espace de départ. Pour ce faire, nous allons construire

une matrice de similarités représentant la similarité entre chaque paire de points, puis nous chercherons l'ensemble des points dans l'espace d'arrivée qui vérifie le même graphe de similarité.

3.3.1 Similarité dans l'espace de départ

Considérons une matrice d'observations X de n observations de dimensions D. Le voisinage d'un point x_i dans X est représenté par la probabilité conditionnelle $p_{i,j} = p(x_i|x_j)$ pour toute observation x_j dans X. Cette probabilité représente la probabilité que x_j soit considéré comme étant un voisin de x_i . Pour que cette définition tienne, cette valeur de probabilité devra bien entendu être entre 0 et 1 et décroître avec la distance entre x_i et x_j . L'objectif de UMAP est de déterminer le nuage de points représenté par les observations réduites y_i de telle sorte à ce que $p(y_j|x_i) \approx p(x_j|x_i)$ pour tous les couples (i,j). Pour concrétiser cette probabilité, UMAP la définit de la façon suivante :

$$p_{j|i} = \exp\left(-\frac{\|x_i - x_j\|_2 - \rho_i}{\sigma_i}\right)$$

ρ_i est définie comme la distance entre x_i et son plus proche voisin, quant à σ_i c'est un coefficient de normalisation qui va forcer la similarité à se situer entre 0 et 1.

3.3.2 Similarité dans l'espace d'arrivée

Comme pour LLE, nous cherchons donc désormais une matrice réduite Y, de dimension $d < D$ de sorte à ce que les probabilités conditionnelles $q_{i,j} = p(y_j|y_i)$ soient les plus proches possible de $p_{i,j}$. Un inconvénient de la similarité utilisée est sa décroissance rapide. Cela permet de facilement distinguer les voisins (proches de x_i) des non-voisins, pour qui la similarité sera proche de 0. Cette propriété est utile dans l'espace de départ mais moins dans l'espace d'arrivée. En effet, supposons que l'on souhaite projeter les données dans un plan, donc un espace à deux dimensions, pour la visualisation, alors les voisins seront agglutinés pour être dans la zone non-nulle de la gaussienne. En effet, pour un même nombre de points, il y a moins de « volume » disponible en 2D que dans un espace à n dimensions. Les points qui sont voisins se retrouveront placés très proches les uns des autres, tandis que les non-voisins pourront être placés n'importe où ailleurs, puisque la similarité vaut de toute façon zéro. Autrement dit, on risque de produire une visualisation avec de nombreux points qui se chevauchent, séparés par de grands espaces de vide. On appelle ce problème le problème de « l'agglutinement » (ou crowding problem).

Pour éviter ce problème d'agglutinement, UMAP utilise une autre définition de la probabilité conditionnelle dans l'espace réduit d'arrivée, qui va encourager les points à se disperser dans le plan, UMAP va utiliser une loi de similarité inspirée de la loi t-Student :

$$q_{i,j} = \frac{1}{1 + a(y_i - y_j)^{2b}}$$

a et b étant des paramètres réglables. En pratique, la fonction idéale n'autoriserait jamais deux points dans l'espace d'arrivée à être plus proche qu'une certaine distance

- **n_neighbors** : Ce paramètre contrôle la manière dont l'UMAP équilibre la structure locale par rapport à la structure globale des données. Pour ce faire, il limite la taille du voisinage local que l'UMAP examinera lorsque l'algorithme tentera d'apprendre la structure des données. Une valeur plus élevée de n_neighbors produit une représentation plus lisse et conserve davantage de structures globales, tandis qu'une valeur plus faible peut capturer des structures locales fines. Cependant, une valeur trop élevée peut entraîner une perte d'informations locales.
- **min_dist** : Comme vu plus haut, ce paramètre sert à prévenir la concentration excessive des points, ce qui peut provoquer un agglutinement. Il permet donc de contrôler la distance minimale entre les points dans l'espace de basse dimension. Une valeur plus faible de min_dist permettra aux points de se regrouper plus étroitement, tandis qu'une valeur plus élevée les séparera davantage.
- Nous devons donc régler ces paramètres de telles façons a préservé la structure globale du graphe qui est d'un cluster très dense, ce qui est normal vu la structure réelle et réseautage des routes européennes.

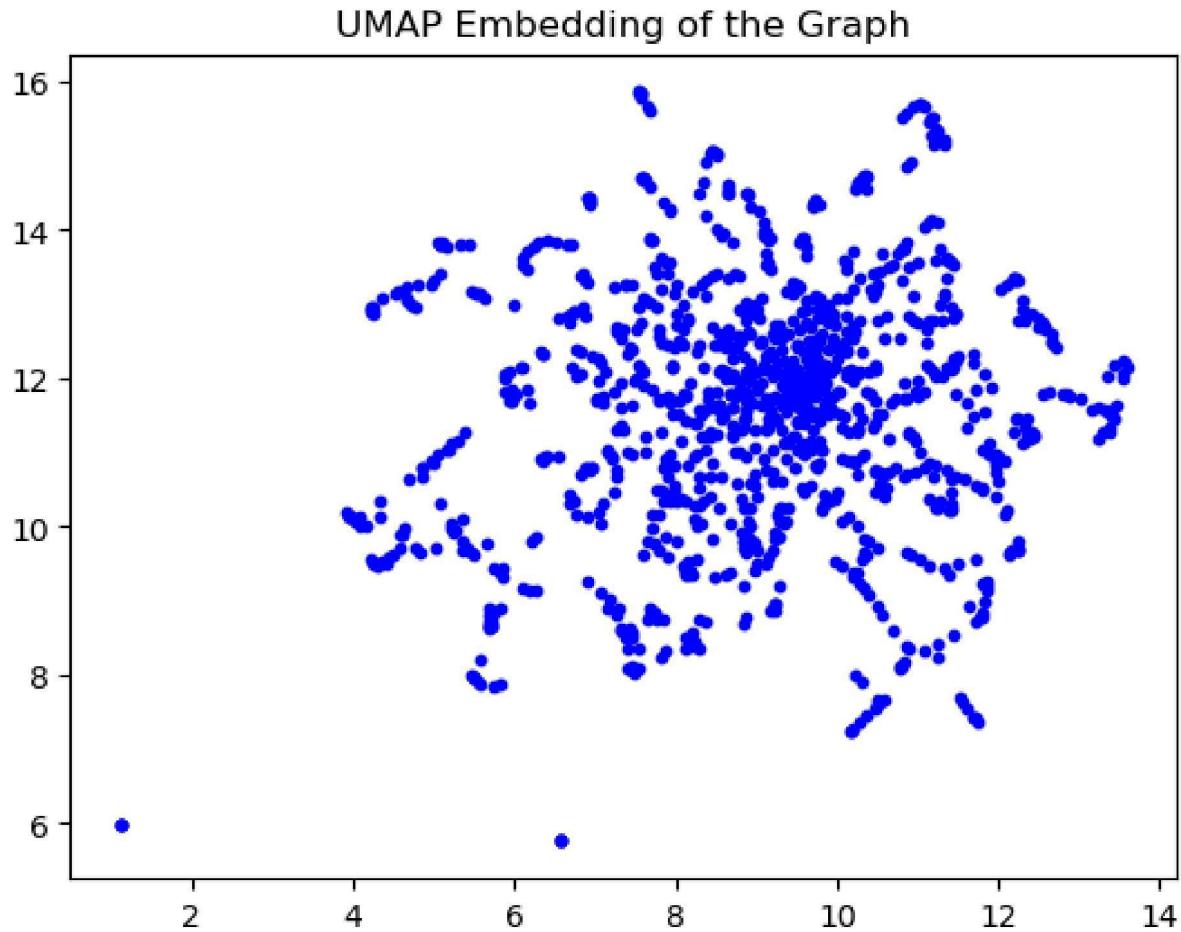


FIG. 3.8 : Representation du graphe en 2D après application d'UMAP

UMAP 3D Embedding of the Graph

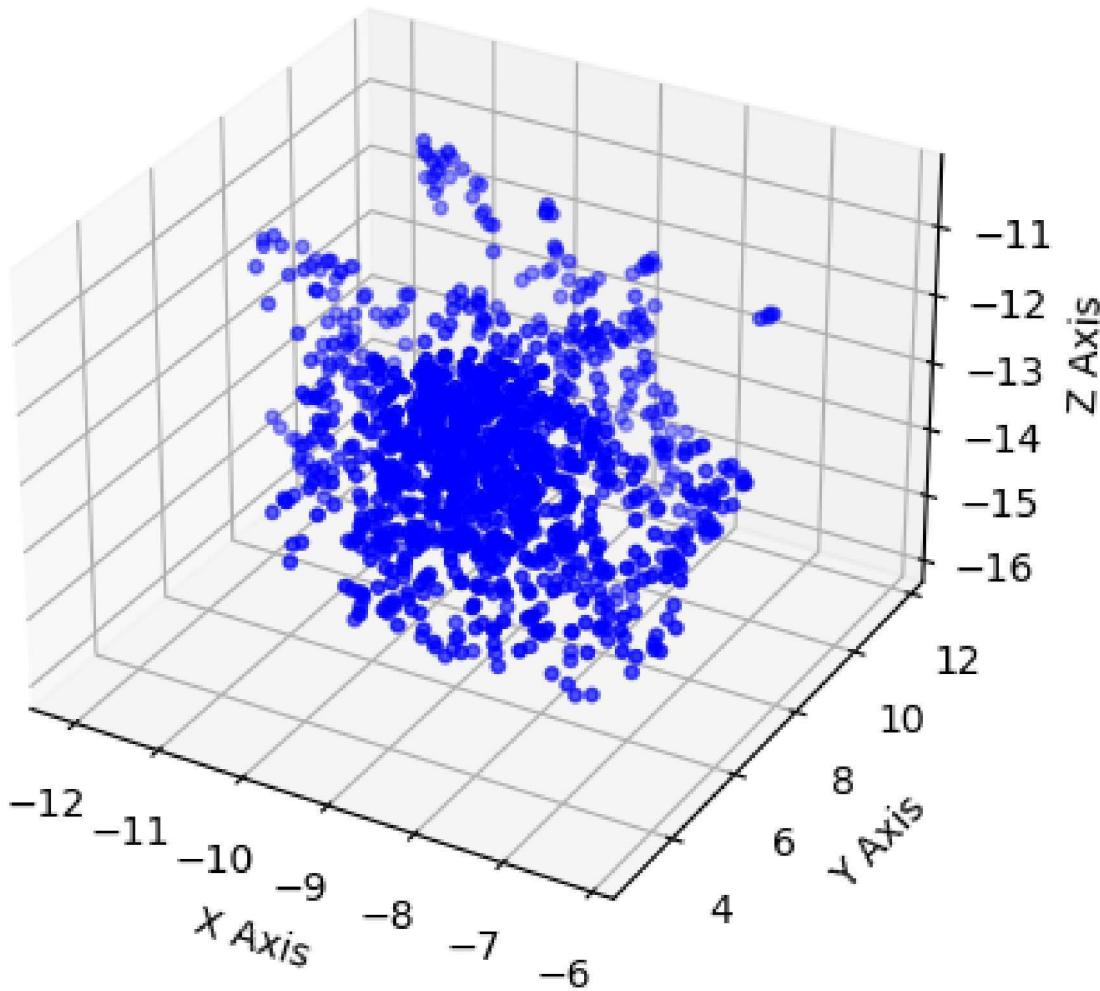


FIG. 3.9 : Representation du graphe en 3D après application d'UMAP

Après visualisation, nous pouvons constater que UMAP a bien conservé la structure du graphe après réduction de sa dimension.

Conclusion et perspectives

Conclusion générale

Dans le cadre de ce projet d'analyse de réseaux sociaux, nous avons exploré différentes méthodes de prédiction de liens en nous concentrant sur des techniques de réduction de dimensionnalité telles que PCA, LLE et UMAP. Notre objectif était de comprendre comment ces méthodes pouvaient être appliquées à des graphes de réseaux pour anticiper les liens futurs entre les noeuds.

Tout au long de ce projet, nous avons réalisé que la prédiction de liens dans les réseaux sociaux revêt une grande importance dans divers domaines d'application. De la recommandation d'amis dans les réseaux sociaux traditionnels à la détection de collaborations potentielles entre chercheurs dans les réseaux académiques, en passant par la personnalisation des recommandations de produits dans le commerce électronique, la prédiction de liens a des implications profondes.

Nous avons également constaté que les méthodes de réduction de dimensionnalité, telles que PCA, LLE et UMAP, offrent des approches innovantes pour analyser et simplifier des graphes de réseaux complexes tout en maintenant la pertinence des informations. PCA se distingue par sa capacité à extraire des composantes linéaires principales, tandis que LLE et UMAP explorent des relations non linéaires plus complexes.

Dans nos analyses, nous avons pu appliquer avec succès ces méthodes à un graphe de réseau réel, en réduisant sa dimensionnalité tout en préservant les caractéristiques importantes de la structure du réseau. Les résultats obtenus montrent que ces méthodes peuvent être précieuses pour la prédiction de liens dans les réseaux sociaux.

En fin de compte, ce projet nous a permis de plonger profondément dans le domaine de l'analyse de réseaux sociaux et de comprendre comment les méthodes de réduction de dimensionnalité peuvent contribuer à résoudre des problèmes de prédiction de liens. Bien que chaque méthode ait ses avantages et ses limites, elles offrent toutes des perspectives prometteuses pour des applications futures dans divers domaines.

En conclusion, ce projet a renforcé notre compréhension de la complexité des réseaux sociaux et des techniques innovantes qui peuvent être utilisées pour en extraire des informations précieuses. Il ouvre la voie à de futures recherches et à l'exploration de méthodes plus avancées pour améliorer la prédiction de liens dans les réseaux sociaux.

Bibliographie

- [1] <https://www.geeksforgeeks.org/principal-component-analysis-pca/>
- [2] <https://medium.com/analytics-vidhya/locally-linear-embedding-lle-data-mining-b956616d24e9>
- [3] <https://towardsdatascience.com/pca-clearly-explained-how-when-why-to-use-it-and-feature-importance-a-guide-in-python-7c274582c37e>
- [4] <https://cedric.cnam.fr/vertigo/Cours/ml/coursReductionDimension.html#id5>
- [5] <https://umap-learn.readthedocs.io/en/latest/index.html>
- [6] Analyse Mathématique de l'algorithme de réduction de dimension UMAP - Antoine Commaret