Rapport – Projet 2 : Analyse de données

# Point 1 : Le graphe du Karaté Club

Nous avons décidé d’utiliser les librairies networkx (pour la manipulation des graphes) et matplotlib (pour leur affichage). Nous importons le graphe avec la fonction karate\_club\_graph() et l’affichons ensuite.

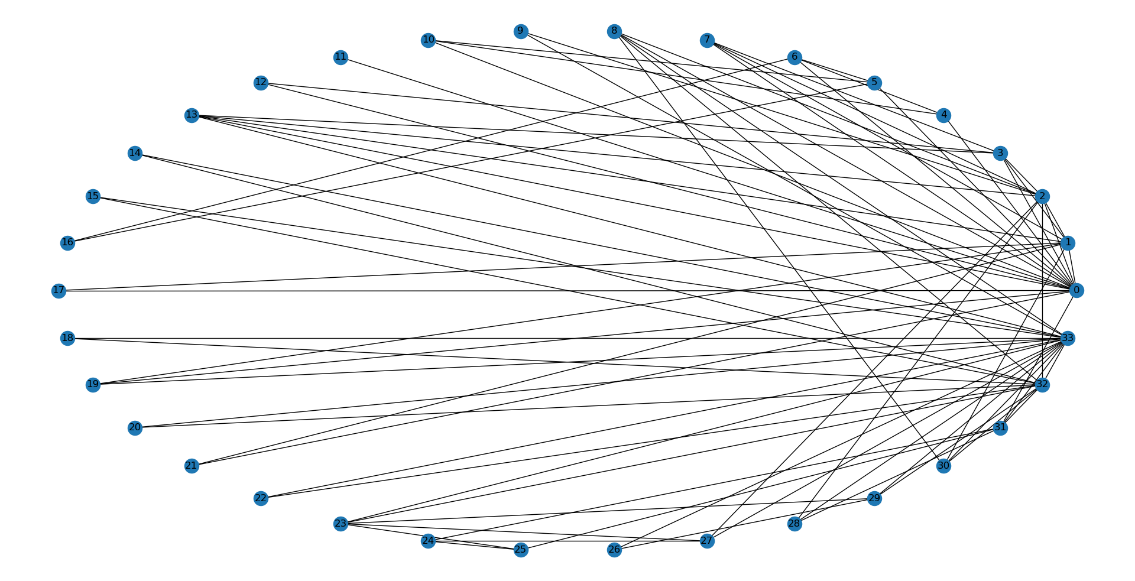


Figure 1 : Graphe du Karaté Club

# Point 2 : La distribution des degrés et de différentes centralités

## La distribution des degrés

Premièrement, nous trions les degrés des nœuds par ordre décroissant. Ensuite, nous stockons cette séquence de degrés dans un objet Counter afin de pouvoir la manipuler plus facilement. A partir de cet objet, nous créons deux listes. L’une contenant l’ensemble des différentes valeurs de degrés du graphe (trié par ordre décroissant) et l’autre, le nombre respectif de degrés de cette valeur présent dans le graphe. C’est-à-dire : (17, 16, 12, 10, 9, 6, 5, 4, 3, 2, 1) et (1, 1, 1, 1, 1, 2, 3, 6, 6, 11, 1). Enfin, nous affichons ces deux listes sous la forme d’un histogramme avec par-dessus le graphe du Karaté Club.

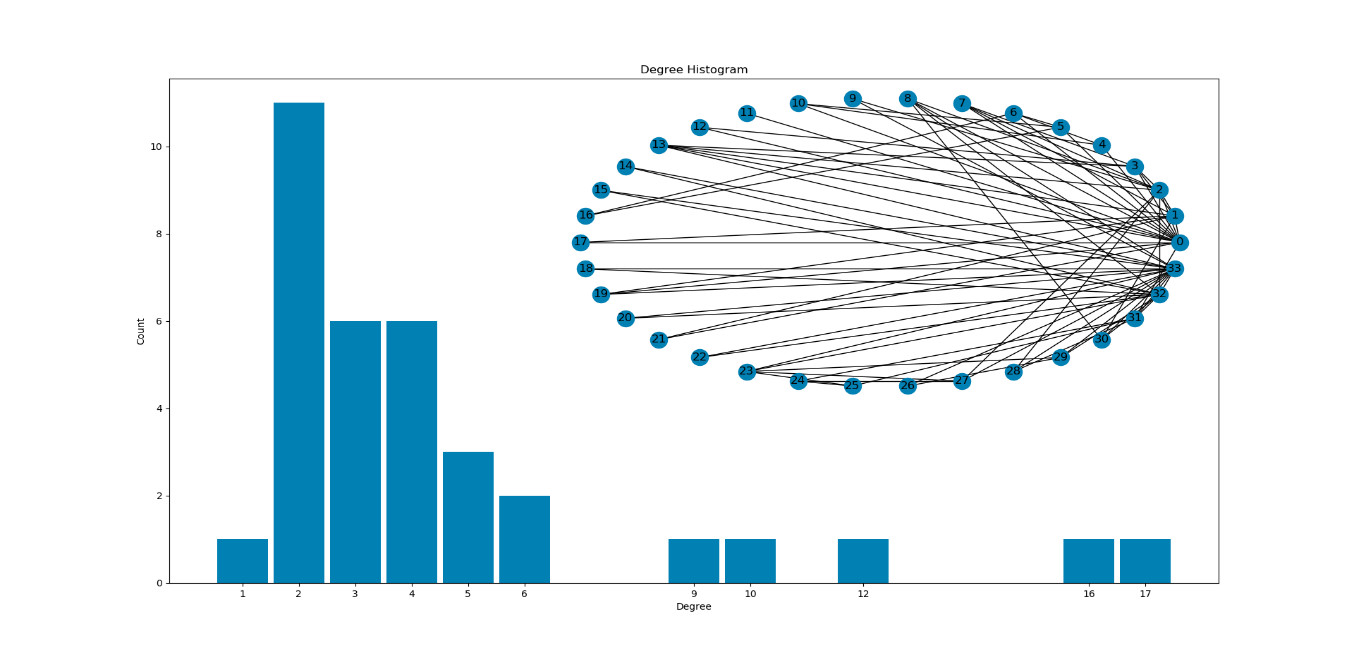


Figure 2 : Distribution de degrés

## La distribution de différentes centralités

Nous calculons les différentes centralités à l’aide des fonctions suivantes closseness\_centrality(), betweenness\_centrality(), katz\_centrality\_numpy(), pagerank(), degree\_centrality().

### Closseness centrality

Cette mesure indique à quel point un nœud est proche des autres. Elle est calculée comme la moyenne des chemins les plus court de ce nœud à tous les autres. Par exemple pour le nœud n°16 la « closseness centrality » est égale à 33 / (2+ 3+ 3+ 3+ 2+ 1+ 1+ 3+ 3+ 4+ 2+ 3+ 3+ 3+ 5+ 5+ 3+ 5+ 3+ 5+ 3+ 5+ 5+ 4+ 4+ 5+ 4+ 4+ 5+ 4+ 3+ 4) = 0.29. Et celle du nœud n°0 est égale à 33 / (1+ 1+ 1+ 1+ 1+ 1+ 1+ 1+ 2+ 1+ 1+ 1+ 1+ 3+ 3+ 2+ 1+ 3+ 1+ 3+ 1+ 3+ 3+ 2+ 2+ 3+ 2+ 2+ 3+ 2+ 1+ 2) = 0.59

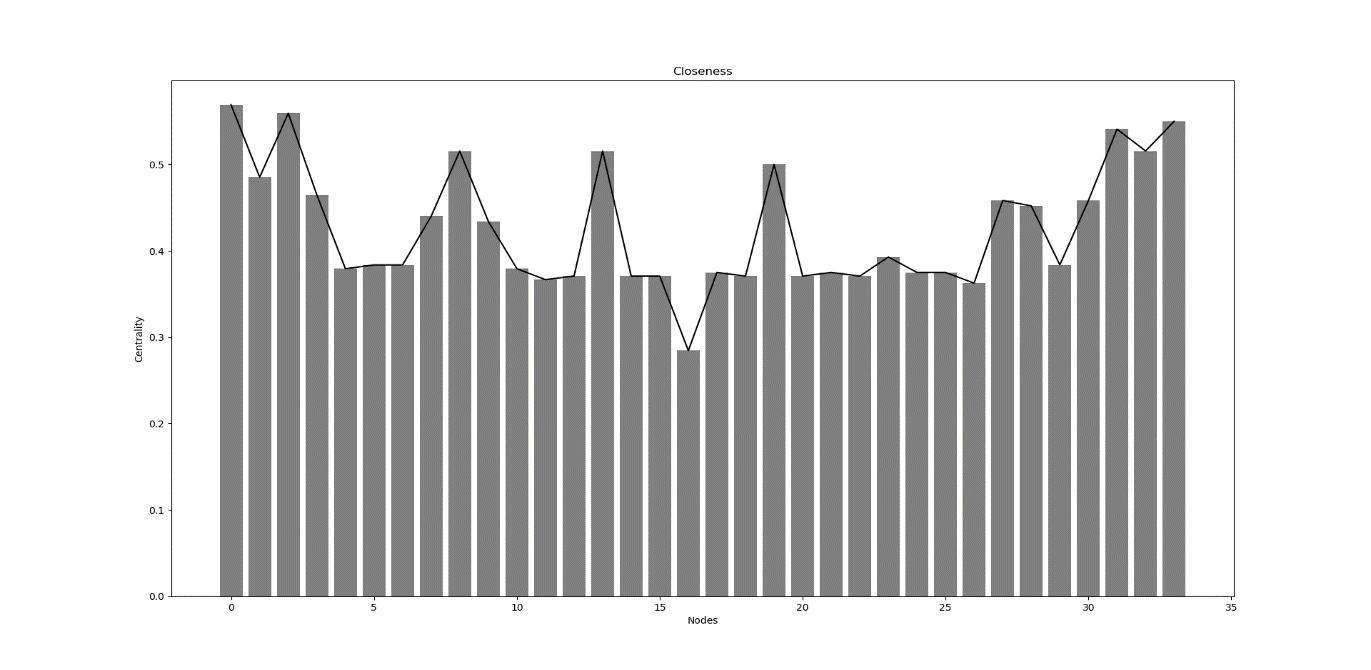


Figure 3 : Closeness centrality

### Betweenness centrality

C’est une autre mesure de la centralité d’un nœud. C’est le nombre de fois qu’un nœud est sur le chemin le plus court entre deux autres nœuds quelconques.

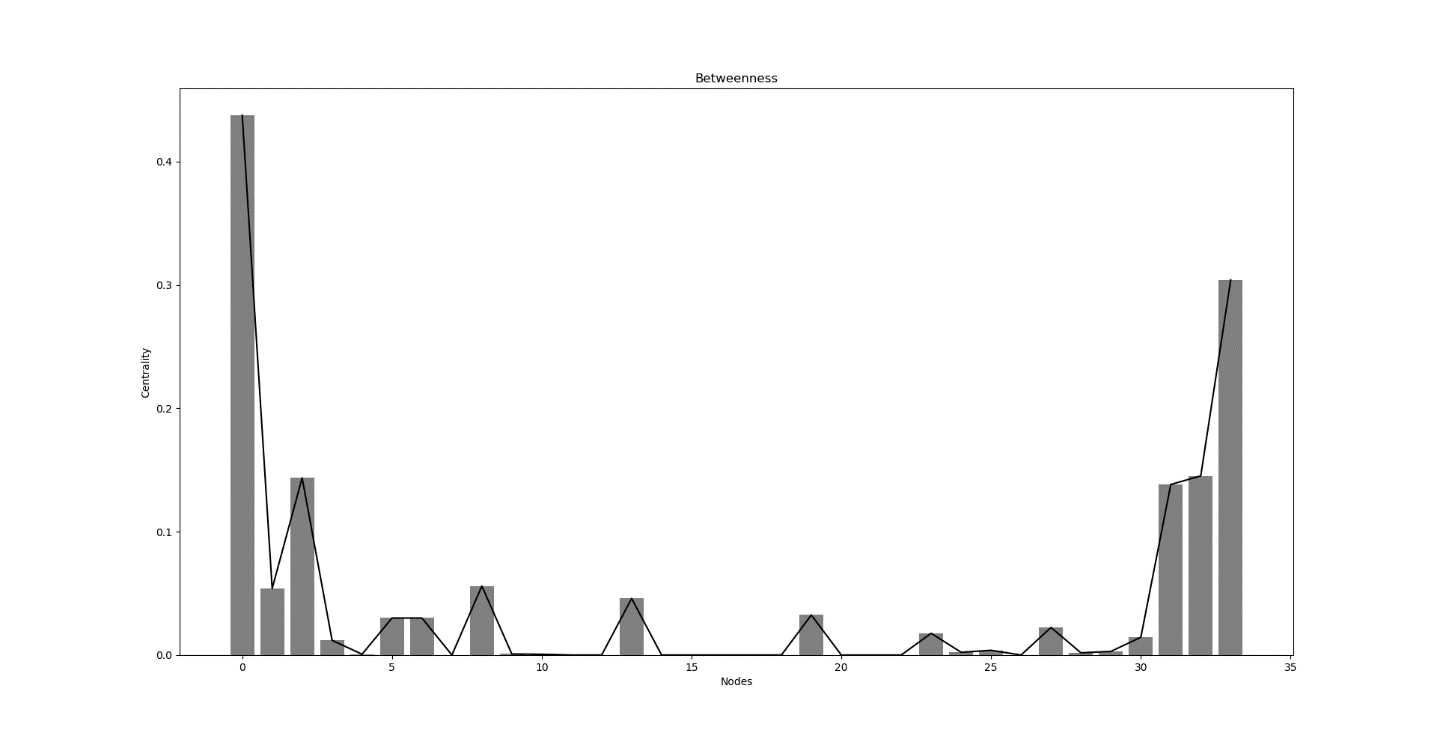


Figure 4 : Betweenness centrality

### Katz centrality

C’est une autre mesure de la centralité d’un nœud, calculée en tenant compte du nombre de voisin direct mais aussi les autres nœuds qui se connecte à ce nœud en passant par les voisins directs.

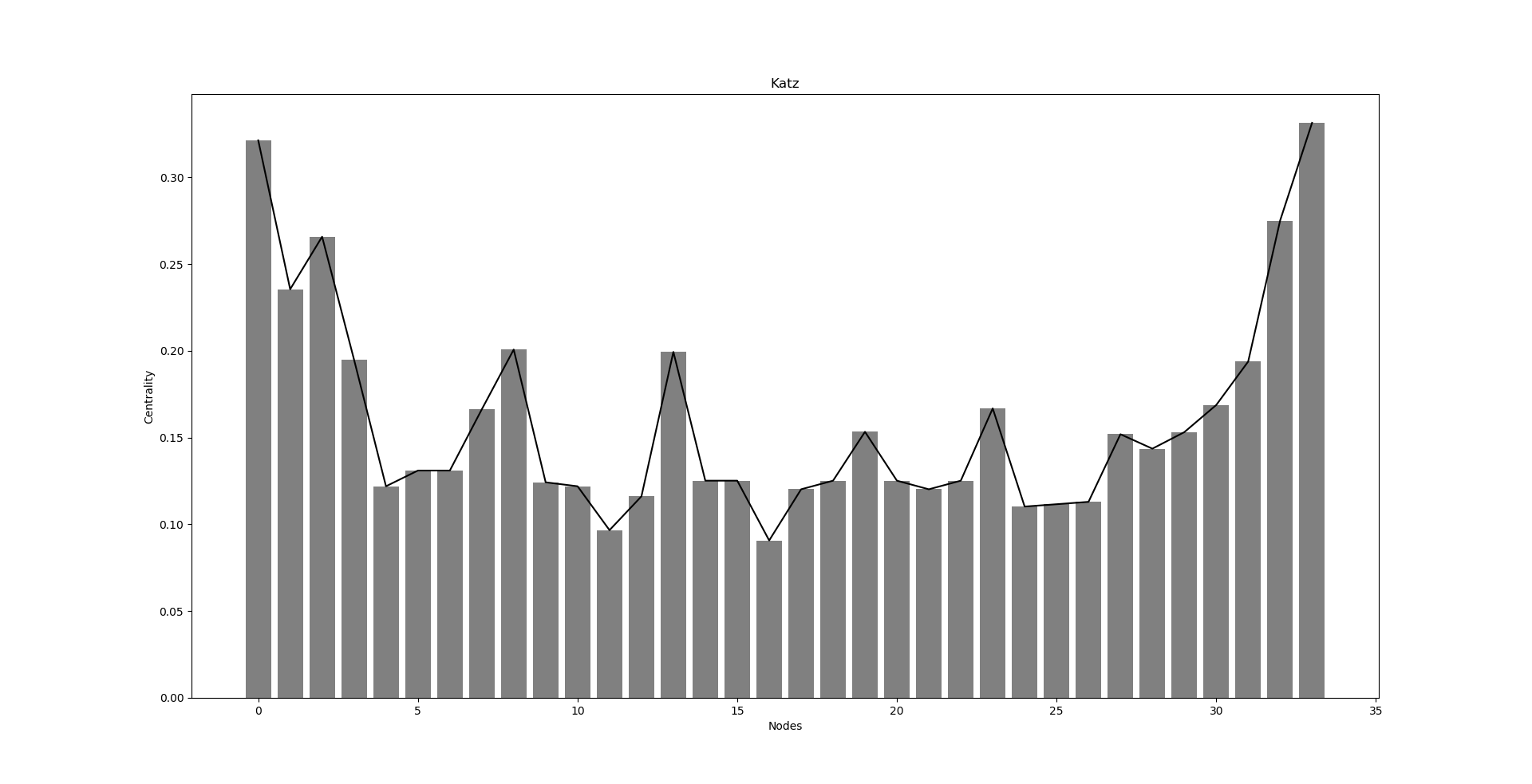


Figure 5 : Katz centrality

### PageRank

PageRank calcule la centralité d’un nœud en prenant en compte, l’importance des nœuds voisin.

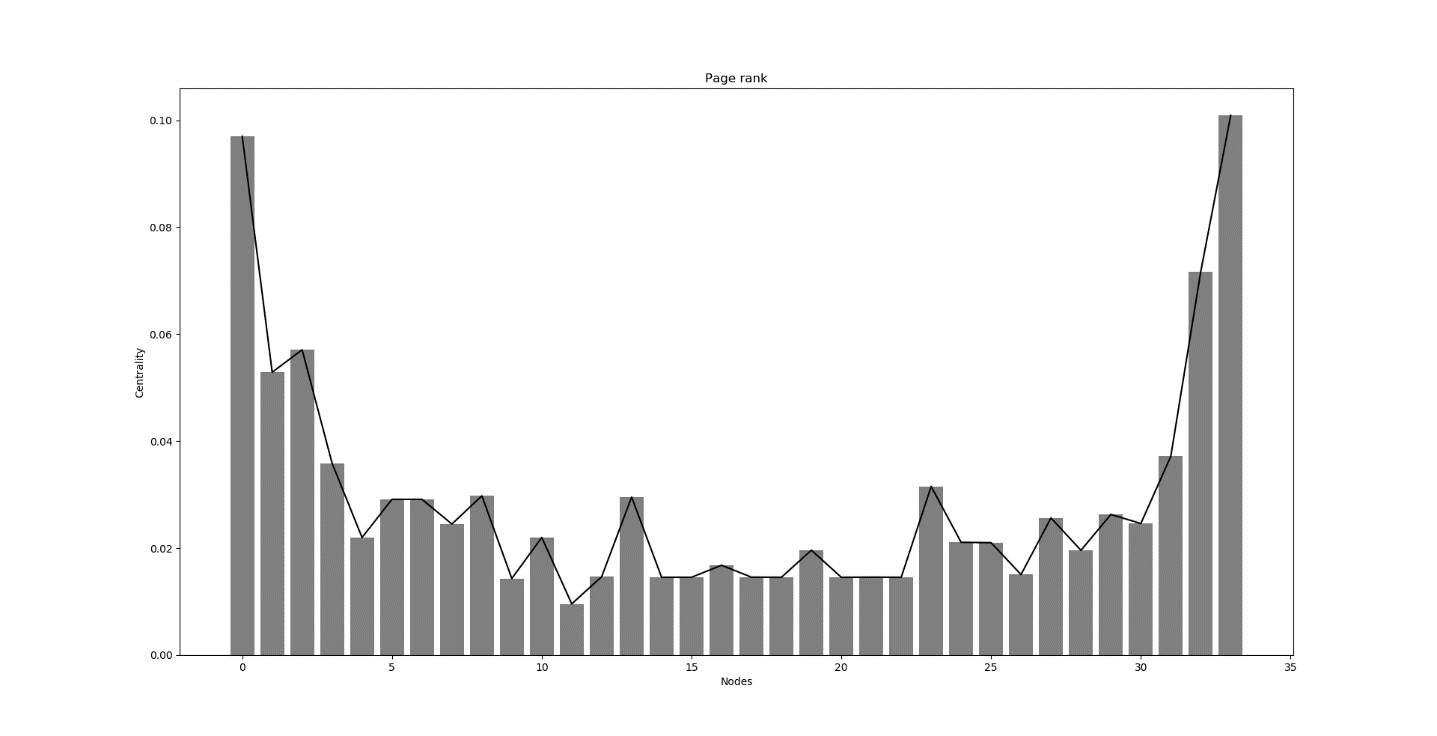


Figure 6 : Page Rank

### Degree centrality

Cette mesure de la centralité est la plus simple car la centralité d’un nœud est simplement égale à la valeur de son degré.

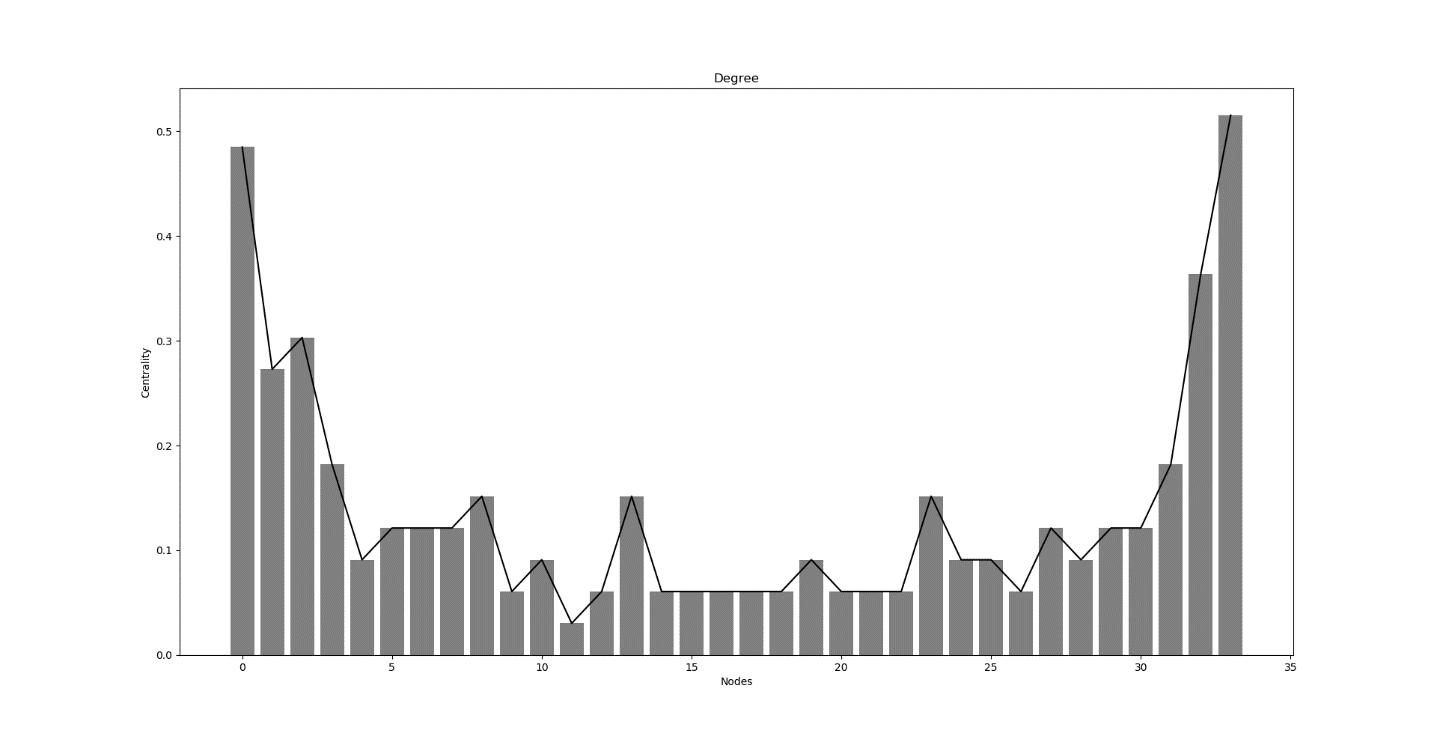


Figure 7 : Degree centrality

# Point 3: Le clustering coefficient (Global)

Ce coefficient mesure le regroupement des nœuds dans un réseau c’est-à-dire à quel point les nœuds voisins sont connecté entre eux.

Dans le graphe du Karaté Club le clustering coefficient est égal à 0,57.

# Point 4 : Le modèle de configuration de degré

Le modèle de configuration de degré est une méthode qui utilise une séquence de degré donnée afin de générer un réseau aléatoire. Il peut donc exister dans ce réseau des boucles et des liens multiples entre deux nœuds. La fraction de liens à éliminer est de l’ordre de 10%, sur 78 liens entre 6 et 10 sont supprimé.

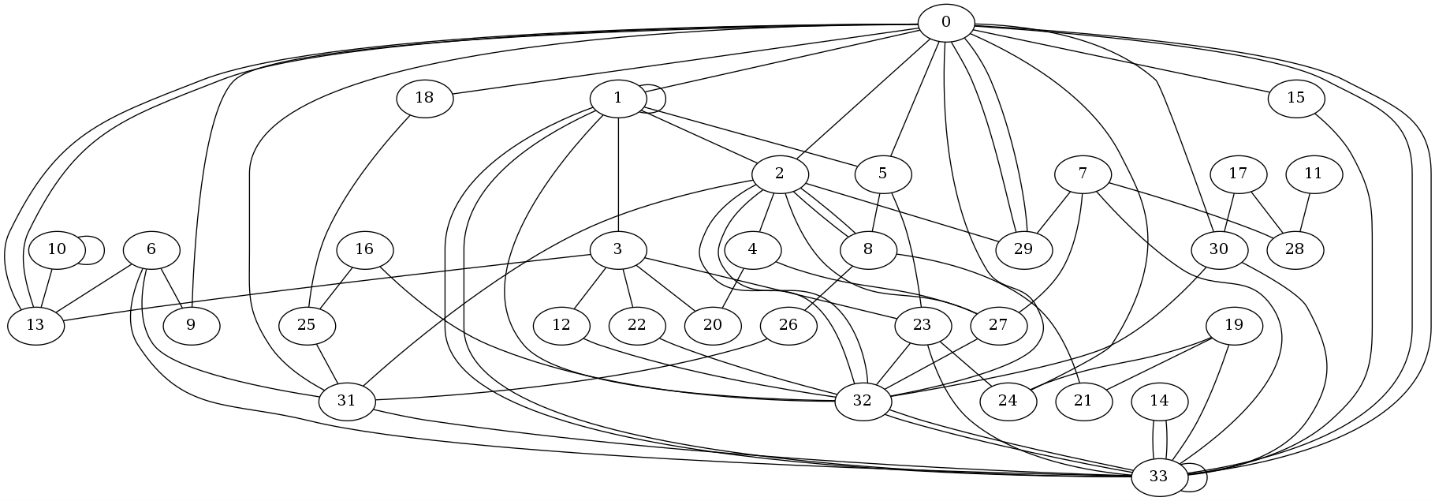


Figure 8 : Graphique généré grâce à un modèle de configuration créé avec les degrés du modèle du karaté club.

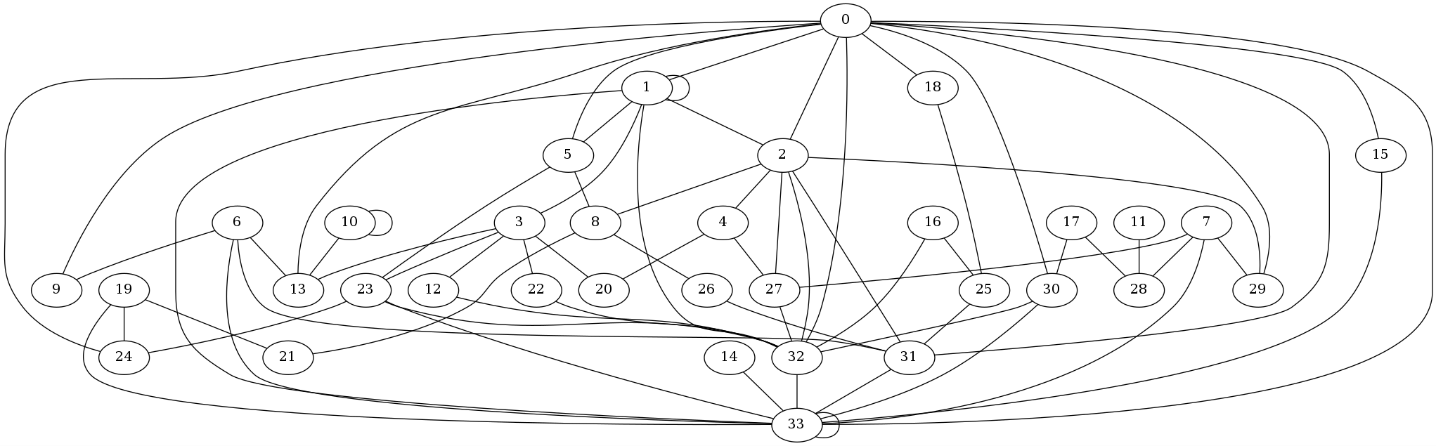


Figure 9 : Le même graphique après élimination des liens multiples

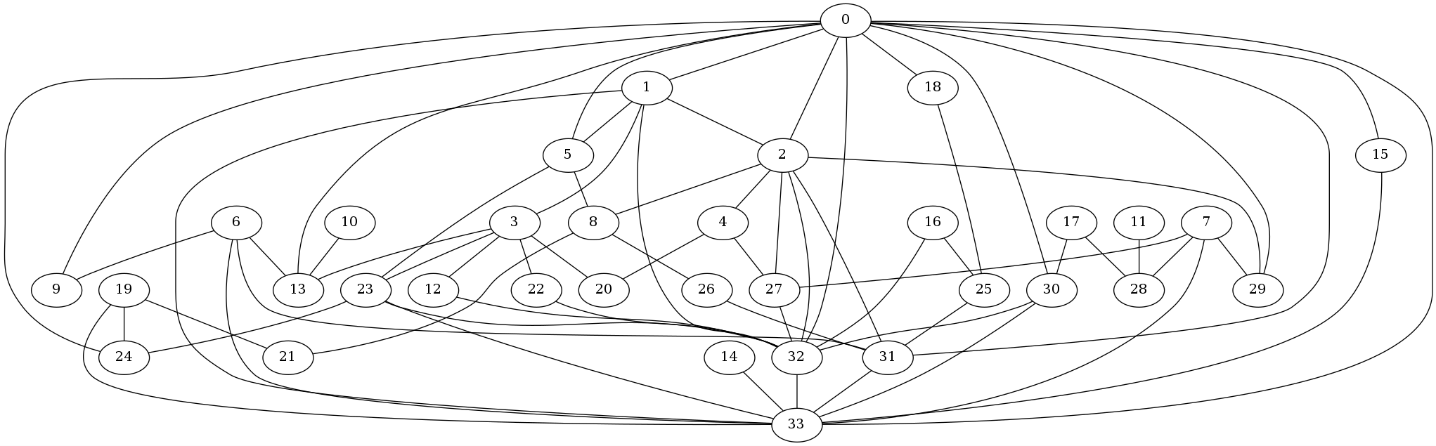


Figure 10 : Le graphique final obtenus après avoir supprimé les boucles

# Point 5 : L’algorithme de Louvain

Cette méthode permet l’extraction de communautés au sein d’un réseau. Nous avons généré 100 graphes aléatoires composé du même nombre de nœuds et de liens que dans le graph du karaté club. Nous avons également fait en sorte qu’il n’y ait pas de boucle, ni des liens multiples dans les graphes aléatoires. On remarque que les graphes ainsi générés ont une modularité comprise entre 0,29 et 0.39. La valeur de modularité du graph du karaté club est-elle égale à 0,42. La modularité est la mesure de la qualité d’un partitionnement donné des nœuds d’un graphe. Les réseaux avec une haute modularité ont des connections plus dense entre les nœuds à l’intérieur d’une même communauté que les nœuds à l’intérieur de communautés différentes.

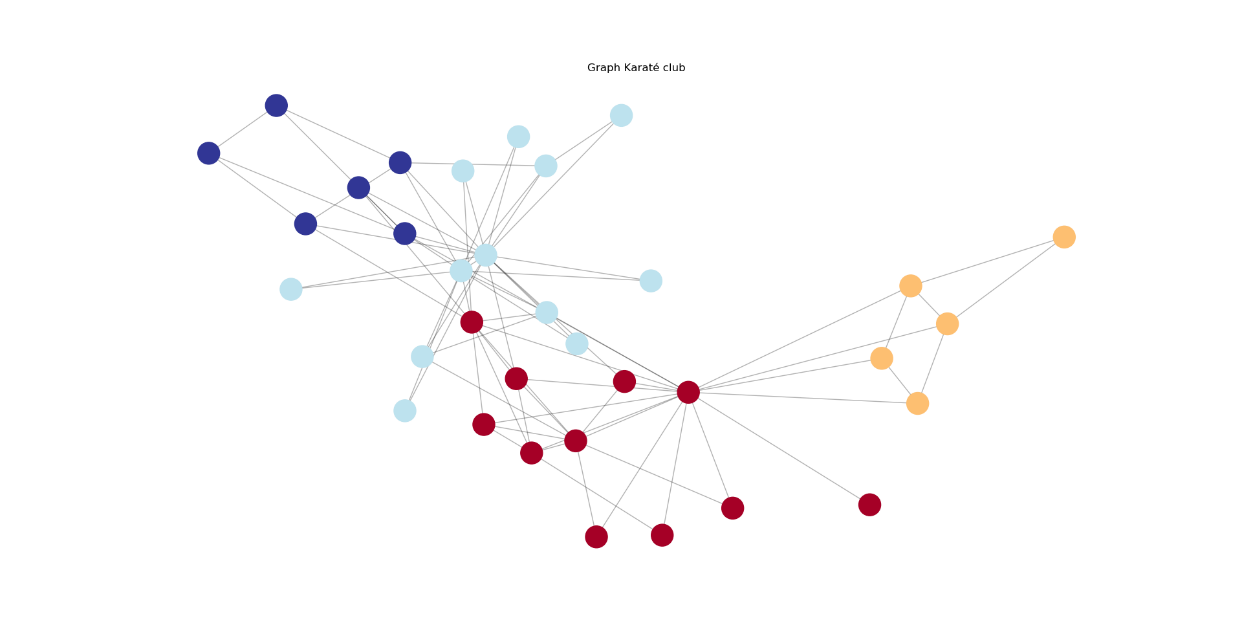


Figure 11 : Représentation des différentes communautés au sein du graph karaté club (modularité : 0,42)

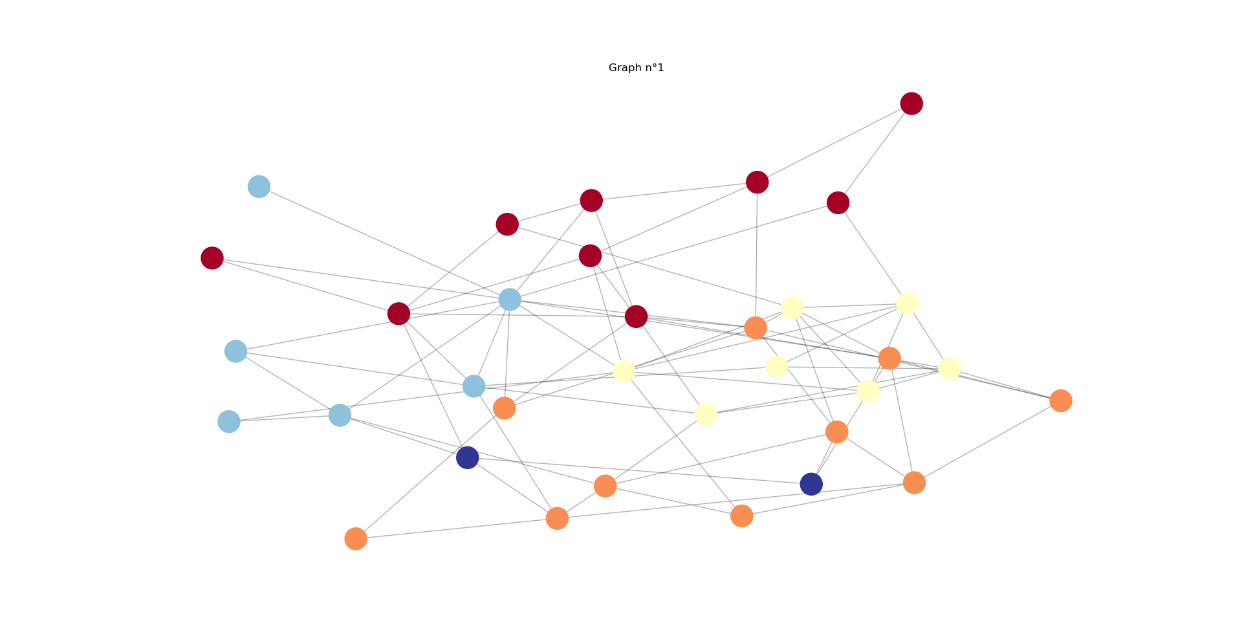


Figure 12 : Représentation des différentes communautés au sein d'un graph aléatoire (modularité : 0,34)

# Point 6 : Ecart-type et clustering moyen

Nous avons calculé le clustering coefficient du graphe karaté club qui est égal à 0,57, celui de sa plus grande communauté vaut 0.76 nous avons ensuite fait la même chose pour 100 graphes aléatoires et en avons fait une moyenne. La valeur moyenne du clustering coefficient pour les 100 graphs est 0.12 et la valeur moyenne pour les plus grandes communautés de ces graphes est 0.18. Ensuite, nous calculons l’écart type de chaque, celui des 100 graphes vaut 0.035 et celui des plus grandes communautés est égal à 0.15.

La première conclusion que nous pouvons en tirer c’est que le clustering moyen des 100 graphes est beaucoup plus faible que celui du graphe karaté club cela est dû au fait que dans le graphe du karaté club certains nœuds sont beaucoup plus connectés que d’autres alors que dans nos 100 graphes aléatoires, les connections entre les nœuds sont assez bien réparties, il n’y a pas de nœud qui est beaucoup plus connecté qu’un autre. On peut dire la même chose pour les plus grandes communautés le fait que les valeurs des plus grandes communautés soient plus grandes que le graphe entier est normal, nous allons chercher une communauté ou cet écart de distributions est plus grand.

À propos de l’écart type, pour les graphes, comme il est très faible, on peut en conclure qu’il n’y a pas de valeur qui sort de l’ordinaire, nous restons toujours assez proche de la moyenne pour chaque graphe. Pour celui des plus grandes communautés on peut voir qu’il est plus élevé cela signifie qu’il y a certaines valeurs qui s’écartent plus de la moyenne.

# Point 7 : Le marcheur aléatoire

Nous créons plusieurs marcheurs aléatoires, un pour chaque nœud (considéré comme le nœud de départ), ces marcheurs parcourent le réseau tant qu’ils n’ont pas parcouru tous les nœuds du réseau. La fonction que nous avons écrit effectue cette opération 2.000 fois et calcul le nombre de pas moyen effectué en fonction du degré du nœud de départ. Nous n’avons pas trouvé de relation entre le nombre de pas effectuer et le degré du nœud initial. Afin de chercher cette relation nous avons diviser les résultats en deux ce qui nous donne les résultats suivants avec donc pour chaque nœud 1.000 marcheurs y ayant commencé.

Voici les box plot correspondant aux deux jeux de données :

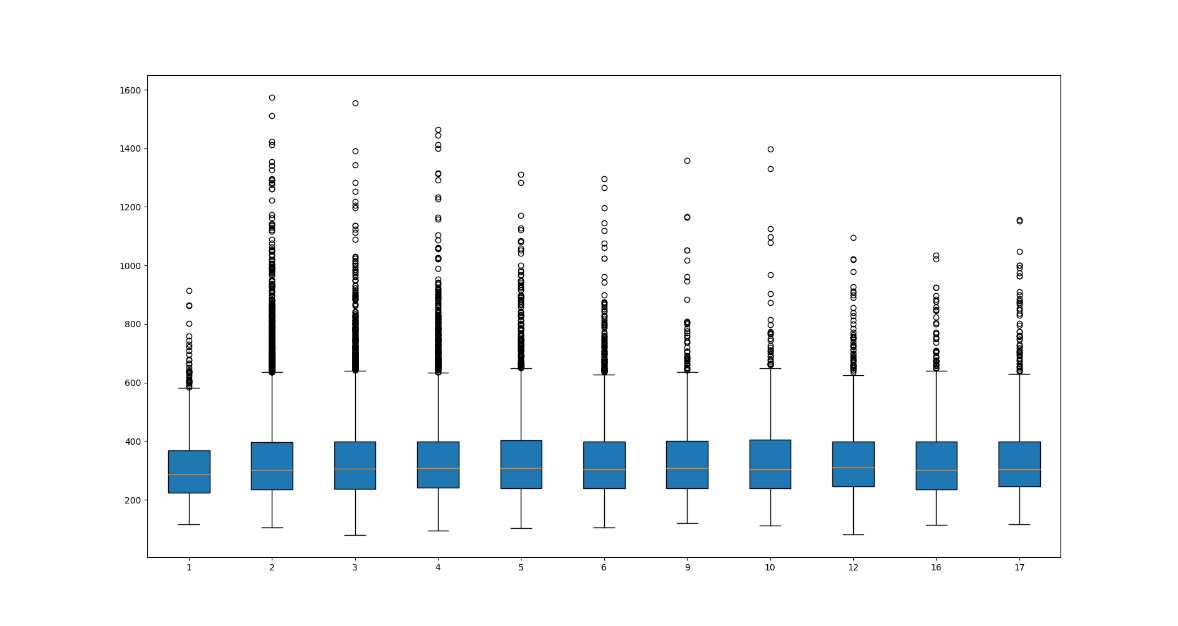


Figure : Box plot des premières 1000 itérations

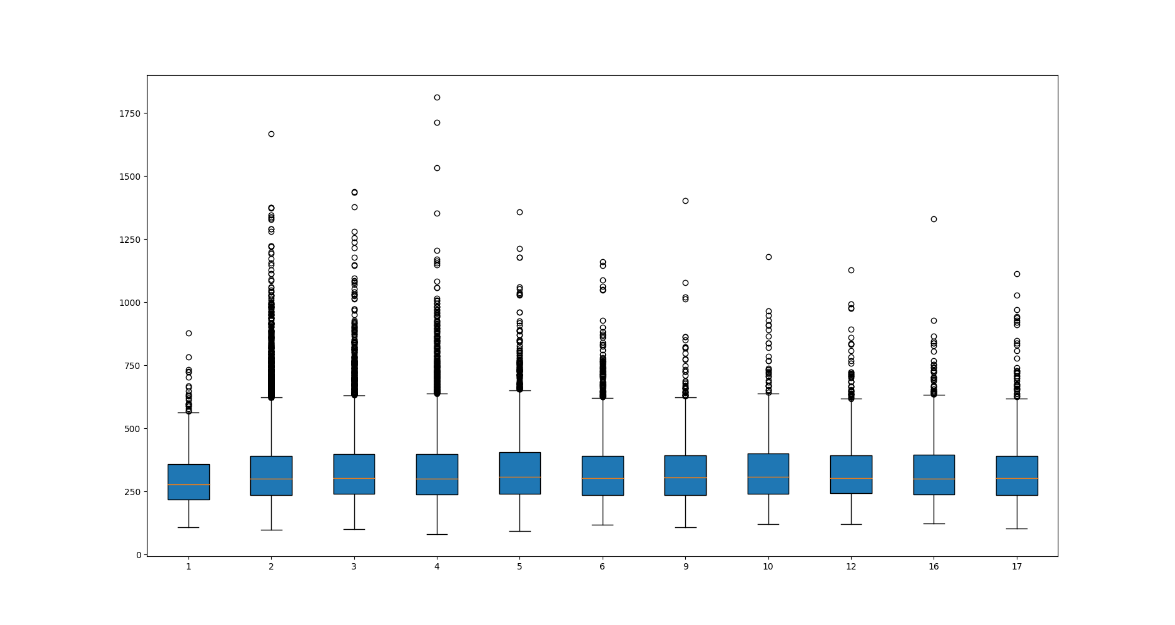


Figure : Box plot des 1000 itérations suivantes

On se rend compte que certains outliers sont très éloignés de la valeur médiane. On peut remarquer que les deux box plot sont similaire. La valeur de la médiane est légèrement inférieure si le nœud de départ est de degré 1. Il est en effet plus difficile pour un marcheur aléatoire de passer par un nœud de faible degré. S’il commence par un nœud de degré un, il n’a plus besoin d’y passer pour terminer son chemin et il a donc plus facile à le terminer.

Sur les graphiques suivants, on se rend bien compte que la médiane du nombre de pas (lorsque le degré du nœud de départ est égal à un) est systématiquement plus basse que les autres. Il n’y a donc pas de relation entre le nombre de pas et le degré du nœud de départ. Sauf si ce dernier est égal à 1 au quel cas le nombre de pas sera plus faible que si le degré du nœud de départ est plus grand que 1.

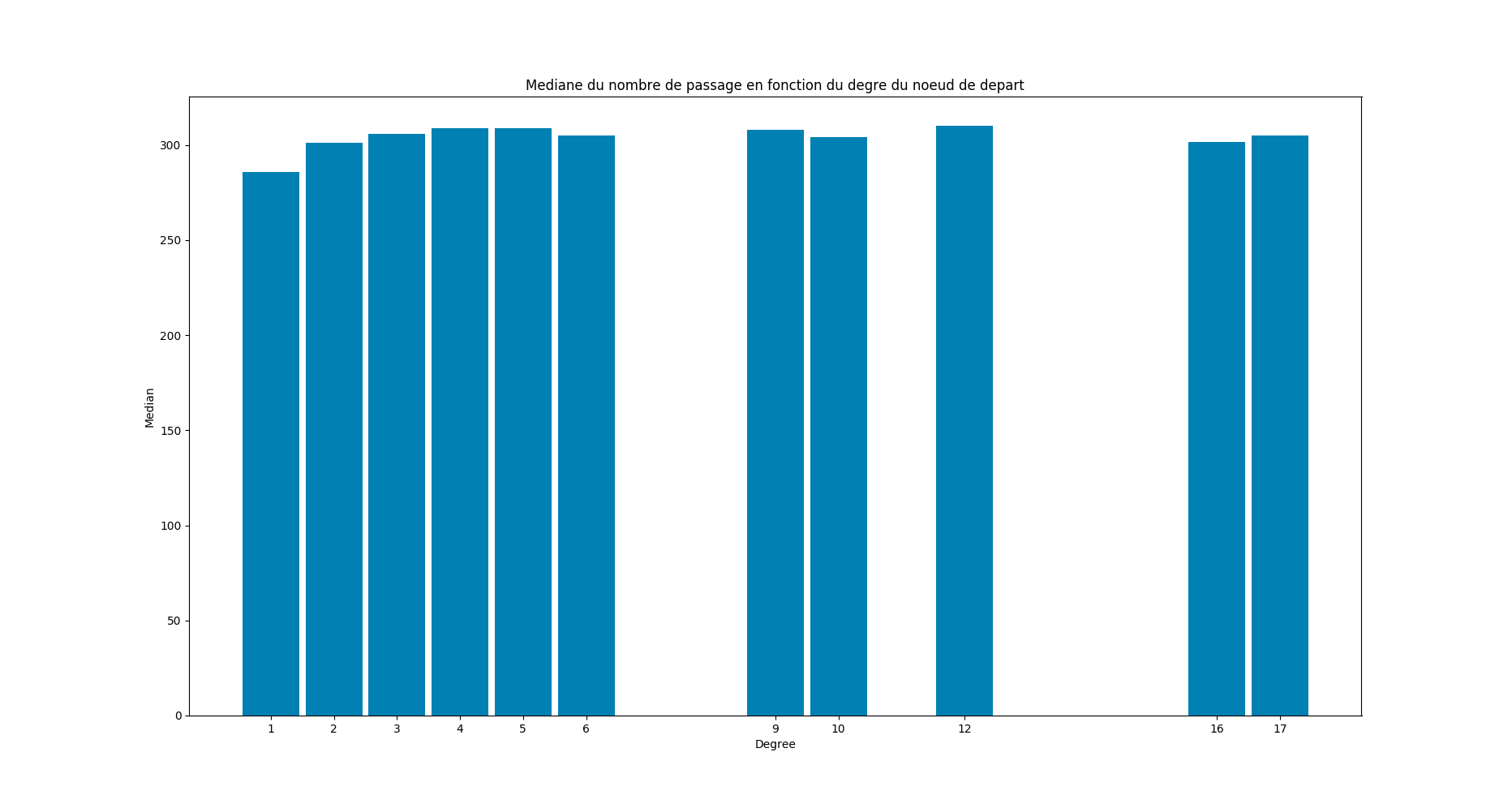


Figure 15 : Graphique représentant les médianes du nombre de pas en fonction du degré du noeud de départ des 1000 premières itérations

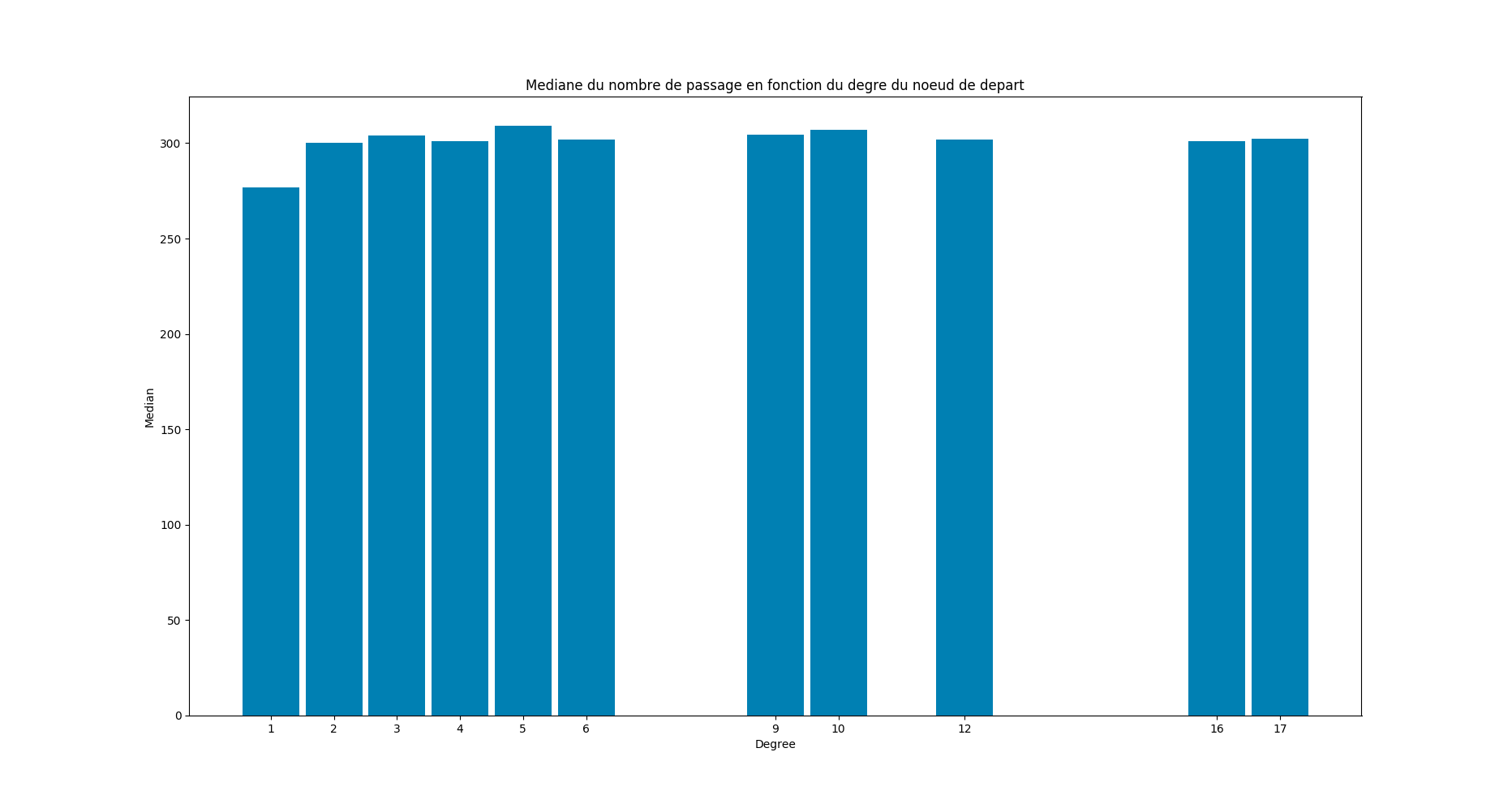


Figure 16 : Graphique représentant les médianes du nombre de pas en fonction du degré du nœud de départ des 1000 itérations suivantes