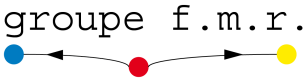
Initiation à Pajek Logiciel pour l'analyse des réseaux sociaux

Laurent Beauguitte

Une courte introduction à Pajek

Laurent BEAUGUITTE - UMR Géographie-cités
beauguittelaurent<at>parisgeo.cnrs.fr

Janvier 2011 - Version 1



Ce livret vise à favoriser une prise en main rapide du logiciel Pajek. Ce dernier a plusieurs avantages : il est gratuit - mais non open source -, tourne sous Windows, Mac (OSX) [1] et Linux, il permet d'étudier des graphes de grande taille (plusieurs milliers de nœuds [2]), il est compatible avec les autres logiciels existants (notamment Ucinet et R). Citons deux autres avantages : un wiki relativement bien fourni (http ://pajek.imfm.si/doku.php ?id=pajek) et de nombreux jeux de données disponibles.

Il présente un inconvénient de taille : il est atrocement mal fichu et sa maîtrise demande quelques pénibles heures d'efforts.

Les commandes sont indiquées en gras dans le texte et répertoriées dans l'index.

Si vous repérez des erreurs ou des approximations, n'hésitez surtout pas à m'en faire part.

1 Principes généraux du logiciel

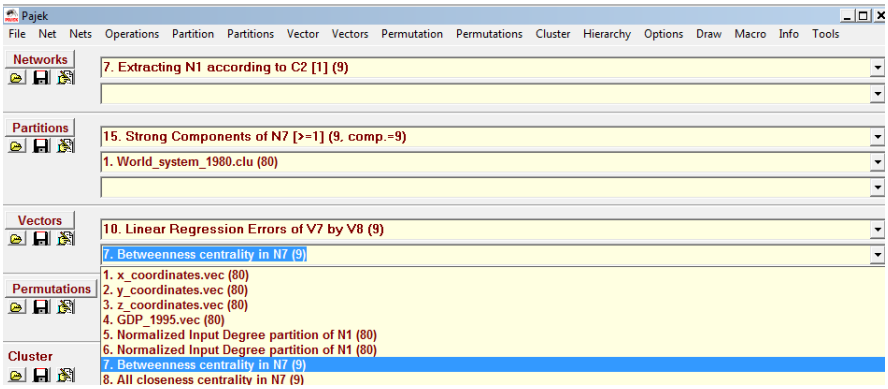
Pajek utilise des menus déroulants à la mode Windows ou Mac. Il convient d'abord de sélectionner les objets à étudier dans la fenêtre principale, puis de dérouler les menus de la barre supérieure pour trouver les commandes appropriées. Les trois principaux objets manipulables avec Pajek sont les réseaux, les partitions et les vecteurs.

Le réseau est une liste de sommets et une liste de liens (éventuellement signés ou valués).

Une partition permet de classer les sommets en différents groupes et, plus généralement, d'attribuer un nombre entier à un sommet. Ainsi, je peux créer une partition codant des villes en 1, 2 et 3 selon leur statut administratif

1. Tutoriel en ligne : http ://vlado.fmf.uni-lj.si/pub/networks/pajek/howto/PajekOSX.pdf
2. Pour les mathématiciens, la taille d'un graphe revoit au nombre de liens, le nombre de sommets désigne l'ordre du graphe.

FIGURE 1 – Aspect du logiciel



(préfecture, sous-préfecture, autre). Si je demande le calcul des degrés, Pajek créera une nouvelle partition nommée degree.

Les vecteurs attribuent à chaque sommet des réels (entiers ou décimaux). Si je fais calculer les degrés normalisés par Pajek, il créera le vecteur correspondant. De manière générale, chaque fois qu'une opération est effectuée, Pajek crée un objet (réseau, partition ou vecteur) qu'il est possible de sauvegarder. La figure 1 montre l'aspect général du logiciel après un petit 1/4 d'heure de travail sur un réseau...

2 Structures des fichiers

Le meilleur moyen pour comprendre comment fonctionne un fichier Pajek est de télécharger l'un des jeux de données disponibles sur le site [3] et de l'ouvrir avec un quelconque éditeur de texte.

Graphe simple

Pajek utilise et produit 3 types de fichiers :
- des fichiers correspondant aux liens (extension .net)
- des fichiers contenant les partitions (extension .clu)
- des fichiers contenant des vecteurs (.vec)

Un seul fichier peut contenir tous ces sous-fichiers, il a alors l'extension .paj. Lorsqu'on débute sur ce logiciel, les erreurs sont fréquentes et il est conseillé de construire un fichier par type (un fichier .net avec les liens, un .clu avec une première partition, etc.).

Une partition assigne à chaque sommet un code numérique entier correspondant (le plus souvent) à une classe. Les vecteurs permettent d'assigner

3. http ://vlado.fmf.uni-lj.si/pub/networks/data/

des valeurs décimales à des sommets.

Deux options existent. La première consiste à modifier ses données afin dans les rendre Pajek-compatibles dans un fichier .txt. Il est également possible de créer directement ses graphes dans Pajek mais la lourdeur de la procédure est telle qu'elle n'est pas détaillée ici [4].

Un graphe est comme chacun sait un ensemble de sommets et un ensemble de liens. La structure minimale d'un fichier utilisable par Pajek contient donc une liste de sommets, et, à la suite, une liste de liens.

Le fichier .txt minimal doit avoir la structure suivante :

```
*Vertices  nombre_de_noeuds
 1 "label" x y  z
 2 "label" x y z
 ...
*Arcs (en cas de graphe orienté)
 numéro_origine numéro_destination valeur
 ...
*Edges (en cas de graphe non orienté)
 numéro_origine numéro_destination valeur
 ...
```

Pour les sommets, les rubriques x, y et z sont optionnelles, elles permettent d'assigner des coordonnées en x (axe horizontal), y (axe vertical) et z (hauteur) comprises entre 0 et 1. Comme dans Netdraw, l'origine des axes xy (0,0) est en haut à gauche de l'écran, les coordonnées (1,1) désignant le coin inférieur droit.

En ce qui concerne arcs et edges, seuls le numéro, l'origine et la destination sont obligatoires. La valeur peut être positive ou négative, entière ou décimale. Dans ce dernier cas, le point (et non la virgule) doit être utilisé comme séparateur. Il est possible de créer des boucles (lien d'un sommet vers lui-même), elles ne sont pas représentées graphiquement mais prises en compte dans les calculs effectués.

Une fois le texte saisi, enregistrez-le et changez l'extension en .net (ou .paj). Si tout va bien, vous pouvez alors l'ouvrir et l'étudier avec Pajek. Attention à bien utiliser des espaces et non des tabulations sinon Pajek refusera obstinément d'ouvrir le fichier. Attention également à ne pas insérer de sauts de ligne dans le fichier source entre la liste des sommets et la liste de liens (et tant pis pour la lisibilité de l'ensemble).

Pour créer ses propres partitions en classe, menu **Partition > Create null partition**. Par défaut, Pajek propose un nombre de nœuds égal à celui du graphe. Répondre oui entraîne la création d'une partition nulle de même ordre que le graphe de départ. Il est alors possible d'ouvrir cette partition et de coder les différents sommets afin de les répartir dans les classes désirées.

4. Les curieux se reporteront à de Nooy *et al.*, 2005, p.22-24.

Les classes sont définies par des valeurs discrètes. Pour créer des propriétés attributaires continues, le principe est le même (1 sommet = 1 valeur) mais l'extension du fichier est .vec.

Dernier point : toutes ces informations (réseau et propriétés attributaires discrètes et continues) peuvent être stockés dans un fichier .paj. Pour l'ouvrir dans Pajek, choisir la commande **File** > **Pajek Project File** > **Read**.

Two-mode network

Pour créer un réseau bipartite, le seul changement concerne la déclaration des sommets. Il convient en effet d'indiquer après le nombre total de sommets combien font partie du premier sous ensemble de sommets. Si jamais il y a une erreur dans la liste des liens, c'est-à-dire si on a déclaré un lien à l'intérieur du même sous ensemble de sommets, Pajek le signale à l'ouverture du fichier (mais l'ouvre malgré tout).

Le mini texte suivant crée un two-mode network entre 2 ensembles de sommets $\{a, b\}$ et $\{1, 2, 3\}$.

```
*Vertices 5 2
1 "a"
2 "b"
3 "1"
4 "2"
5 "3"
*Edges
3  1  1
3  2  1
4  1  1
5  1  1
```

Pour transformer ce réseau en one-mode network, la commande est **Net** > **Transform** > **2-Mode to 1-Mode** et l'on choisit ensuite si on souhaite utiliser les lignes ou les colonnes pour transformer la matrice de départ.

Graphe évolutif

L'un des intérêts de Pajek est de permettre de façon relativement simple de créer des réseaux à différents moments, et de représenter graphiquement leur évolution. Le texte ci-dessous montre les éléments indispensables pour créer un tel fichier, il est suivi d'un exemple montrant un graphe évoluant au fil du temps.

```
*Vertices nombre_de_vertices
1 "label" [x-y] # signale que le sommet 1 est présent du
                  temps x au temps y}
```

```
2 "label" [y]   # le sommet 2 est présent seulement au temps y
3 "label" [x]   # le sommet 3 est présent seulement au temps x
4 "label" [x,z] # le sommet 4 est présent au temps x et
                  au temps z
*Edges
1 3 1 [x]     # il existe un lien entre le sommet 1 et 3
                au temps x
1 2 1 [y]     # il existe un lien entre 1 et 2 au temps y
```

On s'en doute, le risque d'erreur est assez élevé lors de la saisie des données. Pour visualiser l'évolution du script suivant, copiez le dans un fichier .txt puis changer l'extension en .net.

Ouvrez le avec Pajek puis sélectionnez l'option **Net** > **Transform** > **Generate in time** > **All** , indiquez 1 comme "first time point" et 4 comme dernier, 1 comme "step". Pajek produit les 4 moments du graphe. Pour visualiser l'évolution, choisissez **Draw** (ou le raccourci Crtl+G), le layout (algorithme de visualisation) qui vous convient puis **Options** > **Previous**/**Next** > **Apply to** > **Network** et, en cliquant sur **Next**, les étapes devraient apparaître les unes après les autres.

```
*Vertices 4
1 "A" [1-4]
2 "B" [1-4]
3 "C" [1-4]
4 "D" [2,4]
*Edges
1 2 1 [1-4]
2 3 1 [2,4]
2 4 1 [2]
3 4 1 [4]
```

Pour info

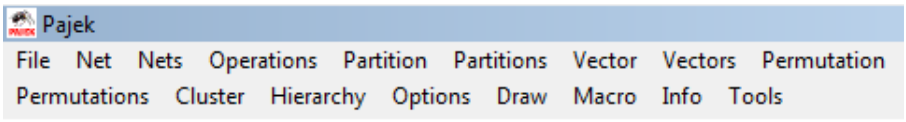
Pajek accepte aussi les données en format matrice en entrée mais, dans la mesure où tout l'intérêt de Pajek est de traiter des graphes comprenant plusieurs centaines ou milliers de nœuds, l'intérêt de manipuler une matrice de cette taille n'apparaît pas évident.

3 Mesures et structures

La ligne de commande du haut de l'écran est particulièrement obscure au premier abord (voir figure 2), quelques repères sont utiles.

Si on souhaite agir sur un réseau (pour l'ouvrir, le transformer...), c'est le menu **Net** qu'il faut dérouler. Si les opérations concernent 2 réseaux (ou

FIGURE 2 – Les menus déroulants



plus), utiliser le menu **Nets**. Pour dessiner : **Draw**. Pour faire des opérations entre différents objets, **Operations**.

Pour connaitre les caractéristiques de base du réseau, menu **Info** > **Network** > **General**. On obtient le nombre de sommets, d'arcs, d'arêtes, de boucles et de liens multiples ainsi que deux mesures de densité (avec et sans boucle).

Chaque fois qu'une manipulation est effectuée, Pajek demande si on souhaite créer un nouveau réseau (**Make new Network ?**). Répondre non entraîne l'écrasement du réseau de départ, il est donc préférable de répondre oui.

Degrés et centralités

Pour obtenir le degré des nœuds, utiliser la commande **Net** > **Partition** > **Degree** puis **Info** > **Partition**. Pajek crée alors une partition correspondant aux degrés et un vecteur pour les degrés normalisés.

Pour connaître le degré moyen, **Partition** > **Make Vector** puis **Info** > **Vector**. Pourquoi faire appel aux partitions et aux vecteurs alors que l'information concerne toujours le degré ? Le degré des nœuds est un entier (donc une partition), le degré moyen est souvent un nombre décimal (donc un vecteur en logique Pajek).

Les mesures de betweenness et de closeness sont accessibles via la commande **Net** > **Vector** > **Centrality**. Il est possible de tester les corrélations (Spearman ou Pearson) entre deux partitions (ou entre deux vecteurs) à l'aide de la commande **Partitions** > **Info**. Les résultats étant donnés sans test de significativité, ils servent uniquement à donner la tendance générale.

Distances

Le diamètre du graphe peut être calculé à l'aide de la commande **Net** > **Paths between 2 vertices** > **Diameter**. [5] Il est sans doute plus judicieux d'utiliser la commande **Net** > **Paths between 2 vertices** > **Distribution of Distances** > **From All Vertices** qui fournit, en plus du diamètre,

5. Si le graphe est non connexe, les données fournis concernent le sous graphe connexe où ces valeurs sont les plus élevées.

la distance moyenne entre sommets. Obtenir la matrice des distances géodésiques entre les sommets du graphe est possible grâce à la commande **Net** > **Paths between 2 vertices** > **Geodesics Matrices*** [6]. Il est également possible de chercher tous les chemins d'une longueur maximale k entre deux sommets avec **Net** > **Paths between 2 vertices** > **Walks with Limited Length**.

Composants

L'information sur le nombre de composant concerne un réseau, elle se détermine donc dans le menu **Net**. Si le graphe est non orienté, les options indiquées ci-dessous donneront le même résultat. Par contre, si le graphe est orienté, le nombre de composants détecté sera généralement plus élevé avec l'option **Weak** qu'avec l'option **Strong**. Les commandes s'obtiennent ainsi :

- **Net** > **Components** > **Strong** puis **Draw** > **Draw - Partition**.
- **Net** > **Components** > **Weak**.

Cycles

La structure globale du graphe concernant la transitivité (si A est lié avec B et C, B et C sont-ils liés ?) - appelée *clustering* par les physiciens - s'obtient avec la commande **Info** > **Network** > **Triadic Census**. Le nombre de triades est fourni par catégorie tout comme le test du chi-2.

k-cores

Pajek est limité pour la recherche des sous-graphes fortement connexes. La recherche des k-cores [7] se fait à l'aide de la commande **Net** > **Partitions** > **Core** > **All**

Blockmodel

Deux algorithmes de partitionnement de graphe basés sur l'équivalence sont disponibles , l'un concerne l'équivalence régulière et l'autre l'équivalence structurale. Les seuls options proposées par défaut concernent le nombre d'itérations souhaitées et le nombre de blocs souhaités (**Operations**> **Blockmodeling*[?]**).

6. Une commande Pajek suivie d'une * signale qu'elle est gourmande en ressources et en temps, il est donc conseillé de la réserver aux graphes ne dépassant pas la centaine de sommets.

7. Un k-core est un ensemble d'au moins 3 sommets qui tous sont voisins d'au moins k autres sommets.

Et quelques mesures supplémentaires

Pour comparer les groupes existants dans un graphe à deux moments différents (ils doivent absolument avoir le même nombre d'acteurs rangés dans le même ordre), sélectionner le graphe dans **Network**, les deux partitions puis **Partition** > **Info** > **Cramer's V, Rajski**.

Pour connaître les effectifs et pourcentages bruts et cumulés des différents indicateurs recherchés (partition ou vecteur), **Info** > **Partition**. La logique des menus déroulants est la même que pour **Net** : pour agir sur une partition, menu **Partition**, pour agir sur plusieurs partitions, menu **Partitions**.

4 Manipulations de grands graphes

Pour y voir plus clair dans un graphe aux sommets multiples, les manipulations suivantes peuvent être utiles :

- extraire d'un graphe un groupe donné
- récupérer les propriétés du groupe extrait
- remplacer les groupes par des sommets

Le jeu de données utilisé en exemple est le fichier World_trade.paj téléchargeable à l'adresse suivante :

http ://vlado.fmf.uni-lj.si/pub/networks/data/esna/metalWT.htm.

Il donne pour l'année 1994 la valeur des produits métalurgiques[8] pour 80 pays[9] . Jeter un œil à sa structure à l'aide d'un éditeur de texte permet de voir qu'il contient les éléments suivants :

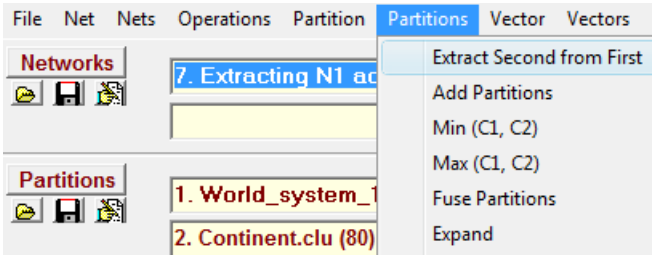
- une liste de 80 sommets ayant des coordonnées x et y
- une longue liste d'arcs valués (en milliers de $US)
- une partition (.clu) donnant la position dans le Système-Monde (centre, semi-périphérie, périphérie, données manquantes codées 999998)
- une partition (.clu) donnant la répartition par continents
- trois vecteurs (.vec) donnant les coordonnées de chaque sommet en x, y et z
- un vecteur (.vec) donnant le PNB en 1995 par sommet

Le graphe contient 1000 liens (**Info** > **Network** > **General**). Une possibilité pour le rendre plus lisible est de déterminer un seuil de représentation. Le seuil se choisit à l'aide de la commande **Net** > **Tranform** > **Remove** > **Lines with Values** > **Lower than**.

Si on souhaite examiner les relations à l'intérieur du continent africain, la première étape consiste à repérer le code associé (1)[10]. La commande **Operations** > **Extract from Network** > **Partition** permet de sélectionner le(s) groupe(s) à extraire. Mais les données attributaires associées doivent

8. *Miscellaneous manufactures of metal* dans le texte original.
9. Ce chapitre synthétise le chapitre 2 de Nooy *et al.*, p.29-57.
10. Sélectionner la partition continent et l'éditer. Algeria est codée 1.

FIGURE 3 – Extraire un groupe et les données associées



être extraites elles aussi. Il convient donc de sélectionner la partition Continent et la partition World System puis d'utiliser la commande **Partitions > Extract Second from First** (voir 3. Il est alors possible de représenter les relations à l'intérieur du continent africain en colorant les sommets en fonction de leur position dans le Système Monde.

On peut aussi souhaiter étudier les relations entre les pays africains et les autres continents. Dans ce cas, il suffit d'utiliser la commande **Operations > Shrink Network > Partition** et de sélectionner quel groupe ne sera pas agrégé.

5 Visualisations

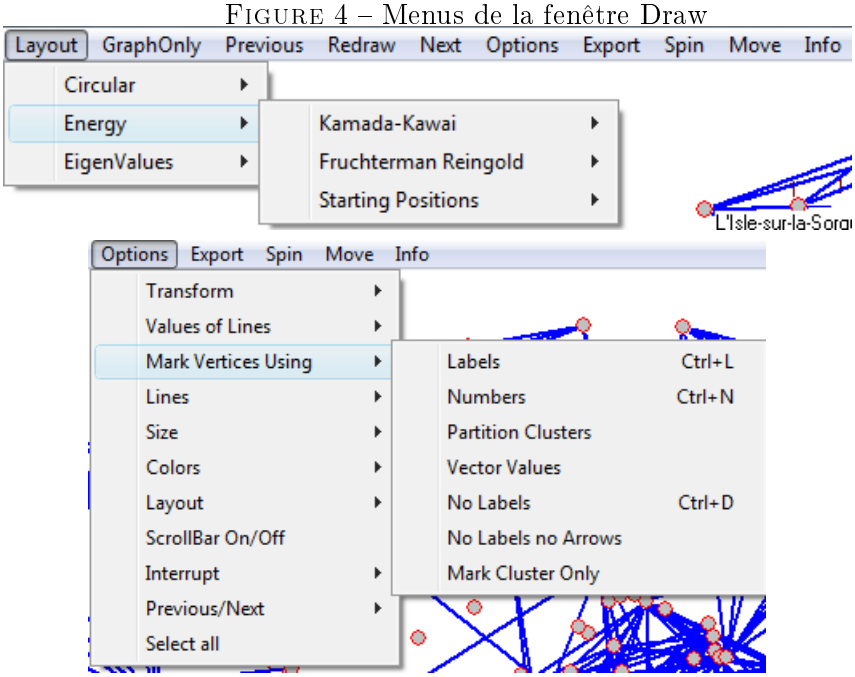
La visualisation des graphes, des partitions et/ou des vecteurs se fait à l'aide du menu déroulant **Draw**. Les commandes disponibles dans la fenêtre **Draw** et les menus déroulants concernant les options et les algorithmes de visualisation sont reproduits dans la figure 4.

Il est possible de zoomer sur une partie du graphe : clic droit et dessiner la zone à agrandir. Pour déplacer tous les acteurs d'un groupe : placer la souris à proximité d'un acteur du groupe, clic gauche maintenu appuyé et déplacer. Pour revenir à l'image de départ, **Redraw**. Les images peuvent être exportées en format .svg .ps .eps ou bitmap (**Options > Export**).

Pour faire varier taille et/ou couleur des sommets en fonction de leur(s) propriété(s) (degré, betweenness), il convient de sélectionner le graphe, la partition et le vecteur souhaité puis de choisir l'une des options du menu **Draw**.

Si les liens sont valués, leur représentation peut se faire de trois manières différentes dans la fenêtre **Draw** : indiquer la valeur du lien (**Options > Lines > Mark Lines > With Values**), faire varier l'épaisseur des traits en fonction de l'intensité du lien (**Options > Lines > Different Widths**) et enfin garder la même épaisseur mais avec un dégradé du blanc au gris foncé (**Options > Lines > GreyScale**).

FIGURE 4 – Menus de la fenêtre Draw



Il est enfin possible de visualiser la matrice d'adjacence du graphe étudié, dans sa forme originale ou dans la forme bloquée [11]. Les commandes respectives sont **File > Network > Export Matrix to EPS > Original** et **File > Network > Export Matrix to EPS > Using Permutations**. Si une partition a été effectuée (blockmodel par exemple), l'*image matrix* correspondante peut également être visualisée. Un article des auteurs disponible en ligne permet de compléter ces quelques indications[**?**].

Atouts et limites

Pajek est un logiciel plutôt rapide et, une fois la logique partition - vecteur comprise, son utilisation est plutôt aisée. Ses limites principales sont au nombre de trois ; le côté clicodrome de l'ensemble est parfait pour donner un cours d'initiation mais se révèle un cauchemar lorsqu'on souhaite appliquer les mêmes analyses à plusieurs matrices [12].

Les algorithmes de visualisation disponibles sont peu nombreux. Enfin, que ce soit en terme de mesures ou de recherche de structure, le choix est limité, excepté en ce qui concerne les mesures de distance.

11. Une matrice bloquée permute lignes et colonnes afin d'obtenir des sous graphes homogènes.

12. Ajoutons qu'il est paradoxal que, pour un logiciel gratuit, le seul mode d'emploi digne de ce nom soit payant (30 euros couverture souple, 62 euros couverture rigide !). Le pseudo manuel librement téléchargeable est d'une utilité relative[**?**].

Au final, un outil pratique pour les graphes de grande taille mais qui demande une longue mise en forme des données [13] et ne permet que des analyses standards. Pour des analyses plus poussées, passer aux différents packages de R [14] est sans doute le choix le plus judicieux.

13. Apprendre à utiliser la fonction RECHERCHERV disponible dans les tableurs permet un recodage relativement rapide des données.

14. Citons notamment 'statnet', 'igraph' ou 'blockmodeling'.

Index

Table des figures

Table des matières