Document de synthèse : Travail sur la base de données – Relations interpersonnelles

Ce document vise à résumer le processus de de création des matrices grâce aux programmes, les quelques questionnements et problèmes rencontrés et de présenter les observations effectuées.

* Matrice formation

Les informations à disposition étant sous une forme mixte (textes divers et dates), cela rend très difficile l’automatisation de la création de la matrice. C’est pourquoi un premier traitement a été effectué pour compiler les références par formation et par promotion. Cela consistait tout d’abord à établir la liste de toutes les formations apparaissant dans les descriptions en étudiant chaque cellule une à une. Puis de filtrer par mots-clés (noms des formations tel que « Polytechnique » pour relever la référence de chaque individu par établissement et enfin de les classer par promotions dans des tableaux. Il s’agit par la suite de coder un programme permettant d’indiquer « 1 » dans les cellules dont les références correspondent aux individus issus d’une même promotion/formation en utilisant une boucle qui, pour chaque cellule de la plage en question, va indiquer « 1 » dans les cellules dont la référence en ligne correspond à celle de la cellule sélectionnée et en colonne à chacune des autres cellules de la plage, et ce jusqu’à la fin de la plage sélectionné. Il faut donc changer la plage concernée dans le programme pour chaque formation/promotion.

Ensuite, la diagonale qui correspond aux cellules (i,i) doit être codée par « 0 » car un individu ne peut pas être en relation avec lui-même. Toutes les cellules vides restantes sont aussi codées par « 0 » grâce à un programme indiquant 0 si la cellule est vide.

Problèmes rencontrés : La définition de « réseau » s’avère complexe et a des implications pour notre traitement des données. On peut considérer que l’on note « 1 », donc que l’on établit une relation entre deux individus dans la matrice, lorsque l’on considère que deux individus ont interagi entre eux. Cela suppose de déterminer dans quels cas on considère qu’il a pu y avoir interaction. Le premier cas a donc été de restreindre aux individus ayant appartenu à une même promotion. Une deuxième définition plus abstraite serait de considérer que deux individus sont liés à travers le passage dans une même institution. Dans ce cas, on établirait une relation entre tous les individus ayant étudié dans un même établissement, indépendamment de l’année de présence. Ainsi, prenant en compte ces considérations, Mr.Deguilhem a proposé de différencier les cas en créant deux matrices distinctes : la première lie les individus par promotions, la seconde de même mais lie aussi les individus appartenant à un même établissement pour lequel nous n’avons pas de date indiquée. Il est raisonnable de penser que des individus d’un même établissement peuvent avoir des liens à travers le réseau de l’école, des systèmes de mentorat, en passant par des individus intermédiaires etc. Il est vrai aussi que cela généralise en quelque sorte la notion de lien car il est plus probable pour deux individus ayant étudié dans la même promotion d’avoir un lien que pour deux individus d’une même école mais d’années différentes.

Cas Particuliers : Les cas particuliers tels que les individus pour lesquels aucun établissement et/ou année n’est précisée n’ont pas été traités (« 0 »). En effet lorsque l’indication est par exemple « licencié en droit », il n’est pas possible d’étendre notre définition pour considérer que tous les individus licenciés en droit ont un lien. Aussi, d’autres cas tels que l’absence partielle de date pour certains individus issus d’une formation ont été traités au cas par cas de la manière suivante pour la matrice étendue :

* lorsque le nombre d’individus n’ayant pas de dates est minoritaire parmi tous ceux d’une même formation, aucune relation n’a été établie pour ces individus dans le but de réduire les erreurs potentielles
* dans le cas contraire (majoritaire) des liens sont établis avec tous les autres (extension)
* dans le cas de l’ENSAE, j’ai observé les dates des autres catégories et remarqué que deux individus présentaient des dates très similaires pour différentes catégories donc il pourrait être raisonnable d’établir une relation entre eux. Les autres individus semblent faire partie d’une même génération car leur date d’entrée à l’IGF appartient à la décennie suivant 2000. Il a donc été décidé d’établir une relation entre ces individus suivant le même raisonnement que précédemment.

Ainsi, on obtient deux matrices formation : la première, très rigoureuse, n’établit des liens qu’entre individus d’une même promotion. La seconde, moins rigoureuse, s’appuie sur l’hypothèse que pour les individus n’ayant pas de date, nous pouvons poser l’hypothèse d’existence de liens entre ces derniers à travers des dispositifs de mentorat ou de réseau propre à l’école par exemple. Il s’agit de la première matrice à laquelle on rajoute le traitement des individus ne présentant pas de date.

* Matrice IGF

Les données étant seulement sous la forme de date ou parfois manquantes (individus n’ayant pas encore atteint le rang d’inspecteur général des finances ou n’ayant jamais atteint ce grade), il était possible d’automatiser entièrement la création de la matrice. Les cases vides ont été marquées comme « NA ». Le programme consiste à retenir une cellule de référence et de rechercher toutes celles dont le contenu est identique à celle de la référence, puis de répéter cela pour toutes les cellules en tant que référence. Après avoir détecté qu’une cellule avait le même contenu, on indique « 1 » dans la matrice grâce à la variable de comptage qui correspond à la référence de la cellule à laquelle on se réfère et à la seconde variable qui représente la référence de la cellule étudiée.

* Matrice Entrée à l’IGF

Il s’agit de procéder de la même manière que pour la matrice IGF mais cette fois-ci, aucune information n’est manquante.

* Agrégation des matrices précédentes

Ici, deux matrices ont été créées : l’une avec l’agrégation des matrices dont celle des formations considérées avec les promotions et la deuxième avec celle des formations élargie. La somme est effectuée pour toutes les cellules ayant les mêmes coordonnées (i,j) de toutes les feuilles concernées.

* Statistiques/Observations

Le programme utilisé consiste à stocker dans des variables le nombre de « 1 », « 2 » et « 3 » si la valeur est trouvée dans une cellule et ce, pour toutes les cellules.

On observe un très faible nombre de relations pour les formations par promotion et l’obtention du grade IGF avec moins de 1% (respectivement 0,49% et 0,641%). Pour l’entrée à l’IGF, nous avons 1,097%. Pour la formation étendue, nous avons 18,83%. Ce nombre élevé s’explique par le fait que pour des formations telles que l’IEP de Paris par laquelle beaucoup sont passés, tous les individus ont été considérés comme affiliés les uns aux autres. Enfin, pour les matrices des sommes, on observe extrêmement peu d’individus ayant des liens sur les trois aspects de leur carrière (formation, entrée à l’IGF et obtention du grade IGF), peu d’individus ayant des liens sur deux aspects et 1,21% d’individus ayant un aspect en commun avec un autre par promotion contre 18.83% pour la matrice somme formation étendue. On peut en conclure que les inspecteurs des finances ont peu de liens interpersonnels entre eux d’après la définition et l’hypothèse que nous avons posée. Les parcours s’avèrent tout de même assez variés (97 formations différentes relevées) avec certains présentant des parcours assez particuliers même si la plupart sont passés par l’ENA, et Sciences Po Paris.

* Vérification des matrices

La vérification des matrices a été effectuée en prenant des cellules au hasard contenant « 1 » pour ensuite rechercher la date ou information en question chez les deux individus concernés (individus en ligne et en colonne) et ainsi vérifier qu’ils coïncident.

Elaboration du réseau – Relations interpersonnelles

Nous avons utilisé les packages principaux *igraph*, *network*, *networkD3* pour la modélisation des réseaux et *knitr*, *markdown* et *rmarkdown* pour pouvoir créer un fichier html affichant le programme et les visualisations.

* Etapes

Après conversion des données csv sous la forme matricielle, on initialise le dataframe dans igraph puis on passe de igraph à networkD3 qui va nous permettre d’obtenir une visualisation dynamique. On distingue les « links » des « nodes », respectivement les liens entre individus et les individus dans notre cas. Il faut définir les différentes variables qui constitueront les paramètres avant d’utiliser la fonction forceNetwork. L’utilisation de « betweeness » à la place de « degree » permet de mettre en valeur les acteurs/intermédiaires les plus influents en prenant en compte le ratio degrés/degrés totaux.

* Décisions (IMPORTANT)

Cela résulte en six réseaux différents : Entrée IGF, année d’obtention du grade IGF, Formation, Formation étendue, SommeFormation, SommeFormation étendue. Suite à la délibération avec M.Deguilhem, M.Bédu et Mme.Granier, nous avons décidé de retenir les matrices Formation et SommeFormation plutôt que la version étendue pour des raisons de rigueur et de lisibilité. Cependant, ce choix implique une moindre prise en compte des potentiels liens entre individus étant passés par l’IEP de Paris ou l’ELSP, les principales institutions par lesquelles passent les inspecteurs des finances. Cela pose un problème conséquent mais nous ne disposons pas des données pour pouvoir traiter de manière adéquate ces formations (absence de dates donc impossibilité d’établir des liens entre promotions). A l’inverse, le fait de retenir les versions étendues risque de relier de nombreux individus sans aucun lien entre eux donc d’obscurcir les informations qui nous intéressent, à savoir la forme du réseau et le poids des différents acteurs dans ce dernier. Il est difficile de distinguer les individus comptant beaucoup de relations en raison de l’hypothèse posée de ceux qui sont réellement d’importants intermédiaires.

* Observations

Dans la matrice Somme finale, on remarque de très importants intermédiaires tels que Jacques Maurice COUVE DE MURVILLE (premier ministre sous De Gaulle) puis Jean Georges Marie AUTIN (inspecteur gén de la France d’Outre-mer, Ministère de la Culture, président de TéléDiffusion de France) mais aussi d’autres intermédiaires importants proches tels que Pierre Rodolphe dit François MORETTE BOURNY (?), Alain Marie André PRATE (Banque de France, Caisse Nationale de Prévoyance), Christian SAUTTER (Secrétaire général adjoint de l’Elysée /s F.Mitterrand, Secrétaire d’Etat chargé du Budget /s J.Chirac puis Ministre de l’Economie, des Finances et de l’Industrie), etc. (voir graphique réseau pour plus de détails).

Dans le graphique étendu, les intermédiaires les plus importants sont sûrement des individus ayant étudié à l’IEP de Paris (puisque nous avons établi un lien entre tous les individus issus de cet établissement). On peut y relever ROGER-MACHART Charles Marie Jacques (député), Pierre Samuel René LEDOUX (« réalise notamment le chantier de la fusion de la BNCI et du CNEP, qui donne naissance à la BNP » *source BNP,* directeur générale de la BNP), par exemple. On retrouve aussi PRATE qui était un intermédiaire important dans la matrice stricte mais aussi dans la matrice étendue.

Autres individus à noter : LEONARD DE JUVIGNY, SAGNE, LALLEMENT, ROBERT, LE PORTZ

Travail préparatoire sur le programme – Relations inter-organisationnelles

* Introduction

Le but ici sera d’établir des liens entre les organisations lorsqu’un individu y est passé. A titre d’exemple, l’individu A passe de l’organisation A à B donc un lien est établi entre l’organisation A et l’organisation B. La base de données initiale consiste en une liste des carrières de chaque individu et une liste de toutes les organisations présentes dans la base de données avec leurs correspondances lorsqu’il y a une nécessité d’agréger. La préparation de cette base de données étant encore en cours, j’ai utilisé la base de données initiale pour tester des programmes sur les liens entre établissements de formation.

* Elaboration du programme et difficultés rencontrées

Tout d’abord, il s’agissait de chercher une fonction qui pourrait nous permettre de déterminer la présence d’une chaîne de caractères précise. Les principales fonctions considérées ont été : InStr, .Find, Filter/AutoFilter. Les tentatives utilisant InStr m’ont confrontée à la nécessité de définir précisément la chaîne dans laquelle chercher et cette méthode ne me paraissait pas optimale. Ensuite, les tentatives utilisant Filter/AutoFilter présentaient une « incompatibilité de type » sans indiquer quel paramètre posait un problème. J’en avais déduit que cela concernait la définition du IfElse car elle utilisait la syntaxe <>Nothing, or l’output de la fonction Filter n’était pas claire.

Enfin, lors de l’utilisation de .Find, il me semblait plus probable d’élaborer un programme efficace en utilisant cette fonction, l’unique problème concernant la définition du terme à rechercher au format ‘’String’’. M.Deguilhem m’ayant indiqué l’existence d’une commande permettant de transformer des données en ‘’String’’ sur R, j’ai pu trouver la commande CStr(). Les autres problèmes rencontrés dans ce programme concernaient des soucis de logique d’emboîtement/articulation des différentes boucles et des variables de comptage tel que le décalage entre la recherche de x et y dont les valeurs ne devaient pas coïncider sous peine d’avoir des doublons dans la matrice finale.

Le programme consiste à stocker le nom de l’organisation A extrait de la liste des organisations puis de rechercher sa présence sur toute la ligne de chaque individu. Si la recherche ne retourne pas Nothing, alors on va rechercher à tour de rôle la présence du nom de l’organisation B dans la ligne, puis C, D etc. Si pour une de ces organisations la recherche aboutit, on rajoute 1 dans la cellule de la matrice correspondant à l’intersection entre ces deux organisations. La boucle s’effectue pour chaque i (individu), chaque x (organisation 1), et chaque y (organisation 2 tel que y = x+1).

* Résultats

Sans grande surprise, les établissements les plus connectés sont l’IEP de Paris et l’Ena car ce sont non seulement les établissements comptant le plus de passage mais les individus y passant ont souvent effectué au moins une autre formation. On peut noter aussi un certain rôle de l’école des HEC et de l’Ecole Polytechnique dans le réseau.

Omissions et problèmes soulevés par la méthode utilisée

* Limites

En raison des données dont nous disposons pour la constitution des réseaux, il semble clair qu’une partie des relations interpersonnelles risque d’être omise lorsque nous choisissons de garder la matrice stricte des formations. Néanmoins, sa version étendue risque d’inclure de nombreuses relations interpersonnelles qui n’ont pas eu lieu : la généralisation est trop forte et s’éloignerait de la réalité. Ainsi, le choix effectué reste pertinent pour l’étude au vu des contraintes imposées par la collecte de données. Nous ne pouvons pas envisager d’interroger tous les inspecteurs des finances pour obtenir une liste exhaustive de leurs relations interpersonnelles, ou même ne serait-ce les informations manquantes touchant à leur parcours en raison du statut et de la temporalité.

Le document sur lequel nous nous appuyons pour y extraire les informations recense les données sur la période 1801-2009. Un biais de temporalité se présente car les informations sont plus complètes pour les périodes antérieures. En effet, nous disposons des informations portant sur l’ensemble de la carrière de ces derniers alors que plus l’on avance dans le temps, plus l’on trouve des individus ayant rejoint l’inspection récemment donc à un point moins avancé de leur carrière. Ce biais de temporalité est un point à ne pas oublier lors du traitement et de l’analyse des données par la suite.

Il faut garder à l’esprit que la méthode de l’interlocking utilisée ici ne permet pas de rendre compte de la différence entre lien fort et lien faible au sens où l’entend la sociologie. Dans une moindre mesure, la matrice des sommes pourrait rendre compte d’un certain degré de relation lorsque deux individus présentent à la fois une formation, une année d’entrée et une année d’obtention du grade d’IGF identiques. Il est fort probable que deux individus s’étant rencontrés à plusieurs reprises à différents stades de leur parcours se connaissent et entretiennent une relation sortant du contexte professionnel/formel. Dans son article « What Do Interlocks Do ? An Analysis, Critique, and Assessment of Research on Interlocking Directorates » paru dans l’*Annual Review of Sociology,* Mizruchi souligne l’importance d’expliciter les implications de la méthode de l’interlocking et de définir précisément les hypothèses qui ont été posées car ces dernières peuvent avoir d’importantes conséquences sur les résultats obtenus et leur interprétation. Ainsi, nous avons tenté de rester rigoureux à ce sujet en explicitant les problèmes posés par nos hypothèses et les limites de la méthode utilisée vis-à-vis du sujet.