

DE LA THEORIE A LA PRATIQUE : QUELQUES ENSEIGNEMENTS DE LA MISE EN ŒUVRE DE LA METHODE « NETWORK SAMPLING WITH MEMORY » POUR ENQUETER AUPRES DES IMMIGRES CHINOIS EN ILE-DE-FRANCE

Aurélie SANTOS, Géraldine CHARRANCE, Marine BONDON, Paul COCHET,
Thomas MERLY-ALPA

Ined, Service des enquêtes et sondages

aurelie.santos@ined.fr
geraldine.charrance@ined.fr
marine.bondon@ined.fr
paul.cochet@ined.fr
thomas.merly-alpa@ined.fr

Mots-clés : Échantillonnage, collecte, population difficile à joindre

Domaine concerné : « Collecte de données d'enquêtes » (4), « Protocole, conception des enquêtes, couverture de populations particulières » (4.1)

Résumé

La méthode d'échantillonnage « Network sampling with memory » (NSM) constitue une nouvelle variante de sondage par chaînage (boule de neige, Respondent Driven Sampling (RDS), etc.). Ces méthodes consistent à enquêter au sein de réseaux en sélectionnant au départ quelques individus appelés « graines ». Seuls ces individus sont désignés par le sondeur. Par la suite, ce sont les enquêtés eux-mêmes qui recrutent/désignent leurs pairs qui seront sollicités à leur tour pour participer à l'enquête. Les méthodes par chaînage permettent en théorie d'atteindre des pans de la population non directement accessibles à des enquêteur-riche-s, mais souffrent d'un biais de sélection très fort. Dans le but de pallier ce défaut, une équipe de l'université de Caroline du Nord a développé la méthode NSM qui présente des avantages théoriques notamment sur la précision des estimations, au prix d'une complexité plus importante. La particularité de NSM est de recréer, au fur et à mesure du terrain, une base de sondage de la population cible composée des personnes citées par les répondants et de tirer aléatoirement les futur-e-s enquêté-e-s dans cette base. Contrairement à la méthode RDS, elle ne cherche pas à enquêter tous les contacts cités mais vise à intégrer une dimension aléatoire dans une méthode de sondage empirique. L'algorithme comprend une première phase exploratoire dite « Search », à la recherche de nouveaux pans du réseau, puis une seconde phase de tirages aléatoires au sein du réseau dévoilé. Selon ses concepteurs, la méthode NSM devrait permettre d'obtenir des estimations d'une précision équivalente à celle d'un sondage aléatoire simple.

Après deux expériences à l'étranger (en Tanzanie et en Caroline du Nord), la méthode a été utilisée pour la première fois en France entre septembre 2020 et juin 2021 dans le cadre d'une enquête menée auprès des immigrés chinois en Ile-de-France (CHIPRe). À l'issue du terrain, 500 questionnaires et quelque 1700 citations ont été collectés. Cette enquête s'est avérée extrêmement délicate à mettre en œuvre. Sa dimension aléatoire, et notamment la nécessité de réaliser des tirages hebdomadaires, a engendré de nombreuses difficultés sur le terrain. La charge de travail individuelle des enquêteur-riche-s s'est révélée difficile à anticiper, puisqu'une forme de filiation citant/cités (conserver le-la même enquêteur-riche pour un-e enquêté-e et les personnes citées dans son « roster » puis échantillonnées) a été privilégiée afin de faciliter les négociations des enquêteur-riche-s sur le terrain. De plus, un paradoxe inhérent à la logique de l'algorithme de tirage dans sa phase « exploratoire » conduisait les enquêteur-riche-s qui collectaient le plus de citations (les plus grands « rosters ») à se voir attribuer parfois moins d'échantillonnés que d'autres enquêteur-riche-s moins performants. En effet, les petits rosters (contenant un petit nombre de citations) étaient plus exposés à l'échantillonnage car interprétés par l'algorithme comme une opportunité d'aller vers des pans inexplorés du réseau, à l'inverse des plus grands rosters, plus souvent délaissés par l'algorithme car associés à la découverte de personnes citées plusieurs fois (doublons), synonymes d'un niveau d'exploration avancé du réseau. Ainsi les « bon-ne-s » enquêteur-riche-s ne voyaient pas toujours leurs efforts récompensés par le tirage des rosters qu'il-elle-s avaient collectés. Cette impossibilité d'anticiper les tirages s'est aussi concrétisée dans le travail de communication auprès des enquêtés, car les enquêteur-riche-s étaient incapables de prédire dans leurs argumentaires si les personnes citées par leurs enquêté-e-s seraient finalement tirées au sort ou non, ni à quelle échéance.

À ces difficultés se sont ajoutées celles de la pandémie. La crise sanitaire et la suspension de la collecte en face-à-face induite par le second confinement ont eu des impacts conséquents sur la collecte. Initialement pensée pour être conduite en face-à-face, nous avons été contraints de mener au moins une partie de l'enquête par téléphone. Les méthodes d'échantillonnage par chaînage reposant sur un lien de confiance entre enquêteur-riche et enquêté-e et contribuant à une réelle implication des enquêté-e-s dans le processus de recrutement, la prise de contact et la passation des questionnaires à distance (par téléphone ou en visio) a rendu la tâche plus ardue. Le distanciel a également rendu plus complexe et moins efficace la gestion des incitations financières (remise de chèques cadeaux).

Abstract

Network Sampling with Memory (NSM) is an innovative method in the field of respondent driving sampling (RDS) methods due to the implementation of a particular sampling algorithm. Among other things, this algorithm makes it possible to create a sampling frame from which respondents are evenly drawn. The NSM method is supposed to prevent the strong selection bias that characterises RDS surveys while obtaining estimators with a precision similar to what we could get with a simple random sampling design. Experimented for the first time in France by the French Institute for demographic studies (Ined) in the context of a survey on Chinese migrants in the Île-de-France region, this work is a restitution of the difficulties encountered during data collection, which can sometimes be contextual and due to the pandemic, but for the most part inherent to the method.

1. NSM, une méthode innovante et ambitieuse pour l'étude de populations cachées

1.1. Les immigrés chinois en Ile-de-France, une population difficile à enquêter

Depuis quelques décennies, on assiste à un élargissement des champs d'études d'enquêtes sociales à des populations dites « marginales » ; Schiltz situe l'émergence de ces études dans la statistique publique dans les années 1990 (Schiltz, 2005). Depuis, la liste de ces populations s'allonge régulièrement dans la littérature, notamment avec l'apparition de nouveaux phénomènes sociaux. Ces populations « en marge » sont en effet des groupes pour lesquels le besoin en données est important : souvent en situation de précarité ou exposées à des risques sociaux ou sanitaires, ces populations ont besoin d'une prise en charge ou d'un accompagnement dans le domaine de la santé, du logement ou encore de la justice.

Cependant, ces populations sont souvent difficiles à enquêter car elles sont de taille relativement faible, ses membres ne sont pas identifiables dans une base de sondage et peuvent même vouloir cacher leur appartenance à la population d'intérêt pour éviter d'être stigmatisés (Marpsat & Razafindratsima, 2010). Le recours à des enquêtes en population générale pour espérer capter ces populations ne permet généralement pas d'avoir un échantillon suffisamment grand pour mener des analyses statistiques. De plus, la possible stigmatisation de ces populations peut conduire à une sous-déclaration dans ce type d'enquête, comme on peut l'observer pour les personnes ayant eu recours à l'avortement (Jones & Kost, 2007). Ces contraintes excluent donc a priori le recours à des enquêtes probabilistes en population générale pour capter et caractériser ces populations. Néanmoins, du fait de leur diversité, les méthodes d'enquêtes doivent être adaptées aux spécificités des populations étudiées. À titre d'exemple, on ne peut adopter la même stratégie pour enquêter des personnes sans-abri, des usagers de drogues ou des personnes recourant à une assistance médicale à la procréation à l'étranger.

À plusieurs égards, les immigrants chinois en Île-de-France s'inscrivent dans cette catégorie des populations difficiles à atteindre : dans l'enquête Trajectoires et Origines (TeO) (Beauchemin et al. 2016), menée en 2008 en population générale, on ne compte que 69 répondants nés en République populaire de Chine ou à Hong Kong¹, ce qui témoigne du fait que les enquêtes en population générale ne permettent pas d'avoir un échantillon suffisamment important pour mener des analyses sur cette population. D'autre part, les sources de données comme le recensement français ou l'échantillon démographique permanent (EDP) ne contiennent que peu d'informations pour caractériser de manière approfondie ces groupes de populations immigrées et peuvent, par ailleurs, souffrir de défaut de couverture (manquer les migrants en situation irrégulière par exemple). Pour ces mêmes raisons de défaut de couverture, les bases de sondage telles que Fidéli ne nous semblent pas adaptées pour tirer un échantillon de notre population d'intérêt.

1.2. Les méthodes d'enquêtes par réseau particulièrement adaptées à ce type de populations

Les méthodes de sondage par chaînage/réseau consistent à enquêter au sein de réseaux en mettant à contribution les répondants pour en recruter de nouveaux. Elles ne nécessitent donc pas de base de sondage.

Le recours à ces méthodes est motivé principalement par trois raisons : (1) une croyance généralement acceptée selon laquelle les répondants sont plus susceptibles de participer et de répondre honnêtement à l'enquête lorsqu'ils sont désignés et invités à répondre par des connaissances ou des amis plutôt que contactés au hasard ; (2) un désir de comprendre la structure des réseaux et le rôle des relations sociales ; (3) une volonté d'enquêter, d'étudier et de comprendre des populations stigmatisées ou cachées, difficiles à atteindre avec les méthodes de sondage traditionnelles.

Ces méthodes ont longtemps souffert de la réputation de produire des estimations biaisées. Cependant, des études récentes basées sur la méthode « Respondent-driven sampling » (RDS) ont montré que sous certaines hypothèses, la probabilité d'être échantillonné pouvait être estimée par le nombre de liens que chaque répondant a avec les autres membres de la population. Les hypothèses sous-jacentes sont les suivantes : (1) la population est formée d'un seul réseau : tous les individus sont connectés entre eux, plus ou moins directement. Cette hypothèse est essentielle car elle permet de mobiliser la théorie de la chaîne de Markov et ainsi argumenter en faveur de la convergence des estimateurs vers des estimateurs sans biais. (2) Les liens entre les individus sont réciproques : si A cite B, B citera nécessairement A. (3) Les individus recrutent aléatoirement parmi leurs connaissances.

Malgré cela, la précision de la méthode reste très sensible aux hypothèses qu'elle formule sur le réseau étudié (Gile et Handcock 2010 ; Goel et Salganik 2010). Un problème fondamental des approches fondées sur la marche aléatoire en général, et du RDS en particulier, est qu'elles ont une variance d'échantillonnage plus élevée que l'échantillonnage aléatoire simple. Alors que la variance d'échantillonnage d'un échantillon aléatoire simple est inversement proportionnelle à la taille de

¹ Aucune surreprésentation de ce groupe n'avait été réalisée ; la dernière édition de l'enquête inclut une telle surreprésentation.

l'échantillon, la précision des RDS est fonction à la fois de la taille de l'échantillon et de la structure du réseau. Dans les réseaux présentant des degrés élevés de regroupement (clusters), le RDS a tendance à rester bloqué dans des composantes ségréguées du réseau, ce qui peut entraîner des échantillons biaisés. Dans certains réseaux, la variance d'un RDS pourrait être trop élevée pour être acceptable, mais le chercheur, ne connaissant pas le réseau dans lequel il échantillonne, serait incapable de distinguer les résultats corrects des résultats incorrects. Compte tenu de cette incertitude fondamentale, la capacité à déduire la prévalence de phénomènes mesurée par l'enquête au niveau de la population à partir d'échantillons fondés sur la marche aléatoire est discutable. En effet, Goel et Salganik (2010) ont trouvé des niveaux très élevés de variance d'échantillonnage pour le RDS dans un échantillonnage simulé utilisant des données de réseaux réelles.

L'objectif de cette communication est donc de mobiliser une nouvelle approche, la méthode « Network sampling with memory » (NSM), qui s'appuie sur des développements récents dans les littératures mathématiques et informatiques sur l'échantillonnage de grands réseaux. L'approche de marche aléatoire de base est améliorée grâce à l'incorporation d'informations sur la topographie locale du réseau et sur les cas récemment échantillonnés pour gagner en efficacité lors de l'échantillonnage (Alon et al. 2008 ; Avin et Krishnamachari 2008 ; Cooper, Frieze et Radzik 2009 ; Ribeiro et Towsley 2010). Comme nous l'expliquons dans cet article, le NSM est supposé améliorer l'efficacité de l'échantillonnage à partir d'un réseau en recueillant des données sur le réseau auprès des répondants dans le cadre de l'enquête, qui sont utilisées pour révéler progressivement la liste des membres de la population. Au fur et à mesure que l'échantillon progresse, la liste des personnes qui ont été désignées dans l'enquête tend à ressembler à la liste complète des membres de la population, ce qui permet théoriquement au NSM de se rapprocher du processus de sondage aléatoire simple.

1.3. Les apports de la variante Network sampling with memory (NSM)

Les concepteurs de la méthode *Network sampling with memory* ont voulu développer une méthode asymptotiquement sans biais (comme l'approche RDS lorsque ses hypothèses sont satisfaites) et surtout dont la variance d'échantillonnage et donc l'erreur absolue moyenne diminuent rapidement avec la taille de l'échantillon collecté.

Pour cela, la méthode NSM recourt à deux modes d'échantillonnage: un mode « List » qui garantit des estimations asymptotiques sans biais (*Naive list mode* et *Even sampling*) et un mode « Search » qui est conçu pour fonctionner en tandem avec le mode « List » afin d'améliorer la précision des estimations. L'intuition de base est assez simple : le processus d'échantillonnage est utilisé pour découvrir la liste des membres de la population, puis l'échantillonnage se fait dans cette liste avec remplacement. Chaque étape du processus d'échantillonnage consiste en un questionnaire avec un répondant qui, en plus des questions de fond de l'enquête, se voit demander de fournir les coordonnées de ses amis ou contacts qui sont membres de la population A (cette liste d'individus est désignée par le terme « roster » par la suite). À chaque étape, de nouveaux membres du réseau sont ajoutés à la liste (roster total), qui comprend tous les membres de la population qui ont été désignés par les répondants.

Les différents modes d'échantillonnage interviennent de manière successive au fil de la collecte : d'abord le *Naive list mode*, puis le *Search*, et enfin l'*Even sampling*. Afin de déterminer le mode d'échantillonnage à mobiliser, on s'appuie sur une batterie d'indicateurs permettant de rendre compte du niveau d'exploration du réseau. Parmi eux, on trouve la taille du réseau dévoilé (L), le nombre d'interviews réalisées (*Step*) et la part de personnes citées une seule fois parmi toutes les personnes dévoilées (P1).

Naive list mode : Démarrer la collecte aléatoirement

La collecte démarre par l'identification de graines, qui seront les premiers individus enquêtés. Leurs réponses fournissent un premier réseau. En début de collecte, on tire aléatoirement un individu parmi les premiers identifiés dans le réseau (appelés « nœuds »). Ayant une connaissance très faible du réseau à ce moment-là, il convient de mobiliser la technique de sondage la plus simple et égalitaire : le sondage aléatoire simple. Le problème du *Naive List mode* est que les nœuds nommés en début de

collecte vont être sur-échantillonnés par rapport à ceux apparus plus tard, les premiers apparus étant éligibles au tirage pendant plus longtemps.

Search mode : Assurer l'exploration du réseau

L'objectif est d'enquêter des pans du réseau encore inexplorés. Pour cela, il faut repérer les répondants les plus susceptibles de nous y conduire : ces individus sont appelés « nœuds-ponts » (ponts entre sous-parties du réseau). On calcule, pour chaque répondant, sa probabilité d'être un nœud-pont, fondée sur la proportion de personnes citées une seule fois parmi leurs « amis ». Après identification des 5 individus les plus susceptibles d'être des nœuds-ponts, on en sélectionne un (proportionnellement à sa probabilité d'être un nœud-pont), puis on tire aléatoirement un de ses amis parmi ceux cités une seule fois et non enquêtés. Il s'agit là d'un tirage à deux degrés. On mobilise ce mode après quelques interviews afin d'orienter au mieux l'échantillonnage vers de nouveaux pans du réseau, et éviter de rester bloqué dans un cluster.

Even sampling mode (List mode) : Homogénéiser les expositions au tirage

Pour atténuer l'effet du sur-échantillonnage des nœuds apparus tôt, le mode *Even sampling* permet d'homogénéiser les CSR (*cumulative sampling rate*). A chaque tirage, quel que soit le mode, le CSR des individus exposés à l'échantillonnage est incrémenté de leur probabilité d'être sélectionné au cours du tirage. Actualisé dès lors que l'individu est candidat à l'échantillonnage, le CSR mesure l'exposition des individus aux tirages successifs. Homogénéiser les CSR consiste donc à exposer au tirage les individus nouveaux ou jamais soumis au tirage jusqu'alors. Pour cela, on exclut du tirage les nœuds les plus exposés au tirage précédemment, et ayant un CSR important. Ici, on recourt à un tirage aléatoire simple parmi les nœuds ayant un CSR inférieur à l'ESR (*Even sampling rate*) (ou les 100 CSR les plus bas si le volume de nœuds remplissant la condition est inférieur à 100).

Figure 1 : Les différents modes d'échantillonnage de la méthode NSM et les critères de passage d'un mode à l'autre

Step	Mode					
	If Step<4					
1	Naive list					
2						
3						
	[Default mode]					
4	Search					
5						
....						
49						
50						
				If Step>50 & P1<A1 (=0.4)		
				If Step>50 & P1>A1 (=0.4)	L[Step-5]=L[Step]	L>200
51	Search	Even Sampling	Even Sampling	Naive List		
52						
...						

La sortie du mode *Search* repose sur trois critères : (1) avoir collecté plus de 50 questionnaires (Step>50), (2) avoir une taille de réseau supérieure à 200 (L>200) ou si les 5 dernières interviews n'ont pas permis de découvrir de nouveaux individus (L[Step-5]=L[Step]) et enfin (3) lorsque le paramètre P1 (nombre d'individus cités une seule fois et non enquêtés) passe en dessous du seuil A1 (seuil fixé par les concepteurs de la méthode). Lorsque P1 est bas, cela signifie que peu d'individus dans le réseau n'ont été cités qu'une fois, autrement dit, beaucoup d'individus ont été cités plusieurs fois, c'est-à-dire que l'on a atteint un certain niveau d'exploration. Il convient alors de passer en *Even sampling* afin de soumettre aux tirages les individus jusqu'alors ignorés, et d'équilibrer les CSR. En pratique, les deux premiers critères ont été assez rapidement atteints. C'est le critère P1<A1 qui nous est apparu comme le plus discriminant, celui qui nous a contraint à rester en mode *Search* jusqu'à la fin de la collecte.

2. Passer de la théorie à la pratique : la complexe mise en œuvre de la méthode NSM sur le terrain

La collecte des données de l'enquête ChiPre s'est déroulée de septembre 2020 à juin 2021 et a mobilisé au total 10 enquêteur-riche-s et une équipe de 5 gestionnaires de collecte (informaticien, statisticien-ne-s, ingénieure d'études). La partie suivante aborde en détail notre expérience du passage en production de la méthode NSM, méthode innovante, non stabilisée et qui ne dispose pas de guide pratique établi pour sa mise en œuvre.

2.1. Finaliser le protocole d'échantillonnage et de collecte « sur le tas »

En dépit d'une intense préparation et d'une automatisation conséquente des processus de collecte (algorithmes de matching et de tirage, application web de collecte et de suivi conçue ad-hoc), la mise en place de cette méthode sur le terrain a nécessité de nombreux ajustements en début et en cours de collecte.

2.1.1. Une élaboration des principes de tirage par « tâtonnement »

La première difficulté à laquelle s'est trouvée confrontée l'équipe gestionnaire a été l'élaboration de règles encadrant les échantillonnages. La méthode implique de suivre le dévoilement progressif d'un réseau, et donc nécessite de renouveler régulièrement les échantillonnages afin d'alimenter la liste des individus « à enquêter » et de progresser au sein de ce réseau. La méthode ayant très peu été testée sur le terrain, nous ne disposions pas de recommandations sur les fréquence et taille optimales de ces échantillonnages successifs.

Concernant la fréquence, nous avons commencé par produire de très petits tirages (entre 5 et 10 individus) deux fois par semaine. Cette stratégie a permis de suivre de près l'évolution du terrain et du réseau au démarrage de la collecte mais s'est rapidement avérée trop chronophage et nous avons alors décidé de ne réaliser qu'un seul tirage plus conséquent par semaine (et exceptionnellement de gros tirages à la veille de périodes de congés).

Pour déterminer le nombre d'individus tirés à chaque phase d'échantillonnage, nous mobilisons les paramètres suivants : la charge d'enquête restante pour chaque enquêteur-riche ; leur charge de relance et d'identification des rosters incomplets (un enquêté donne des citations dans son roster mais souhaite d'abord demander à ses amis leur accord pour transmettre leurs coordonnées, il faut alors le relancer régulièrement pour obtenir ces informations) ; la taille du réseau au moment du tirage par rapport à sa taille lors du précédent. Après diverses expérimentations, nous avons décidé de produire des tirages de taille modulable sans jamais dépasser la moitié des nouveaux arrivés dans le réseau depuis le précédent tirage. Si la collecte de la semaine avait permis de découvrir 50 nouvelles personnes dans le réseau (50 nouvelles lignes uniques ajoutées dans le roster total), alors le tirage pour la semaine suivante ne devait pas dépasser 25 échantillonnés. L'inconvénient de cette méthode est que la taille des tirages ne peut pas être anticipée. Il faut d'abord identifier les doublons afin de connaître le nombre de nouveaux individus ajoutés dans le roster total.

Une fois le tirage réalisé, reste ensuite à répartir les échantillonnés entre enquêteur-riche-s. Nous avons démarré le terrain avec un objectif fort de respect de l'équité entre enquêteurs : donner à chacun le même nombre de nouveaux cas pour égaliser les chances d'avancer dans le réseau. Ce choix s'est rapidement montré intenable car il était en contradiction avec le respect des « filiations » sur le terrain. Les enquêteur-riche-s souhaitaient suivre « leur » réseau, c'est-à-dire enquêter les cités de « leurs » enquêté-e-s (si échantillonnés). Or, une répartition équitable d'échantillonnés entre eux impliquait des « transferts de filiation », qui se sont révélés beaucoup trop coûteux sur le terrain (il faut négocier avec un répondant que son réseau soit enquêté par un autre enquêteur inconnu !). Nous avons donc opté pour un principe de strict suivi des « filiations » au moment de la répartition des échantillonnés.

Avec ces diverses contraintes, il est possible de programmer un nombre total d'échantillonnés, mais il est impossible d'anticiper la charge nouvelle que recevra chaque enquêteur-riche à chaque tirage. Cela induit une forte complexité dans le management de l'équipe sur le terrain.

2.1.2. Des modifications de l'algorithme d'échantillonnage en cours de terrain

La mise en production de l'algorithme en collecte réelle s'est accompagnée de diverses surprises que des simulations sur des réseaux fictifs ne permettaient pas d'anticiper.

Sans entrer dans le détail de toutes les modifications mises en œuvre, nous avons adapté les critères de calcul permettant de déterminer la liste des rosters éligibles à chaque tirage en mode *Search*, en excluant les nœuds déjà échantillonnés auparavant afin, entre autre, de limiter la surexposition artificielle au tirage des rosters déjà candidats au tirage précédemment.

Face au rythme très lent d'avancée dans le réseau et au regard de nos objectifs et de la durée de l'enquête, nous avons décidé d'augmenter le seuil A1 (cf. partie 1) afin d'avoir plus de chances de l'atteindre et d'expérimenter la « sortie » du mode *Search*. Ce seuil n'a pas été décidé arbitrairement, il a été tiré de l'article fondateur de la méthode NSM, dans lequel diverses versions de ce seuil sont présentées (Mouw, 2012).

2.1.3. Des adaptations du protocole face aux réalités du terrain

Face à la lente progression dans le réseau, nous avons dû prolonger la durée de l'enquête, qui devait initialement prendre fin en mars 2021, jusqu'en juin 2021. Cette prolongation s'est accompagnée d'un renouvellement partiel de l'équipe des enquêteur·rice·s et donc d'une nouvelle formation en cours d'enquête.

En raison des difficultés d'accès à certains groupes sociaux et régionaux, et pour compenser l'extinction progressive de certaines branches du réseau, de nouvelles graines ont été recrutées et ajoutées en cours de collecte. Les personnes originaires de la région de Wenzhou, les étudiants (dont la forte présence dans le recensement se reflétait mal dans notre réseau), et les personnes les plus précaires ont fait l'objet d'une attention particulière à cet égard.

La réalité du terrain nous a rapidement imposé plus de souplesse dans l'attribution des chèques cadeaux pour remercier les enquêté·e·s. Le protocole prévoyait un chèque d'un montant de 15 euros pour le questionnaire, et de 20 euros pour le remplissage d'un roster dans lequel 6 contacts de l'enquêté·e sont renseignés. Il nous a semblé plus réaliste de distribuer le chèque roster à partir de 3 contacts cités.

Finalement, la mise en œuvre de la méthode NSM, sans guide pratique existant pour son passage en production, a induit de nombreux tâtonnements pour cerner le fonctionnement de l'algorithme, pour concevoir peu à peu une stratégie de tirages proportionnés, un suivi des graines adapté et des protocoles plus réalistes. Cette expérimentation a nécessité une vigilance permanente aux évolutions du réseau collecté et une forte capacité d'adaptation pour toutes les équipes impliquées. Une fois les bases du fonctionnement de la collecte mises en place, l'équipe en charge des échantillonnages a toutefois dû faire face à des difficultés liées aux spécificités de l'algorithme utilisé.

2.2. Découvrir un algorithme aux effets contre-intuitifs et parfois désincitatifs

2.2.1. La préférence paradoxale pour des rosters petits et sans doublons, allant à l'encontre d'une dynamique de progression dans le réseau

Tandis que nos consignes aux enquêteur·rice·s insistaient sur l'importance de collecter de grands rosters (« collectez le maximum de citations dans vos rosters »), nous avons peu à peu réalisé que l'algorithme, lui, favorisait à l'inverse les petits rosters lors des échantillonnages. Selon notre interprétation, cet « effet pervers » est dû au fait que l'algorithme considère les petits rosters comme décrivant probablement des réseaux de personnes plus isolées, ou de potentiels nœuds-ponts vers d'autres sous-réseaux plus difficiles d'accès.

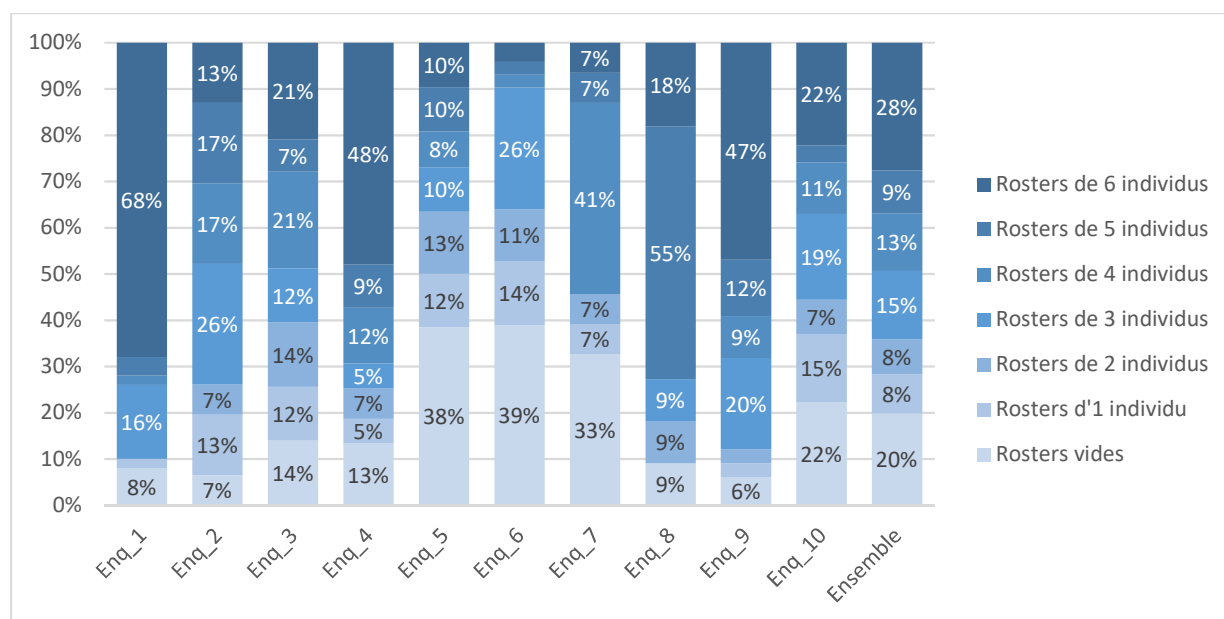
Autre découverte : l'algorithme défavorise de surcroît les rosters contenant des doublons (un doublon est un individu déjà cité au préalable par un·e autre enquêté·e et identifié comme tel à l'issue de notre

processus de matching²). La quête des doublons fait toutefois partie des objectifs de l'enquête : plus on a de doublons, plus l'indicateur de saturation du réseau baisse (P1, cf partie 1) et plus on a de chances de passer sous le seuil de sortie du mode *Search* (A1) pour basculer en *Even sampling*. Les enquêteur-riche-s sont donc formé-e-s à collecter suffisamment d'informations pour rendre l'identification des doublons réalisable grâce à l'algorithme de matching conçu spécialement pour l'enquête (informations sur l'identité des cités).

2.2.2. ...Et induisant des stratégies et des « biais enquêteurs » non négligeables

Ces tendances de l'algorithme sont contre-intuitives et surtout contre-productives car les enquêteur-riche-s finissent par s'en rendre compte et par adapter leur comportement pour minimiser les efforts de collecte. Une enquêtrice a rapidement compris ces deux effets et s'est mise à collecter de petits rosters contenant de 1 à 3 citations, s'assurant ainsi de limiter la présence de doublons et maximisant également ses chances de voir ses rosters sélectionnés au tirage suivant (voir figure 2, enq 6). Elle était donc certaine d'avoir une liste de « à enquêter » toujours remplie, tout en fournissant le minimum d'effort de négociation des rosters. Or l'avancée dans le réseau, et à terme le passage d'un mode d'échantillonnage à l'autre, repose sur un effort collectif visant à dévoiler le maximum de liens dans le réseau enquêté. De plus, la structure du réseau en ressort biaisée : ses enquêté-e-s disposaient sans doute de réseaux plus étendus que ce dont leur roster témoigne en définitive, et ses enquêté-e-s au roster vide avaient probablement un réseau, plus difficile à négocier, mais pas inexistant.

Figure 2 : Distribution des rosters collectés en fonction de leur taille (à l'exclusion des rosters des graines) par enquêteur-riche



Sans aller jusqu'à ce niveau de rationalisation, le biais « anti-doublon » reste désincitatif : la présence d'un doublon n'est pas perçue comme une bonne nouvelle par les enquêteur-riche-s car elle implique que leur roster ne sera pas prioritaire au tirage. Le travail de collecte des informations identifiantes (les plus difficiles à négocier) qui a permis la détection du doublon n'est donc pas du tout valorisé. Bien que coûteux à collecter, les doublons sont nécessaires, et l'algorithme n'incite pas les enquêteur-riche-s à les « produire ».

² Le matching est réalisé en deux étapes : un algorithme de matching attribue un score de similarité à chaque ligne de roster, puis une vérification manuelle permet de trancher sur les cas ambigus au score « moyen » ne permettant pas mécaniquement de conclure à l'existence ou non d'un doublon.

2.2.3. Une temporalité d'échantillonnage distante de la réalité du terrain

Enfin, la temporalité de l'algorithme n'est pas la même que celle du terrain. Certains rosters collectés en décembre ont pu être échantillonnés en mars, demandant un travail de négociation difficile sur le terrain pour les enquêteur·rice·s.

Ces résultats de l'échantillonnage sont d'autant moins acceptés par les enquêteur·rice·s qu'ils se cumulent à d'autres, décrits précédemment : si l'algorithme mobilise des rosters anciens moins exploités, c'est aussi parce que l'offre de rosters en cours est moins propice à être échantillonnée, en raison notamment des doublons. Au découragement lié à la non sélection des rosters de taille conséquente avec doublons bien identifiés, s'ajoute la difficulté d'aller négocier la poursuite de la collecte auprès d'enquêté·e·s sollicité·e·s plusieurs mois en amont.

Pour conclure, l'algorithme produit des effets contre-intuitifs rendant le terrain et la gestion de la collecte ardu. Les effets pervers de l'échantillonnage induisent des comportements d'enquêteur·rice·s qui, par stratégie ou découragement, peuvent aller à l'encontre de l'objectif initial de collecter des rosters de grande taille et d'identifier correctement les doublons. Un effort de pédagogie constant de la part des équipes gestionnaires a été nécessaire pour rendre le processus de collecte acceptable et, autant que possible, cohérent avec les principes sur lesquels repose la méthode NSM. Enfin, la méthode nécessite impérativement un cadre de passation en face-à-face, ce dont nous a privé la pandémie de Covid-19.

2.3. Expérimenter un contexte sanitaire très défavorable aux enquêtes en face-à-face

La crise sanitaire en général, et la suspension du face-à-face induite par le second confinement et la reprise de l'épidémie à l'automne, ont eu des impacts conséquents sur la collecte entre septembre 2020 et juin 2021.

2.3.1. Une population d'autant plus difficile à joindre dans ce contexte de crise

Les immigrés chinois en Île-de-France ont très tôt montré une grande méfiance liée aux risques sanitaires, étant informés du déroulé de l'épidémie en Chine. De plus, les retours du terrain ont montré qu'une partie d'entre eux avaient d'autres préoccupations en tête, après avoir essuyé de lourdes pertes avec la fermeture de commerces et de restaurants, ou parce qu'il fallait maintenir un lien avec les membres de la famille en Chine, souvent eux-mêmes en difficulté à cause de l'épidémie.

Les enquêteur·rice·s ont aussi témoigné d'un effet de contraction du réseau de sociabilité qui a encore complexifié le travail de collecte des rosters : les enquêté·e·s ne fréquentent plus autant leurs amis et connaissances qu'avant la crise, et il est d'autant moins approprié qu'ils les sollicitent à l'occasion d'une enquête par questionnaire.

Certaines sous-populations se sont montrées particulièrement inaccessibles en distanciel. Les Wenzhou semblent appartenir à des réseaux très fermés, peu disponibles et peu intéressés par la problématique de recherche mise en avant. Les argumentaires pour les convaincre ont été difficiles à trouver et l'absence d'échange en face-à-face a nécessité de nombreux efforts de la part des enquêteur·rice·s.

De même, les immigrés sans-papiers constituent une sous-population vulnérable, méfiante et peu prompt à citer des connaissances, surtout à distance. Ils vivent souvent dans des logements partagés, dans des contextes où l'intimité fait défaut pour répondre correctement à l'enquête par téléphone. Une forte implication sur le terrain pour tisser un lien de confiance avec ces enquêté·e·s a été nécessaire en début de collecte et l'arrêt du face-à-face n'a pas permis de faire se développer ce réseau autant qu'attendu.

2.3.2. Des leviers négociation largement altérés par l'abandon du face-à-face

La négociation des coordonnées des cités (identifiant *WeChat*³ ou numéro de téléphone) est beaucoup plus difficile à distance. Collecter des lignes de roster repose sur la confiance tissée lors de la passation du questionnaire, et la phase de négociation pour obtenir les coordonnées des amis des enquêté·e·s, avec l'insistance requise, est plus simple à aborder en face-à-face qu'en distanciel. En effet, l'enquêteur·rice peut demander à l'enquêté·e d'appeler ses amis dans l'immédiat et l'accompagner dans l'argumentaire, il peut montrer le flyer de l'enquête ou encore dire à l'enquêté·e d'envoyer les supports de communications à ses amis via *WeChat*. Si le « contrôle » sur la négociation de l'enquêté·e avec ses amis est envisageable en face-à-face, il ne l'est plus du tout en distanciel. En conséquence, obtenir un roster complet demande à l'enquêteur·rice de relancer l'enquêté·e ultérieurement et prend plus de temps. Tout le rythme de collecte s'en trouve ralenti (puisque les échantillonnages des futurs candidats dépendent de la collecte des coordonnées des amis cités par les enquêté·e·s précédent·e·s).

Ne pas pouvoir remettre le chèque cadeau immédiatement en main propre à l'issue de la passation est une autre conséquence de l'arrêt du face-à-face. De plus, le chèque-cadeau papier n'était pas utilisable sur internet alors que nous étions confinés, certains s'en sont plaint (les étudiants surtout). Cela a rendu notre incitation financière moins efficace et la recommandation auprès des pairs sans doute moins motivée.

La pandémie a complexifié notre enquête auprès d'une population déjà difficile d'accès au départ et éprouvée par les stigmates d'un virus identifié comme provenant de son pays natal. La collecte des rosters a été impactée par la faiblesse des liens relationnels entretenus à cette période, et par l'impossibilité de négocier en face-à-face, chèque-cadeau et autres stratégies à l'appui. Enfin, le mode distanciel est peu adapté à ce type de protocole reposant essentiellement sur la recommandation par les pairs, et donc la confiance et les liens établis sur le terrain. Toutefois, le maintien sur le terrain pendant presque 10 mois a permis d'aboutir au dévoilement d'un réseau de près de 1700 individus. La partie suivante permet d'avoir un aperçu de l'échantillon ainsi collecté.

3. Traitements post-enquête et premiers résultats

Malgré ces nombreuses difficultés, 501 questionnaires ont été collectés, conduisant à un taux de réponse sur les personnes échantillonnées de 60 % (844 individus échantillonnés et non identifiés comme hors champ). En plus des questionnaires, 1689 « liens » ont été recensés via les rosters, soit 3,37 individus cités par interview (le protocole permettait un maximum de 6 citations par individu). Malgré le recours à des incitations financières, la collecte d'informations sur le réseau de sociabilité des enquêté·e·s s'est avérée difficile. Cette base de sondage constituée au fil de la collecte a permis de recenser 1 523 individus distincts dans le réseau (cf. Figure 3).

Du point de vue de la méthode, nous n'avons jamais rempli les conditions pour sortir du mode *Search*. L'indicateur P1, correspondant à la proportion d'individus non enquêtés et cités une seule fois, n'est jamais descendu en dessous de 0,6 alors que nous avons fixé le seuil A1 à 0,4. Les concepteurs de la méthode avaient fait des simulations avec différentes valeurs pour A1 (0,2 ou 0,4). Nous nous sommes posés des questions sur les conséquences du non-passage en mode *Even sampling* sur la qualité et la représentativité de l'échantillon, et sur la validité de la méthode dans ce cas de figure.

En termes de post-traitement, le calcul des poids (tel que proposé dans l'article de Ted Mouw de 2012) s'est avéré assez simple à mettre en œuvre dès lors qu'anticipé. Les poids de sondage dépendent du taux de sondage cumulé de chaque individu (somme des taux de sondage sur chaque échantillonnage, soit le CSR du dernier tirage) et de la probabilité d'être nommé au moment du tirage effectif. Ces informations doivent donc être stockées au fil de la collecte car elles ne peuvent être calculées a posteriori. Cependant, les poids de sondage calculés ainsi sont très dispersés, avec un ratio du poids maximum sur le poids minimum très important (supérieur à 400), mettant en péril la validité des

³ WeChat est une application mobile de messagerie textuelle et vocale développée par le géant chinois Tencent Holdings Limited. Elle permet également les appels audio et vidéo. L'application est très populaire en Chine et compte plus d'un milliard de comptes dans le monde en mars 2018.

estimations produites. Les concepteurs de la méthode ont, en pratique, assez peu mobilisé la formule simple et préféré calculer les poids de sondage par bootstrap ; nous n'avons pas investi cette piste, qui pose des questions sur les hypothèses sous-jacentes au processus de bootstrap.

Nous avons ensuite mobilisé les données de la base de sondage pour réaliser une correction de la non-réponse totale. Une correction de la non-réponse par GRH a été tentée, mais cette opération conduisant à une détérioration du système des poids sans pour autant améliorer la structure de l'échantillon, nous avons donc préféré abandonner cette piste.

Nous avons finalement réalisé un calage de l'échantillon sur les marges du recensement de la population 2018. Comme discuté en partie 1, nous pouvons considérer que les données du recensement sont imparfaites (en raison d'un potentiel biais de couverture). Cependant, nous avons décidé de les mobiliser pour caler notre échantillon pour corriger d'une part les biais liés à la non-réponse, et d'autre part l'effet de structure du graphe dévoilé, qui présente une forte surreprésentation des jeunes, des personnes très diplômées et des personnes arrivées très récemment en France.

Ce constat ne remet cependant pas en cause l'intérêt de la méthode, qui permet d'une part de capter des personnes potentiellement absentes du recensement (et donc inaccessibles si un échantillon était tiré dans cette base), mais aussi de constituer une base des liens entre individus de cette population, permettant des analyses de réseau, de proximité et d'homogénéité de celui-ci. En revanche, le fort biais que présente le réseau dévoilé nous conduit à nous interroger sur les paramètres de la méthode, et notamment sur le choix des graines et son implication sur la structure finale de l'échantillon. Il est possible que la structure des graines (surreprésentation des étudiants) conduise à ce même biais parmi notre échantillon final de 501 répondants. Enfin, la méthode n'étant jamais sortie du mode *Search*, il est possible que cela explique ces biais, qui auraient été corrigés lors d'une phase *Even sampling*, malheureusement jamais atteinte. Cependant, les estimateurs que permet de construire la méthode NSM sont supposés sans biais même avant cette dernière étape. Il est alors possible que la non-réponse des individus conduise à ce biais. D'autres tests seraient ainsi nécessaires pour conclure sur ce point.

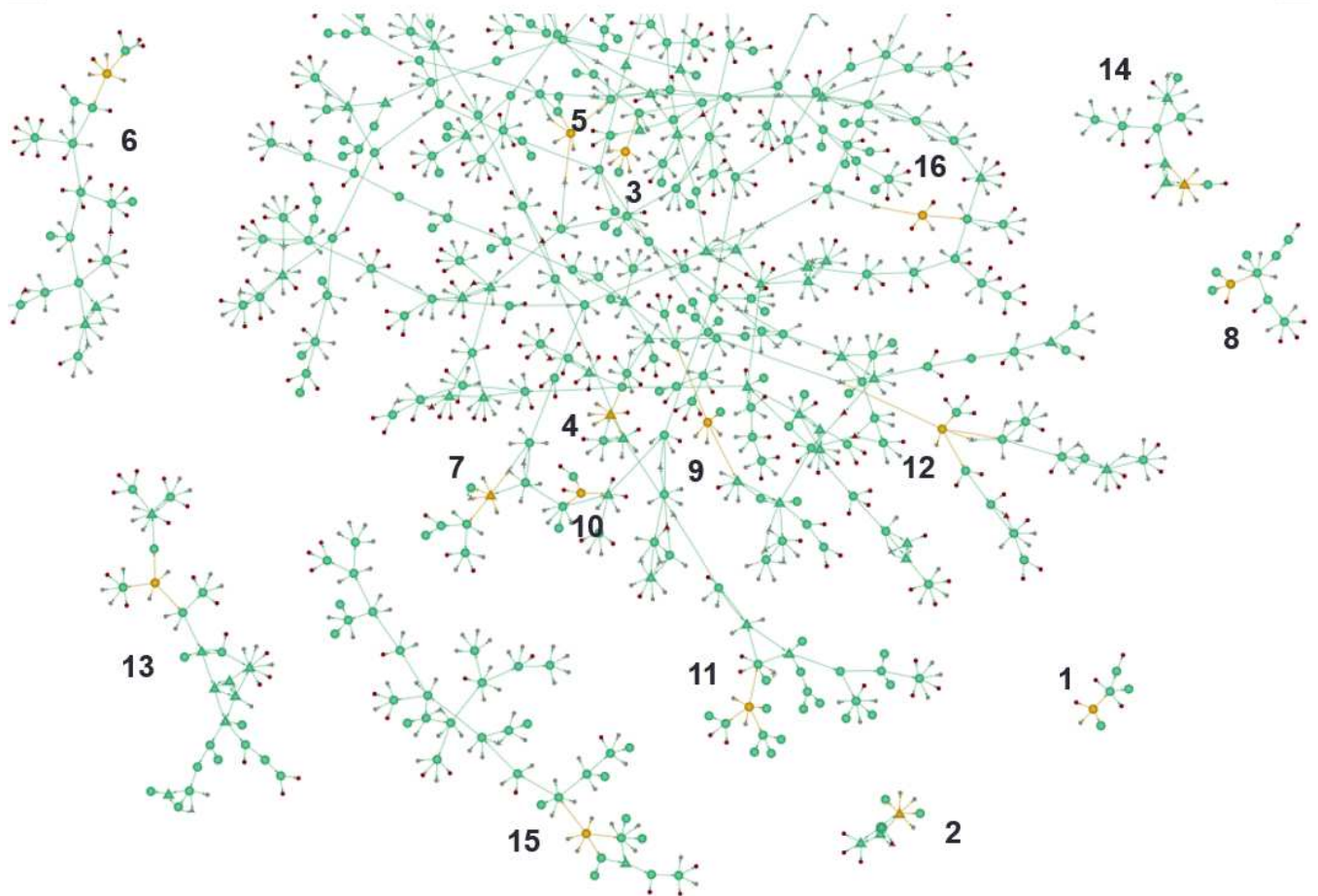
4. Conclusion

Il est difficile de prédire ce qu'auraient été nos résultats si notre collecte n'avait pas été affectée par la pandémie. Le contexte sanitaire et le passage au distanciel ont eu un impact fort sur l'acceptabilité de l'enquête auprès de notre population cible, sur la recommandation par les pairs pourtant centrale dans les méthodes par chaînage, sur le rythme de collecte et sur les conditions de travail des enquêteur·rice·s.

Toutefois, certaines difficultés de terrain semblent intrinsèquement liées à la méthode NSM. L'imprévisibilité de la « charge enquêteur », la préférence pour les petits rosters et l'aversion pour les doublons sont des effets qui vont à l'encontre de l'objectif initial de collecter des rosters de taille conséquente pour faire avancer l'enquête, et qui altèrent l'acceptabilité de la méthode par ceux qui doivent l'éprouver au quotidien. NSM apparaît ainsi comme une méthode conçue à partir de réseaux théoriques ou déjà dévoilés mais qui peine à mesurer le coût pratique du processus même de dévoilement du réseau. Notre article constitue un apport en ce sens : il permet un éclairage concret sur la mise en application d'une telle méthodologie.

En définitive, la méthode nous semble prometteuse, mais sans doute plus adaptée pour une expérimentation auprès d'une population un peu mieux connue et de taille plus modeste. Une population plus restreinte permettrait de tester la sortie du mode *Search* et l'application du mode *Even Sampling* dans un temps de collecte raisonnable. Cela permettrait de mieux rendre compte des avantages de la méthode en termes de qualité d'échantillon collecté, par rapport à un RDS simple.

Figure 3 : Graphe du réseau découvert en fin de collecte



Lecture : les nombres correspondent les graines.

Bibliographie

- [1] Alon, N., C. Avin, M. Koucky, G. Kozma, Z. Lotker, and M. R. Tuttle. 2008. Many Random Walks Are Faster Than One. Pp. 119–28 in Proceedings of the Twentieth Annual Symposium on Parallelism in Algorithms and Architectures. New York: Association for Computing Machinery
- [2] Avin, C. and B. Krishnamachari. 2008. 'The Power of Choice in Random Walks: An Empirical Study. Computer Networks 52(1):44–60.
- [3] Beauchemin C. et al. Trajectoires et Origines. Enquête sur la diversité des populations en France, Ined Editions, 2016
- [4] Cooper, C., A. Frieze, and T. Radzik. 2009. 'Multiple Random Walks and Interacting Particle Systems. Lecture Notes in Computer Science 5556:399–410
- [5] Gile, K. J., & Handcock, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. Sociological Methodology, 40, 285-327.
- [6] Goel S, Salganik M.J, «Assessing respondent-driven sampling», PNAS April 13, 2010 107 (15) 6743-6747
- [7] Jones K J, Kost K. Underreporting of induced and spontaneous abortion in the United States: An analysis of the 2002 National Survey of Family Growth. Studies in family planning, Vol. 38, Issue 3, Septembre 2007 : 187-197
- [8] Marpsat M, Razafindratsima N. Les méthodes d'enquêtes auprès des populations difficiles à joindre: Introduction au numéro spécial. Methodological Innovations Online. 2010;5(2):3-16.

- [9] Merli G, Mouw T, Stolte A, Le Barbenchon C, Florey-Eischen F, Using Multiple Modes of Data Collection to Recruit Migrant Samples With Network Sampling With Memory: The Chinese Immigrants in Raleigh-Durham (ChIRDU) Study, 2019 Annual Meeting, Austin, TX April 10-13
- [10] Merli M G, Verdery A., Mouw T, Li J., «Sampling Migrants from their Social Networks: The demography and Social Organization of Chinese Migrants in Dar es Salaam, Tanzania», *Migr Stud.* 2016 Jul;4(2):182-214.
- [11] Merli et al. 2015. "Challenges To Recruiting Representative Samples Of Female Sex Workers In China Using Respondent Driven Sampling." *Social Science & Medicine* 125.
- [12] Mouw T, Chavez S, Edelblute H, Verdery A., «Binational Social Networks and Assimilation: A Test of the Importance of Transnationalism», *Social Problems* Vol. 61, No. 3 (August 2014), pp. 329-359
- [13] Mouw T, Verdery AM. « Network Sampling with Memory: A proposal for more efficient sampling from social networks». *Sociol Methodol.* 2012 Aug;42(1):206-256.
- [14] Ribeiro, B., P. Wang, and D. Towsley. 2010. On Estimating Degree Distributions of Directed Graphs through Sampling. University of Massachusetts CMPSCI Technical Report UMCS-2010-046.
- [15] Schiltz M-A. Faire et défaire des groupes : L'information chiffrée sur les « populations difficiles à atteindre ». *Bulletin de Méthodologie Sociologique.* 2005;86(1):30-54.