Annexe 7 - Évolutions de la méthode d'échantillonnage

Récapitulatif des modifications

	Sur quoi porte la modification ?	Descriptif
05/11/2020	Méthode échant.	- Ajouter le critère d'avoir au moins un échantillonnable dans son roster pour être sélectionné parmi les 5 nœuds-ponts - Ne sont échantillonnables que les individus dont RSTATUS vaut 1 ou 2.
06/11/2020	Calcul CSR	- Suppression des tirages rejetés dans le calcul des CSR (cumul des SRS)
08/11/2020	Méthode échant.	- Actualiser le RSTATUS des individus immédiatement pour qu'ils n'apparaissent pas comme échantillonnable lors du tirage suivant (appartenant à la même session de tirage)
13/11/2020	Calcul CSR	- Figer les CSR des individus dès lors qu'ils sont échantillonnés
17/11/2020	Calcul CSR	- Recalculer les SRS des candidats au tirage en tenant compte du nombre d'échantillonnables dans le roster au moment du tirage (correspond à la probabilité de tirage au 2^{e} degré)
07/12/2020	Méthode échant.	- Modifier le calcul des cj pour chaque nœud enquêté et donc les indicateurs qui en découlent
18/12/2020	Méthode échant.	- Modification du seuil A1 (de 0.2 à 0.4)
30/03/2021	Réflexion sur la citation de son citant	- Un cité peut-il citer son citant ? Peut-on faire le choix de supprimer cette citation pour donner toutes ses « chances » au roster d'être tiré ?

05/11/2020 - Modification du mode Search, appliquée dès le tirage du 06/11/2020

Méthode initiale (jusqu'au tirage du 3/11/2020)

1/ On sélectionne les premiers 5 nœuds ponts (enquêtés) après les avoir classés sur la variable p_bridge (probabilité d'être un nœud pont)

2/ On sélectionne un individu parmi les individus cités une fois et non enquêtés du nœud-pont sélectionné.

Pourquoi ce changement ?

Je propose de modifier cette méthode car, en situation réelle, elle présente quelques défauts. En effet, Depuis quelques tirages, on a un roster comptant 1 seul individu et cet individu a été échantillonné et est actuellement "A ENQUETER (SANS CONTACT)", mais le roster reste quand même dans les 5 noeuds ponts. Sa présence réduit donc nos capacités de tirage. Il ne sortira du tirage que quand un nouveau roster aura une probabilité plus importante d'être un noeud pont.

Plus généralement, quand on réalise un tirage, on a potentiellement pas mal de tirages que j'appelle "rejetés" car pointent sur des individus déjà échantillonnés mais pas encore enquêtés, ou encore des refus/HC/imp.

Nouvelle méthode (à partir du tirage du 6/11/2020)

1/ Au niveau du tirage au 1er degré : Ajouter un critère pour faire partie des 5 rosters les plus probablement nœuds-ponts retenus pour le tirage : celui de compter au moins 1 individu échantillonnable dans leur roster (échantillonnable = Ligne complete (RSTATUS=1) ou Sans contact (RSTATUS=2)). On sera alors sûr de sélectionner les 5 rosters les plus "nœuds-ponts" mais nous menant à des échantillonnables.

2/ Au niveau du tirage au 2e degré : Ne conserver que les amis de l'enquêté sélectionné cités une seule fois et présentant les modalités 1 ou 2 de RSTATUS.

Questionnements

Que ce soit dans l'ancienne ou la nouvelle méthode de tirage, lors des tirages dits « rejetés » (c'est-à-dire lorsqu'on tire un individu dont le RSTATUS est différent de 1 ou 2), on incrémente le CSR de tous les individus candidats à ce tirage. Je m'interroge sur le bien-fondé de ce choix. Ne devrait-on pas incrémenter le CSR qu'en cas de tirage validé ? Le re-calcul des CSR en excluant les tirages rejetés est possible, et peut-être fait jusqu'à l'entrée dans l'evensampling, mais au-delà, il sera trop tard car l'evensampling tient compte des CSR de chaque individu pour identifier les candidats au tirage.

06/11/2020 – Modification des CSR passés

J'ai donc procédé, comme énoncé précédemment au re-calcul des CSR en soustrayant les SRS des tirages rejetés (qui étaient comptabilisés jusqu'à maintenant. On compte 45 tirages rejetés sur les 117 réalisés). Cela permet effectivement d'homogénéiser les CSR entre les individus (réduction du coefficient de variation et du rapport max/min).

Parmi les individus présentant un CSR important, on compte des individus soumis récemment à l'échantillonnage, notamment des personnes échantillonnées aujourd'hui. Il est normal que leur taux de sondage soit plus élevé en moyenne que ceux ayant été soumis au tirage lors des sessions précédentes car avec la nouvelle méthode, on a moins de « dilution » vu qu'on ne tire plus que parmi les échantillonnables, donc un champ plus restreint. Mécaniquement, le taux de sondage est plus élevé.

Cependant, pour être plus carré et cohérent avec ce nouveau tournant, il faudrait également qu'au sein d'une même session de tirage (au cours de laquelle on tire entre 8 et 10 individus), les nouveaux échantillonnés soient retirés immédiatement du vivier, ce qui n'était pas le cas du tirage du 06/11 et qui a conduit à produire un CSR important sur les individus soumis au tirage aujourd'hui. Avec cette nouvelle modification, dès lors qu'un individu sera échantillonné, son CSR n'évoluera plus car ne sera plus éligible au sondage.

Pour mimer cela sur les tirages passés, nous (avec Aurélie) avons exploré l'option de figer le CSR d'un individu dès lors qu'il est échantillonné. Cependant, cela ne permet pas de tout corriger, puisque les taux de sondage pour les autres personnes soumises au tirage sont incorrects. Au moment du tirage, ces individus en réalité non échantillonnables étaient bien présents dans la base de sondage, ils ont donc impacté le taux de sondage des membres de leur roster. Pour cela, il me semble compliqué a posteriori de corriger tout cela.

Mais peut-être que figer le CSR d'un individu dès lors qu'il est échantillonné peut quand même conduire à corriger un peu le système de CSR et offrir un système de poids plus stable in fine, même s'il ne reflète pas totalement la réalité des choses.

08/11/2020 - Modification du mode Search, appliquée dès le tirage du 09/11/2020

Conformément aux adaptations mentionnées ci-dessus, le tirage en mode search a été révisé une nouvelle fois afin d'exclure les individus du tirage dès qu'ils sont échantillonnés, même au sein d'une même session de tirage. A ce stade, il n'existe donc plus de tirage rejetés. Le tirage se fait parmi les amis échantillonnables des 5 nœuds points comptant au moins 1 échantillonnable, et le statut d'échantillonnable est mis à jour immédiatement, et s'applique dès le tirage suivant.

13/11/2020 – Figer les CSR après échantillonnage

Pour être cohérent avec la modification mise en œuvre à partir du 9/11/2020, nous avons décidé de figer les CSR des personnes dès lors qu'elles sont échantillonnées sur les tirages passés.

16/11/2020 – Modification des SRS passés

L'objectif ici est de recalculer les SRS des individus soumis au tirage en corrigeant leur probabilité de tirage (2e degré) en recalculant le nombre d'individus échantillonnables dans le roster au moment du tirage pour tous les tirages ayant eu lieu avant le 09/11/2020.

Par exemple :

Au tirage 17, le roster 117 est candidat. Il est d'ailleurs sélectionné parmi les 5 nœuds ponts. Le roster compte alors 5 individus, tous échantillonnables.

ID_CITANT	RID	RSTATUS	Proba citant (1er degré)	Proba dans le roster (2e degré)	SRS (produit des deux probas)
117	223	1. Complète	0,9225	0,2 (=1/5)	0,1845
117	224	1. Complète	0,9225	0,2 (=1/5)	0,1845
117	225	1. Complète	0,9225	0,2 (=1/5)	0,1845
117	226	2. Sans contact	0,9225	0,2 (=1/5)	0,1845
117	227	1. Complète	0,9225	0,2 (=1/5)	0,1845

L'individu 223 est échantillonné au tirage 17.

Au tirage 18, ce roster est encore candidat, ses membres sont encore soumis au tirage.

Avec l'ancienne méthode de tirage, les individus déjà échantillonnés restaient candidats au tirage et s'ils étaient de nouveau échantillonnés, cela produisait un tirage rejeté. Dans la nouvelle méthode, les individus échantillonnés sont exclus automatiquement du tirage dès lors qu'ils sont échantillonnés.

Il faut donc corriger la probabilité de tirage dans le roster de ces individus et le SRS qui en découle.

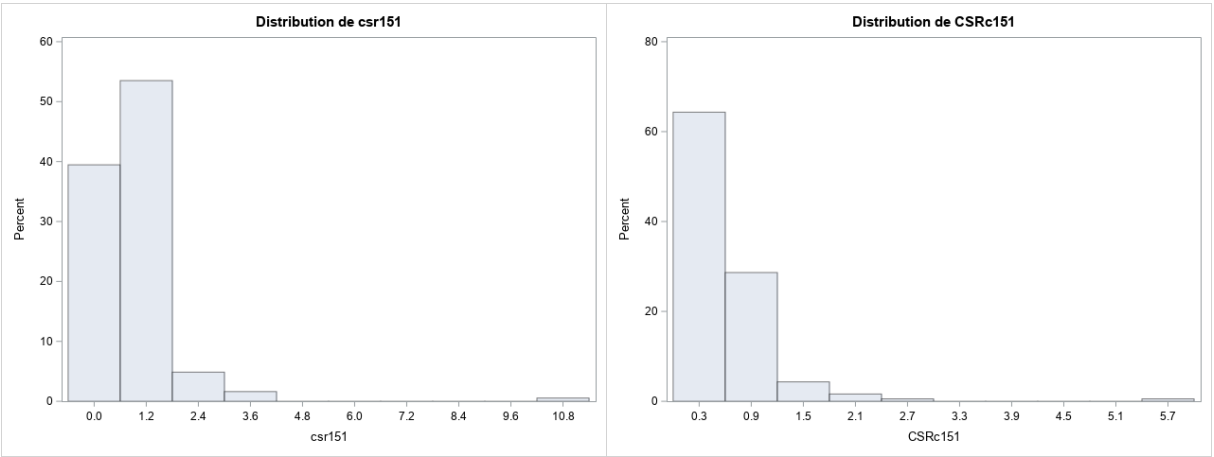
				Avant correction		Après correction	
ID_CITANT	RID	RSTATUS	Proba citant (1er degré)	Proba dans le roster (2e degré)	SRS (produit des deux probas)	Proba dans le roster (2e degré)	SRS (produit des deux probas)
117	223	1. Complète	0,9225	0,2 (=1/5)	0,1845	0	0
117	224	1. Complète	0,9225	0,2 (=1/5)	0,1845	0,25 (=1/4)	0,2306
117	225	1. Complète	0,9225	0,2 (=1/5)	0,1845	0,25 (=1/4)	0,2306

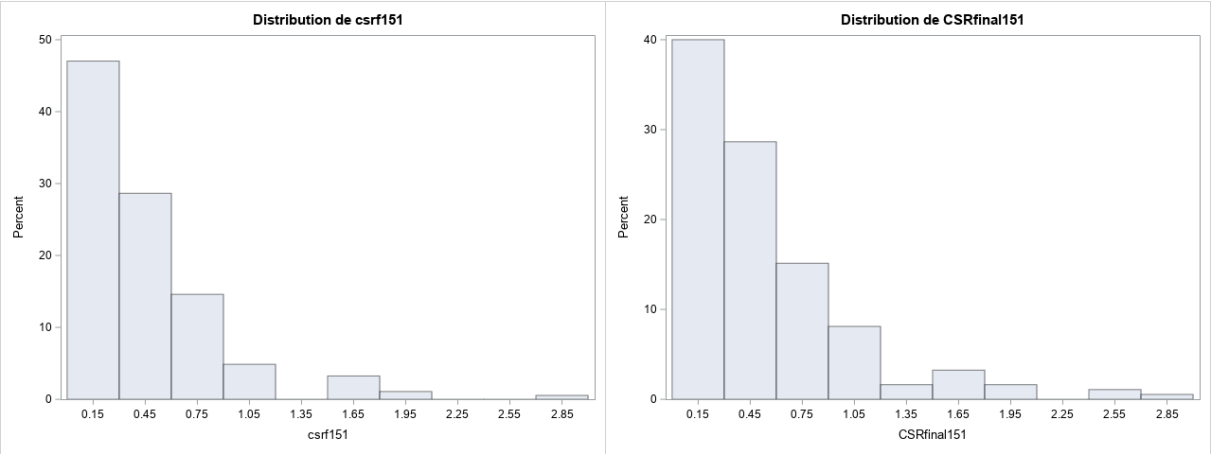
117	226	2. Sans contact	0,9225	0,2 (=1/5)	0,1845	0,25 (=1/4)	0,2306
117	227	1. Complète	0,9225	0,2 (=1/5)	0,1845	0,25 (=1/4)	0,2306

Le SRS des individus restants est donc augmenté. Il faut ensuite recalculer les CSR à chaque tirage.

Voici l'évolution des statistiques des CSR à chaque étape (CSR151 – dernier tirage):

	Moyenne	Coefficient de variation	Max (le min étant toujours 0)
Méthode de calcul initiale	0,82	119,3%	10,87
En soustrayant les SRS des tirages rejetés	0,57	102,7%	5,84
En figeant les CSR des ind. dès qu'ils sont échantillonnés	0,45	95,3%	2,9
En corrigeant les SRS des tirages passés	0,53	95,1%	2,9





Les différentes étapes de correction tendent donc à homogénéiser les CSR, ce qui joue en faveur de la qualité des pondérations qui en découleront.

Entre le CSR initial et le CSR finalement calculé, la valeur est augmentée pour 26 personnes, reste la même pour les 51 individus n'ayant jamais été soumis au tirage, et a diminué pour les 108 restants.

Si ce mode de calcul vous convient, il faudra donc remplacer la base CSR qu'on actualise au fil des tirages pour que ce soit ces nouveaux CSR qui soient désormais incrémentés, et pris en compte au moment de l'evensampling.

07/12/2020 – Modification de la méthode d'échantillonnage

A l'issue du tirage du 04/12/2020, une nouvelle exploration du programme a été faite par Marine & Aurélie car elles avaient décelé un souci sur le traitement des questionnaires « en cours » (qui n'étaient pas pris en compte dans le calcul du cj).

A cette occasion, dans le programme, elles ont repéré qu'il était noté que cj correspondait au nombre d'amis cités 1 fois et non échantillonnés. Or, nous critiquions la méthode SEARCH en se disant qu'au lieu de « non enquêtés », il faudrait utiliser le critère « non échantillonnés ».

Le programme a donc été revu ce jour pour que cj corresponde au nombre d'individus cités par j qui ne sont ni cités par ailleurs, ni échantillonnés.

Jusqu'à présent la méthode employée a contribué à surexploiter les rosters contenant déjà des échantillonnés. Les rosters avec des échantillonnés n'ont pas été disqualifiés à tort, entrainant un manque de diversité dans les rosters exploités, vu que le critère d'être un nœud pont ne dépendait pas du nombre d'individus échantillonnés dans le roster, mais juste du nombre

d'individus enquêtés. Certains rosters n'ont donc pas été exploités à tort. Peut-être que la correction réalisée ce jour va permettre à d'anciens rosters d'être exploités

Lors de cette révision, nous nous sommes interrogées de nouveau sur le calcul de P1. Jusqu'à présent, l'indicateur P1, utilisé pour déterminer le mode d'échantillonnage à appliquer, était calculé comme suit : nombre d'individus cités une fois & non enquêtés rapporté à la taille totale du réseau (nombre de RID sans doublon). En parcourant l'article de Ted, il est noté que P1 correspond au nombre d'individus cités une seule fois sur la taille totale du réseau. Dans cette configuration, les « non-enquêtés » ne sont pas exclus du numérateur. Cette modification a pour conséquence de remonter considérablement le niveau de P1. Il passe de 63% à 95%. Cela a pour conséquence de maintenir le mode SEARCH encore très longtemps (le seuil pour passer en evensampling est fixé à 0,2 et peut-être élevé à 0,4). Par ailleurs, en note de fin d'article, une nouvelle définition de P1 est fournie : il est indiqué que les individus échantillonnés sont considérés comme des individus cités plusieurs fois, ce qui signifie qu'au numérateur, ne sont comptabilisés que les individus cités 1 fois et non échantillonnés.

Compte tenu des multiples définitions de P1 présentes dans l'article, nous avons décidé de maintenir notre calcul de P1 en plaçant au numérateur le nombre d'individus cités une fois et non enquêtés, qui nous semble une définition intermédiaire aux deux définitions de Ted.

Nous suspectons une ambiguïté dans le discours de Ted. La frontière entre « sampled » et « interviewed » n'est pas claire étant donné que l'article se base sur des simulations sur des réseaux existants (facebook) et la non-réponse est donc absente du processus. Dans son cas, sampled=Interviewed.

Davantage d'informations sur la méthode en conditions réelles (CHIRDU par exemple) auraient vraiment été bénéfiques.

18/12/2020 – Modification de la méthode d'échantillonnage

Lors d'un point avec l'équipe CHIRDU, la question du seuil A1 a été posé à Ted. Dans CHIRDU, ils ne sont jamais passés en EvenSampling. Selon lui, le seuil A1 à 0.4 est plus « realistic » qu'à 0.2.

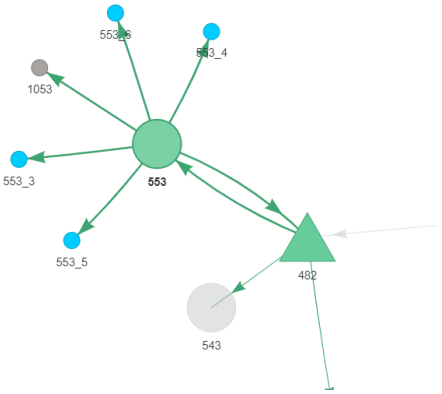
Le programme d'échantillonnage a donc été modifié dans ce sens.

A ce jour, P1 est à 0.60. Le passage en EvenSampling est donc envisageable dans l'enquête ChIPRe.

30/03/2021 – Un cité peut-il citer son citant ?

Nous réalisons qu'aucune consigne claire n'a été donnée aux enquêteurs concernant la possibilité, pour un échantillonné, de donner le nom de son propre citant. C'est un point délicat étant donné que le cité n'est pas toujours au courant de l'identité de son citant (parfois même de la volonté du citant).

Ce problème a émergé quand nous avons constaté qu'un enquêté (553) comptait parmi son roster de 6L son propre citant (482). A une période où nous apprécierions échantillonner de nouveaux rosters et ne pas « taper dans les vieux », nous constatons que cette citation de citant diminue fortement pour ce roster ses chances d'être candidat à l'échantillonnage. Plus généralement, la citation du citant semble reposer seulement sur la compréhension de la consigne par les enquêtés et non pas sur le réel lien entre eux. Autrement dit, il semble que beaucoup de gens pourraient citer leur citant mais ne le font pas forcément parce qu'ils jugent cela inutile, donc l'absence du citant dans les cités ne voudra pas nécessairement dire que leur lien n'est pas réciproque.



A donc été discutée la possibilité pour nous de supprimer manuellement cette citation, qui n'apporte *a priori* rien et pénalise inutilement ce roster.

Dans l'idéal, il aurait fallu se fixer une règle dès le départ pour ces cas : si les gens sont au courant de l'identité de leur citant, leur préciser qu'il ne faut pas le citer en retour (quitte à avoir un roster vide), sinon supprimer manuellement ces cas.

En effet, nous constatons au fil de la collecte que cette info (avoir cité son citant) est à la fois difficile à interpréter et crée de la confusion quant à l'exploration du réseau :

- l'absence de cette info (citer son citant) ne témoigne finalement pas du lien réel entre cité et citant. Dans l'analyse il sera impossible de conclure que tous les enquêtés qui n'ont pas cité leur citant à leur tour ont un lieu faible avec lui ! il nous sera donc difficile d'analyser la réciprocité des liens
- la présence de cette info crée des doublons qui nous semblent en partie « artificiels » (à moins de considérer que citer son citant revient à « avoir un réseau limité » et que ce type de doublon peut participer légitimement à la construction de l'indicateur de « saturation » du réseau, notre P1)

Actuellement on compte 20 cas de liens réciproques (soit 40 liens au total).