



N° 11-522-XIF au catalogue

**La série des symposiums internationaux
de Statistique Canada - Recueil**

**Symposium 2004 : Méthodes
innovatrices pour enquêter
auprès des populations
difficiles à joindre**

2004



ÉCHANTILLONNAGE EN RÉSEAU SELON UNE APPROCHE BAYÉSIENNE

Mosuk Chow et Steven K. Thompson¹

RÉSUMÉ

Dans le cas des populations difficiles à joindre, l'échantillonnage en réseau est parfois le seul mode pratique d'obtention d'un échantillon assez nombreux pour une étude efficace. Dans le présent document, nous examinons le problème de l'estimation des propriétés de réseau social à partir d'échantillons. Nous considérons un plan d'échantillonnage par dépistage de liens. Nous décrivons le cadre d'une approche bayésienne pour l'estimation des propriétés de réseau social lorsque la population et sa structure de réseau social est modélisée selon un graphe stochastique. Nous présentons un programme Mathematica et calculons les estimations de Bayes en prenant un exemple.

MOTS CLÉS : Approche fondée sur un modèle, distributions a priori bêta, échantillonnage en réseau, estimation bayésienne, plans d'échantillonnage par dépistage de liens.

1. INTRODUCTION

1.1 Contexte

On étudie couramment les liens qui unissent les gens ou d'autres entités sociales en termes de réseaux. Les données de réseau social comportent des mesures à propos des relations entre les gens ou d'autres entités sociales, de même que des mesures à propos des entités mêmes. Dans les plans d'échantillonnage par dépistage de liens, les liens sociaux sont suivis d'un répondant à l'autre afin d'obtenir l'échantillon. Au départ, on a utilisé la méthode d'échantillonnage en réseau, ce traitement étant inévitable lorsqu'il était difficile de définir des règles de dénombrement unitaire, les populations d'intérêt étant rattachées à une multitude d'établissements. Dans les années 1960, une grande enquête d'un fournisseur médical visant à estimer la prévalence d'une maladie mortelle, la fibrose kystique, dans trois états de la Nouvelle-Angleterre (Kramm et coll., 1962) a été menée. La fibrose kystique est une maladie grave et les gens qui en étaient atteints se rendaient habituellement dans plus d'un centre médical pour y être traités. Au National Center of Health Statistics (NCHS), Birnbaum et Sirken ont proposé plusieurs estimateurs sans biais, dont l'estimateur de multiplicité largement utilisé par la suite et les estimateurs de réseau de Horvitz-Thompson, pour résoudre le problème décrit par Birnbaum et Sirken (1965). C'est ainsi qu'on a reconnu les possibilités qu'offre l'échantillonnage en réseau d'accroître le « rendement » d'enquêtes-ménages portant sur des caractéristiques rares et d'obtenir des estimations entachées d'erreurs d'échantillonnage inférieures à celles des méthodes classiques. Ainsi, Nathan (1976) a utilisé les règles de multiplicité pour estimer le nombre de naissances en Israël. Dans le cas de populations humaines cachées et difficiles d'accès, de tels plans de sondage sont parfois le seul mode pratique afin d'obtenir un échantillon assez nombreux pour une étude efficace. On peut citer l'exemple des études de comportements humains risqués. Thompson et Collins (2001) ont formulé la méthode de l'échantillonnage adapté sous forme de graphe ou dans un cadre spatial pour la recherche sur les comportements à risque.

¹ Mosuk Chow, Département de statistique, Université d'État de Pennsylvanie, 326, immeuble Thomas, University Park, PA 16802, États-Unis; Steven K. Thompson, Département de statistique, Université d'État de Pennsylvanie, 326, immeuble Thomas, University Park, PA 16802, États-Unis, et Los Alamos National Laboratory, Los Alamos, NM 87544, États-Unis.

Les plans d'échantillonnage par dépistage de liens se présentent sous bien des formes. Un de ces plans qui est fréquemment appliqué est celui de l'échantillonnage boule de neige, comme l'a baptisé Goodman (1961). C'est là un plan de sondage où on demande aux gens faisant partie d'un échantillon initial d'identifier des gens qu'ils connaissent; ceux-ci font de même à leur tour et la démarche se poursuit un nombre déterminé de fois. Il importe de pouvoir estimer les propriétés de réseau à partir d'échantillons. Snijders (1992) a employé le terme « échantillonnage boule de neige » pour inclure les plans où on fait le suivi seulement d'un sous-échantillon de liens à partir de chaque nœud. Frank et Snijders (1994) considèrent l'estimation selon le plan et l'estimation fondée sur un modèle d'une taille de population cachée à l'aide de l'échantillonnage boule de neige. Thompson et Frank (2000) utilisent une approche fondée sur un modèle pour effectuer des inférences avec des plans d'échantillonnage par dépistage de liens. Dans leur étude, ils calculent les fonctions de vraisemblance par un modèle tant symétrique qu'asymétrique, puis appliquent des estimateurs de maximum de vraisemblance aux paramètres de modélisation et aux prédicteurs des quantités réalisées de population en graphe. Ils comparent enfin l'estimateur de maximum de vraisemblance aux estimateurs classiques et aux statistiques sommaires.

1.2 Approche bayésienne à l'analyse de réseau

Ces derniers temps, il y a eu beaucoup d'intérêt à utiliser l'approche bayésienne pour l'analyse de réseau. Nous mentionnons ici quelques-unes des études consacrées à la question. Wong (1987) prend le modèle p_1 proposé par Holland et Leinhardt (1981) pour analyser les graphes orientés des études de réseaux. Dans cette étude, une approche bayésienne qui utilise une distribution a priori normale échangeable pour les paramètres représentant les caractéristiques d'attraction et d'extension des nœuds est proposée. Snijders et Nowicki (1997) recommandent pour leur part diverses méthodes statistiques, dont une approche bayésienne, à des fins d'estimation et de prévision par modèle stochastique en blocs pour des graphes à structure latente de blocs. Plus récemment, Nowicki et Snijders (2001) proposent une méthode statistique applicable à une modélisation a posteriori en blocs pour les graphes orientés. Le modèle probabiliste suppose que les sommets du graphe orienté se divisent en plusieurs classes inobservées et que les probabilités de relation entre deux sommets dépendent seulement des catégories auxquelles ils appartiennent. Un estimateur bayésien fondé sur un échantillonnage de Gibbs est proposé dans l'étude. Dans Hoff, Raftery et Handcock (2002), on développe une catégorie de modèles où la probabilité d'une relation entre acteurs dépend des positions respectives des individus dans un « espace social » inobservé. Les inférences relatives à l'espace social se font dans des cadres d'estimation de maximum de vraisemblance et d'estimation bayésienne. On propose les méthodes de Monte Carlo à chaîne de Markov pour les inférences relatives aux positions latentes et aux effets des covariables observées.

1.3 Approche bayésienne dans l'échantillonnage par dépistage de liens

Comme nous l'avons mentionné à la section 1.1, Thompson et Frank (2000) calculent la fonction de vraisemblance d'un modèle graphique tant symétrique qu'asymétrique. Dans Chow et Thompson (2003), on propose une approche bayésienne d'estimation et de prévision lorsque la population et sa structure de réseau social sont modélisées sous forme de graphe stochastique. Nous présentons ici un programme Mathematica pour l'évaluation du cas présenté dans cet article. Nous voulons démontrer que la méthodologie est accessible au praticien et que les calculs sont simples et rapides. Pour le traitement de tels problèmes, l'approche bayésienne peut largement être adoptée.

2. NOTATION

2.1 Fonction de vraisemblance d'un modèle en graphe entier

Pour un graphe de N nœuds désignés $1, 2, \dots, N$ et la variable d'intérêt Y_u liée au u^e nœud, nous notons tout l'ensemble de nœuds par $U = \{1, 2, \dots, N\}$ et $Y = (Y_1, Y_2, \dots, Y_N)$. La variable indicatrice A_{uv} est égale à un s'il existe une relation directionnelle entre u et v et à zéro dans tout autre cas où u et v sont des nœuds distincts. Comme dans Chow et Thompson (2003), nous supposons que Y_1, Y_2, \dots, Y_N sont des variables aléatoires de Bernoulli indépendantes et identiquement distribuées (i.i.d), que $P(Y_u = 0) = \theta_0$ et que $P(Y_u = 1) = \theta_1$ avec $\theta_0 + \theta_1 = 1$. Conditionnellement aux valeurs de nœud (Y_1, Y_2, \dots, Y_N) , les dyades (A_{uv}, A_{vu}) sont indépendantes pour $1 \leq u \leq v \leq N$. La probabilité conditionnelle est donnée par $P[(A_{uv}, A_{vu}) = (k, l) \mid Y_u = i, Y_v = j] = \lambda_{ijkl}$ pour toutes les combinaisons de

$i=0,1, j=0,1, k=0,1$ et $l=0,1$. Pour toutes les combinaisons de i et j , les sommations sur k et l sont désignées par $\lambda_{ij\cdot}$ et sont égales à un. Pour obtenir des probabilités de graphe qui ne dépendent pas des identités des nœuds, nous devons poser les conditions suivantes de symétrie naturelle : $\lambda_{1110} = \lambda_{1101}, \lambda_{1011} = \lambda_{0111}, \lambda_{1010} = \lambda_{0101}, \lambda_{1001} = \lambda_{0110}, \lambda_{0010} = \lambda_{0001}$ et $\lambda_{1000} = \lambda_{0100}$. Soit N_i désignant le nombre total de nœuds à valeur i dans le graphe, de sorte que $N_0 + N_1 = N$. Désignons par M_{ijkl} le nombre total de dyades du type (i,j,k,l) . Dans ce cas, la fonction de vraisemblance du modèle en graphe entier ayant les paramètres (θ, λ) est

$$L(\theta, \lambda; Y, A) = \left(\prod_i \theta_i^{N_i} \right) \left(\prod_i \prod_j \prod_k \prod_l \lambda_{ijkl}^{M_{ijkl}} \right).$$

2.2 Fonction de vraisemblance avec les données d'échantillon du modèle symétrique entier

La méthodologie bayésienne est démontrée pour le modèle symétrique entier. Dans ce modèle, $\lambda_{ijkl} = 0$ pour $k \neq l$, $\lambda_{ijkk} = \lambda_{ijkk}$ pour $i,j,k=0,1$ et $\lambda_{ij00} + \lambda_{ij11} = 1$. Employons $\beta_{i+j} = \lambda_{ij11}$ et, ainsi, β_k désigne la probabilité de relation bidirectionnelle entre deux nœuds ayant une valeur totale k , où k peut prendre les valeurs 0, 1 ou 2. Considérons le plan d'échantillonnage par dépistage de liens où un échantillon initial s_0 est sélectionné et où tous les liens de s_0 sont suivis pour ajouter l'ensemble s_1 de nœuds qui ne se trouvent pas en s_0 , mais qui sont adjacents à des nœuds de s_0 . L'échantillon entier est $s = s_0 \cup s_1$. On peut exprimer tout l'ensemble de nœuds de la population comme l'union de trois ensembles disjoints $U = s_0 \cup s_1 \cup \bar{s}$, où \bar{s} désigne les nœuds non échantillonnés. Dans le présent document comme dans Chow et Thompson (2003), le plan où la décision de suivre les liens à partir du nœud u dépend de la valeur de nœud y_u . Il peut ainsi se formuler : $P(s|y_s, a_{s_0U})$. Les données sont $d = (s, y_s, a_{s_0U})$. Comme la décision dépend des valeurs de y et a seulement par les données observées, le plan se factorise hors de la fonction de vraisemblance et se divise hors de la distribution a posteriori de Bayes, si bien que l'inférence de Bayes dépend seulement du modèle posé. Soit $n_i(s)$, $n_i(s_0)$ et $n_i(\bar{s})$ désignant les nombres respectifs de nœuds du type i dans l'échantillon entier s , l'échantillon initial s_0 et les nœuds non échantillonnés \bar{s} . Soit r_{ij} désignant les comptes de dyades où le premier indice représente la somme des valeurs de nœud et le second, la somme des valeurs de lien. La fonction de vraisemblance est donnée par :

$$L(\theta, \beta; d) = P(s|y_s, a_{s_0U}) \theta_0^{n_0(s)} (1-\theta_0)^{n_1(s)} \beta_0^{r_{0,0}} (1-\beta_0)^{r_{0,1}} \beta_1^{r_{1,2}} (1-\beta_1)^{r_{1,0}} \beta_2^{r_{2,2}} (1-\beta_2)^{r_{2,0}} [\theta_0 (1-\beta_0)^{n_0(s_0)} (1-\beta_1)^{n_1(s_0)} + (1-\theta_0)(1-\beta_1)^{n_0(s_0)} (1-\beta_2)^{n_1(s_0)}]^{n(\bar{s})}$$

3. ESTIMATEURS DE BAYES

3.1 Distributions a priori des paramètres

Comme les paramètres $\theta_0, \beta_0, \beta_1, \beta_2$ prennent des valeurs entre 0 et 1, il est tout à fait courant qu'on leur applique une distribution a priori bêta. Ajoutons que, comme il n'y a pas de contraintes spécifiques sur ces paramètres, on peut poser qu'il s'agit de distributions a priori indépendantes. À noter que les distributions a priori bêta forment une catégorie plutôt riche de distributions pouvant donner une bonne approximation de la plupart des distributions unimodales sur l'intervalle $[0,1]$. La distribution a priori est donnée par :

$$\pi(\theta_0, \beta_0, \beta_1, \beta_2) \propto \theta_0^{a-1} (1-\theta_0)^{b-1} \beta_0^{c-1} (1-\beta_0)^{d-1} \beta_1^{e-1} (1-\beta_1)^{f-1} \beta_2^{g-1} (1-\beta_2)^{h-1}.$$

S'il y a connaissance antérieure de la moyenne et de la variance de ces paramètres, on peut résoudre pour les valeurs de a, b, c, d, e, f, g et h en posant les valeurs correspondantes égales à la moyenne d'une distribution bêta $(a, b) = a/(a+b)$ et à la variance d'une distribution bêta $(a, b) = ab/((a+b)^2(a+b+1))$. Dans bien des cas, la

réalisation d'une étude pilote peut aider à avancer les valeurs de a, b, c, d, e, f, g et h . En cas d'absence d'information, on peut songer à utiliser des distributions a priori non informatives.

3.2 Distribution a posteriori

En analyse bayésienne, on spécifie non seulement la fonction de vraisemblance, mais aussi une distribution a priori pour les paramètres inconnus. L'inférence relative à ces paramètres est alors fondée sur la distribution a posteriori. Pour notre problème et la distribution a priori que nous avons spécifiée, la distribution a posteriori est la suivante :

$$\begin{aligned} \pi(\theta_0, \beta_0, \beta_1, \beta_2 | d) &\propto \theta_0^{n_0(s)+a-1} (1-\theta_0)^{n_1(s)+b-1} \beta_0^{r_{0,2}+c-1} (1-\beta_0)^{r_{0,0}+d-1} \\ &\quad \beta_1^{r_{1,2}+e-1} (1-\beta_1)^{r_{1,0}+f-1} \beta_2^{r_{2,2}+g-1} (1-\beta_2)^{r_{2,0}+h-1} \\ &\quad [\theta_0 (1-\beta_0)^{n_0(s_0)} (1-\beta_1)^{n_1(s_0)} + (1-\theta_0)(1-\beta_1)^{n_0(s_0)} (1-\beta_2)^{n_1(s_0)}]^{n(\bar{s})} . \end{aligned}$$

La distribution a posteriori étant établie, les estimations de Bayes de chaque paramètre sont la moyenne a posteriori et peuvent facilement se trouver. Pour établir la moyenne a posteriori de θ_0 par exemple, on prend :

$$\begin{aligned} q(\theta_0, \beta_0, \beta_1, \beta_2) &= \theta_0^{n_0(s)+a-1} (1-\theta_0)^{n_1(s)+b-1} \beta_0^{r_{0,2}+c-1} (1-\beta_0)^{r_{0,0}+d-1} \beta_1^{r_{1,2}+e-1} (1-\beta_1)^{r_{1,0}+f-1} \beta_2^{r_{2,2}+g-1} (1-\beta_2)^{r_{2,0}+h-1} \\ &\quad \beta_2^{r_{2,2}} (1-\beta_2)^{r_{2,0}} [\theta_0 (1-\beta_0)^{n_0(s_0)} (1-\beta_1)^{n_1(s_0)} + (1-\theta_0)(1-\beta_1)^{n_0(s_0)} (1-\beta_2)^{n_1(s_0)}]^{n(\bar{s})} . \end{aligned}$$

Comme $\int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx = B(\alpha, \beta)$ est la fonction bêta, nous pouvons obtenir ce qui suit :

$$\begin{aligned} M_1 &= \int_0^1 \int_0^1 \int_0^1 \int_0^1 q(\theta_0, \beta_0, \beta_1, \beta_2) d\theta_0 d\beta_0 d\beta_1 d\beta_2 \\ &= \sum_{i=0}^{n(\bar{s})} \binom{n(\bar{s})}{i} B(n_0(s) + a + i, n(\bar{s}) + n_1(s) + b - i) B(r_{0,2} + c, i n_0(s_0) + r_{0,0} + d) \\ &\quad B(r_{1,2} + e, i n_1(s_0) + (n(\bar{s}) - i) n_0(s_0) + f) B(r_{2,2} + g, (n(\bar{s}) - i) n_1(s_0) + h) . \end{aligned}$$

Nous obtenons aussi :

$$\begin{aligned} M_2 &= \int_0^1 \int_0^1 \int_0^1 \int_0^1 \theta_0 q(\theta_0, \beta_0, \beta_1, \beta_2) d\theta_0 d\beta_0 d\beta_1 d\beta_2 \\ &= \sum_{i=0}^{n(\bar{s})} \binom{n(\bar{s})}{i} B(n_0(s) + a + 1 + i, n(\bar{s}) + n_1(s) + b - i) B(r_{0,2} + c, i n_0(s_0) + r_{0,0} + d) \\ &\quad B(r_{1,2} + e, i n_1(s_0) + (n(\bar{s}) - i) n_0(s_0) + f) B(r_{2,2} + g, (n(\bar{s}) - i) n_1(s_0) + h) . \end{aligned}$$

L'estimation de Bayes de θ_0 est simplement :

$$E(\theta_0 | d) = \frac{M_2}{M_1} .$$

3.3 Programme Mathematica pour l'évaluation de l'estimation de Bayes

Bien que les distributions a priori bêta ne soient pas des distributions a priori conjuguées pour la fonction de vraisemblance, elles produisent bel et bien des moyennes a posteriori qui sont relativement faciles à évaluer. Nous allons en donner un exemple ici avec un programme Mathematica démontrant la facilité du calcul des estimations de Bayes. Considérons l'estimation en pourcentage des consommateurs de drogues injectables dans un certain milieu.

Pour la commodité de l'exposé, soit θ_0 désignant la proportion de non-consommateurs de ces drogues dans le milieu et $1 - \theta_0$, la proportion de consommateurs. Pour mettre en corrélation cet exemple et l'étude de Chow et Thompson (2003), nous emploierons les mêmes valeurs numériques. Dans une collectivité de 200 personnes, on échantillonne au hasard 22 unités sans remise, 5 étant des consommateurs et 17, des non-consommateurs. On demande aux consommateurs de nommer ceux avec qui ils prennent des drogues injectables. Comme des liens sont possibles seulement entre consommateurs, le suivi des liens ne peut qu'ajouter des consommateurs à l'échantillon. L'échantillon initial fait 12 renvois au total, dont 10 ne figurent pas dans l'échantillon de départ. Ainsi,

$$n_1(s_0) = 5, \quad n_0(s_0) = 17, \quad n_1(s) = 15, \quad n_0(s) = 17, \quad n(\bar{s}) = 168,$$

$$r_{22} = 12, \quad r_{20} = 93.$$

Il convient de noter que β_0 est la probabilité d'un lien bidirectionnel entre deux unités de consommation de drogues injectables et β_1 , la probabilité d'un tel lien entre une unité de consommation et une unité de non-consommation. Ces deux probabilités devraient être égales à zéro. Dans ce cas, les deux paramètres inconnus qui restent sont seulement θ_0 et β_2 . Dans notre exemple, β_2 est la probabilité d'une relation bidirectionnelle entre deux consommateurs. Dans le programme Mathematica que nous présentons, nous calculons les estimations de Bayes de θ_0 et β_2 correspondant à la distribution a priori non informative $a = b = g = h = 1/2$. On peut également démontrer que les estimations de Bayes correspondant aux deux autres distributions a priori non informatives $a = b = g = h = 0$ et $a = b = g = h = 1$ ont des valeurs semblables à celles qui sont calculées ci-après. On peut en conclure que, pour cet ensemble de données, les estimations de Bayes ne sont pas sensibles à la spécification de ces trois distributions a priori.

Programme Mathematica pour l'évaluation des estimations de Bayes de θ_0 et β_2 en cas de distribution a priori non informative avec $a = b = g = h = 1/2$

```

n1s0=5;
n0s0=17;
n1s = 15;
n0s = 17;
nsbar = 168;
r22=12;
r20=93;
a=0.5;
b=0.5;
g=0.5;
h=0.5;
AA= n0s + a;
BB= n1s + b;
GG= r22 +g;
HH= r20 + h;
k = nsbar;
sumzi = Sum[Binomial[k,i]Beta[AA + i,k + BB - i]Beta[GG,HH + (k - i)n1s0];
sumtheta = Sum[Binomial[k,i]Beta[AA + 1 + i,k + BB - i]Beta[GG,HH + (k - i)n1s0];

```

$$\begin{aligned}
\text{thetabayes} &= \text{sumtheta} / \text{sumzi} \quad // N \\
\text{sumbeta} &= \sum_{i=0}^k \text{Binomial}[k, i] \text{Beta}[AA + i, k + BB - i] \text{Beta}[GG + 1, HH + (k - i)nl0]; \\
\text{betabayes} &= \text{sumbeta} / \text{sumzi} \quad // N
\end{aligned}$$

Dans ces calculs, on obtient l'estimation de Bayes de θ_0 qui est appelée « thetabayes » dans le programme. On trouve également l'estimation de Bayes de β_2 appelée « betabayes ». Pour cet ensemble de données, le résultat est le suivant :

$$\text{thetabayes} = .728465, \quad \text{betabayes} = .0438741$$

4. CONCLUSION

Dans cet exposé, nous présentons le programme Mathematica choisi pour notre exemple et montrons la simplicité et la rapidité du calcul des estimations de Bayes. Dans des problèmes avec échantillonnage par dépistage de liens, il arrive assez souvent qu'on dispose déjà de renseignements sur la caractéristique à estimer. L'intégration de ces renseignements est facile dans l'approche bayésienne et aussi avec le programme Mathematica. On n'a qu'à changer les valeurs de a , b , g et h dans le programme en fonction de cette connaissance antérieure. Précisons enfin que, dans le traitement bayésien, l'obtention d'estimateurs d'intervalles et l'évaluation de la précision des estimateurs peuvent se faire sans plus de difficulté.

RÉFÉRENCES

- Birnbaum, Z.W. et Sirken, M.G. (1965), "Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates", *Vital and Health Statistics*, PHS Publication No1, Series 2, No.11, US Government Printing Office, Washington.
- Chow, M. et Thompson, S. K. (2003), "Estimation avec plans d'échantillonnage par dépistage de liens – une approche bayésienne", *Techniques d'enquête*, 29, n° 2, pp. 197-205.
- Frank, O. et Snijders, T.A.B. (1994), "Estimating the Size of Hidden Populations Using Snowball Sampling", *Journal of Official Statistics*, 10, pp. 53-67.
- Goodman, L.A. (1961), "Snowball Sampling", *Annals of Mathematical Statistics*, 20, pp. 572-579.
- Hoff, P.D., Raftery, A.E., et Handcock, M.S. (2002), "Latent Space Approaches to Social Network analysis", *Journal of the American Statistical Association*, 97, pp. 1090-1098.
- Holland, P.W. et Leinhardt, S. (1981), "An Exponential Family of Probability Distributions for Directed Graphs", *Journal of the American Statistical Association*, 76, pp. 33-65.
- Kramm, E.R., Crance, M.M., Sirken, M.G., et Brown, M.L. (1962), "A Cystic Fibrosis Pilot Survey in Three New England States", *American Journal of Public Health*, 52, pp. 2041-2057.
- Nathan, G. (1976), "An Empirical Study of Response and Sampling Errors for Multiplicity Estimates with Different Counting Rules", *Journal of the American Statistical Association*, 71, pp. 803-815.
- Nowicki, K. et Snijders, T.A.B. (2001). "Estimation and Prediction for Stochastic Block-structures", *Journal of the American Statistical Association*, 96, pp. 1077-1087.
- Snijders, T.A.B. (1992). "Estimation on the Basis of Snowball Samples: how to weight", *Bulletin de méthodologie sociologique*, 36, pp. 59-70.

- Snijders, T.A.B. et Nowicki, K. (1997). "Estimation and Prediction for Stochastic Blockmodels for graphs with latent block structure", *Journal of Classification*, 14, pp. 75-100.
- Thompson, S.K. et Collins, L.M. (2001). "Adaptive Sampling in Research on Risk-related Behaviors", *Drug and Alcohol Dependence*, 68, pp. S57-S67.
- Thompson, S. K. et Frank, O. (2000), "Estimation fondée sur un modèle et comportant des plans d'échantillonnage à dépistage de liens", *Techniques d'enquête*, 26, n° 1, pp. 87-98.
- Wong, G.Y. (1987). "Bayesian Models for Directed Graphs", *Journal of the American Statistical Association*, 82, pp. 140-148.