Analyse de graphe et modèles statistiques
Du modèle p_1 à l'ERGM

Laurent Beauguitte, CNRS, UMR IDEES
beauguittelaurent<at>hotmail.com

Mai 2012 - Version 1

groupe f.m.r.

Introduction

Une grande partie des méthodes utilisées en analyse de réseaux, et ce quel que soit le champ disciplinaire, est descriptive et s'appuie sur des mesures intuitives (évaluer l'importance *via* le degré, la popularité *via* le degré entrant...). Une autre approche consiste à comparer un réseau donné, ou plus précisément certains indicateurs relatifs à ce réseau, aux résultats d'un modèle statistique théorique.

Tous les modèles présentés dans cette synthèse du groupe fmr (flux, matrices, réseaux) ont été développés en *Social network analysis* depuis une trentaine d'années et, à notre connaissance, les autres disciplines ont peu testé ces méthodes. Le nombre de chercheur-e-s francophones ayant testé ces méthodes est plus réduit encore (voir cependant les travaux de Lazéga [8] ou de Éloire en sociologie [4], Lemercier en histoire[10] ou Boschet et Rambonilaza en économie spatiale [2]).

La structure du document suit l'ordre chronologique d'apparition de ces modèles statistiques probabilistes, du modèle dit p_1 aux modèles ERGM (*Exponential Random Graph Model*) et Siena (*Simulation investigation for empirical network analysis*). De courts programmes en annexe permettent de tester ces deux derniers modèles avec le logiciel R.

1 Le modèle p_1

Holland et Reinhardt développent dès la fin des années 70 un modèle statistique d'analyse de réseau [7] où l'unité d'analyse est la dyade pour laquelle quatre modalités sont possibles (lien mutuel, lien de i vers j, lien de j vers i et absence de lien). Toutes les relations entre acteurs y sont supposées

statistiquement indépendantes les unes des autres. Dans le modèle p_1, la probabilité de distribution pour une dyade D_{ij} est égale à

$$Pr(Y_{ij} = 1) = exp\lambda_{ij} + k\alpha_i + k\beta_j + l\alpha_j + l\beta_i + (k + l)\theta + kl\rho$$

où λ_{ij} permet de contrôler que la somme des probabilités pour la dyade est égale à 1.

Le sens des paramètres est le suivant : α_i et α_j désignent les degrés sortants des sommets i and j, β_i et β_j les degrés entrants, enfin θ, l et ρ désignent la tendance aux relations réciproques.

Pour l'exprimer de façon verbale, le modèle estime la probabilité qu'un lien existe entre deux sommets en se basant sur les critères suivants : densité, nombre de liens mutuels, degrés entrants et sortants des deux acteurs étudiés.

Ce modèle, s'il est à la base des modèles développés actuellement, n'est plus utilisé aujourd'hui pour deux raisons principales. Tout d'abord, le nombre de paramètres à estimer est au minimum égal au double du nombre de sommets, ce qui devient vite problématique. Ensuite, ce modèle ne permet pas la prise en compte d'effets triadiques (transitivité notamment) [1].

2 Les généralisations du modèle p_1

2.1 Le modèle p_2

Le modèle p_2 permet l'analyse des graphes booléens orientés. Il prend en compte les degrés (entrants et sortants), la densité et les effets de réciprocité (dyades mutuelles ou asymétriques). Il permet également d'introduire des variables décrivant les sommets (genre, âge etc.). Son avantage principal par rapport au modèle p_1, outre cette possibilité d'inclure des attributs comme variables explicatives, est qu'il diminue de façon très nette le nombre de paramètres à estimer : en effet, là où p_1 demande deux paramètres de degré par sommet, p_2 utilise une régression linéaire pour estimer ces paramètres. Par contre, les effets triadiques ne peuvent être pris en compte.

Ce modèle a été utilisé notamment par Lazéga et Van Duijn [9] pour expliquer comment fonctionnaient les demandes de conseil au sein de trois cabinets d'avocats nord-américains (71 individus au total). Les variables prises en compte dans le modèle sont les suivantes : statut (partenaires et associés), ancienneté (trois modalités), genre, cabinet (trois modalités), spécialités (deux modalités) et écoles (trois modalités). Le modèle testé suppose évidemment de construire des hypothèses plausibles (ex. on ne demande pas conseil à un plus jeune ou à quelqu'un formé dans une université moins prestigieuse).

L'extrait du tableau 1, tiré de l'article, montre le modèle et les effets retenus. Seuls les effets significatifs sont présents. Cela permet de montrer

1. Pour plus de précisions sur ce modèle, voir notamment [20] et [5]. Le modèle p_1 est le seul modèle statistique probabiliste implémenté dans le logiciel Ucinet.

Tableau 1 – Modèle p^2 d'après Lazéga et Van Duijn

	Parameter	Empty model	Final model
Sender	Variance σ_A^2	0.58	0.75
	Partner seniority Level 1		-0.92
Receiver	Variance σ_B^2	0.76	0.49
	Associate seniority level		-0.50
Density	μ	-1.87 (0.12)	-3.98
	Similarity status		0.89
	Superiority seniority		-0.29
	Similarity office		1.79
	Similarity specialty		1.60
	Similarity gender		0.29
	Similarity lawschool		0.20

que seule l'ancienneté agit, et de façon négative (signe du coefficient), sur l'émission ou la réception de conseil. Inversement, la densité est sensible positivement à des variables plus nombreuses et notamment le fait de travailler dans le même cabinet ou d'avoir la même spécialité.

Le modèle p_2 est régulièrement qualifié dans la littérature de modèle multi-niveau dans la mesure où il permet la mise en évidence des variables explicatives au niveau du lien, au niveau de l'acteur et enfin au niveau du réseau dans son ensemble.

2.2 Le modèle p^*

Le modèle p^* proposé par Wasserman et Pattison en 1996 [21] permet la prise en compte d'effets triadiques. Il s'agit également d'un modèle probabiliste pouvant s'écrire sous la forme

$$Pr(X = x) = \frac{exp\{\theta' z(x)\}}{k(\theta)} = \frac{exp\{\theta_1 z_1(x) + ... + \theta_r z_r(x)\}}{k(\theta)}$$

où θ est le vecteur des r paramètres à estimer du modèle, $z(x)$ le vecteur des r variables explicatives testées et k une constante permettant de normaliser la somme des probabilités.

Une variante plus économique en temps de calcul s'exprime sous la forme logistique suivante :

$$w_{ij} = log\left(\frac{Pr(X_{ij} = 1|X_{ij}^c)}{Pr(X_{ij} = 0|X_{ij}^c)}\right) = \theta'[z(x_{ij}^+) - z(x_{ij}^-)]$$

où z_{ij} désigne les variables explicatives du lien entre i et j.

Tableau 2 – Exemple de modèle p^*, tiré de Anderson *et al.*, 1999

Effet	Variable explicative	Paramètre estimé
Choix	L^{same}	-2.26
	L^{differ}	-4.17
Mutualité	M	3.27

L^{same} désigne la tendance à citer des ami-e-s de même genre, L^{differ} la tendance à citer des ami-e-s de genre différent. La probabilité qu'un lien soit présent quand deux enfants sont de même genre est 6.75 (exp(-2.26-(-4.17))=exp(1.91)) plus forte que quand les enfants sont de genre différent. La probabilité qu'un lien soit présent est 26 fois supérieure (exp(3.27)) en raison de la tendance aux liens réciproques.

Les variables les plus couramment testées sont, au niveau dyadique le choix (mesuré par la densité) et la réciprocité (nombre de liens mutuels), au niveau triadique la transitivité, l'intransitivité, les cycles, les *2-in-stars*, *2-out-stars* et *2-mixed-stars*.

Soit trois sommets i, j et k dans un graphe orienté : on a une relation transitive si les liens $i \to j$, $i \to k$ et $k \to j$ sont présents et intransitivité si $i \to j$ et $j \to k$ sont présents mais que $i \to k$ est absent. Un cycle sera formé par les liens $i \to j$, $j \to k$ et $k \to i$. Enfin, une *2-in-stars* possible est formée par les liens $i \to j$, $k \to j$; une*2-out-stars* possible par les liens $i \to j$ et $i \to k$ (voir la figure 1).

Les degrés entrants et sortants sont également souvent introduits dans le modèle.

Enfin, il est possible, comme dans le modèle p_2, d'introduire des variables attributaires relatives aux sommets comme variables explicatives.

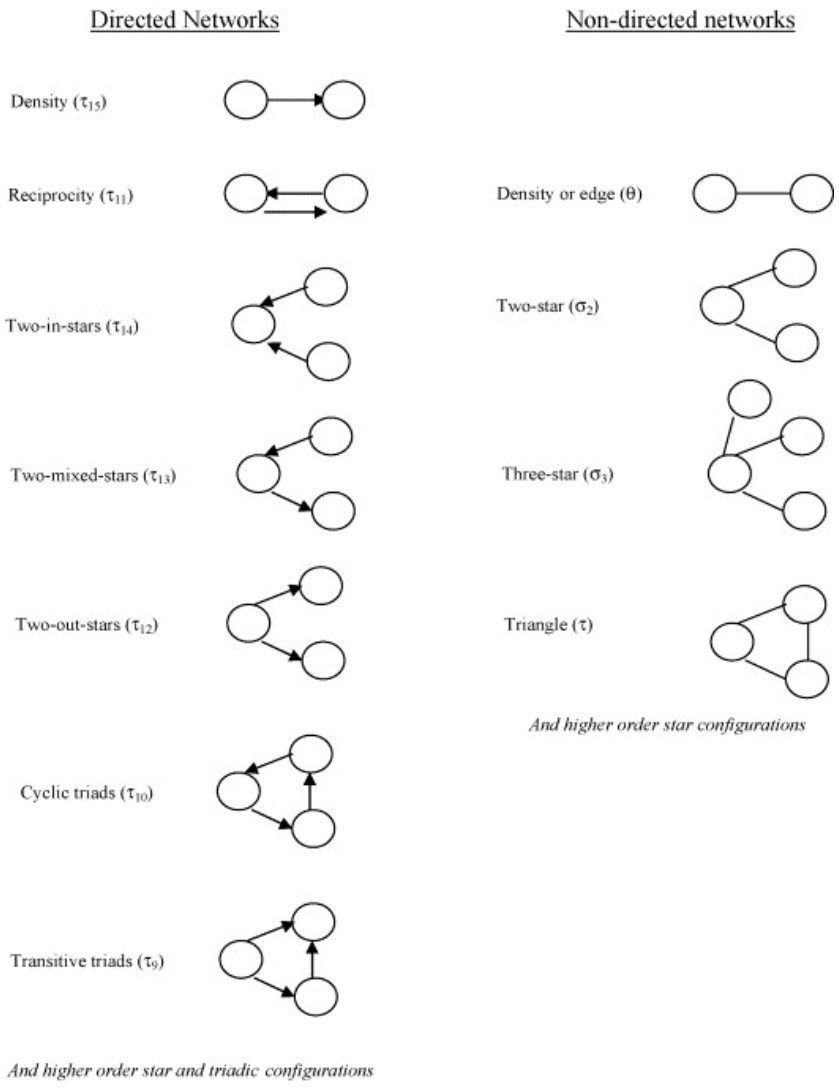
Les résultats se lisent ainsi : quand un paramètre est positif, la probabilité qu'un lien soit présent est supérieure à la probabilité qu'il soit absent. Dans la mesure où il s'agit d'une régression logistique, chaque paramètre peut s'interpréter toutes choses égales par ailleurs. L'article d'Anderson *et al.* [1] fournit l'exemple de relations au sein d'une école primaire. Les données concernent trois classes de niveaux différents. Les effets testés sont le genre, l'âge et l'appartenance à la classe. Les résultats montrent notamment que la réciprocité des liens joue un rôle essentiel (tableau 2).

Des adaptations de ce modèle aux multigraphes [11], aux graphes valués [13] et aux graphes bipartis [15] ont été proposées dans les années suivantes.

2.3 Le *blockmodel* stochastique

Les modèles probabilistes d'analyse de réseau sont également utilisés pour mettre en évidence les groupes pertinents au sein d'un graphe. Plusieurs

FIGURE 1 – Quelques configurations des modèles p^* et *ERGM*



Tiré de Robins *et al.*, *Social Network*, Elsevier, 2007.

pistes ont été proposés depuis le début des années 80 et on peut distinguer deux grandes approches, le *blockmodel a priori* et le *blockmodel a posteriori*. Dans le premier cas, on suppose qu'une variable donnée est pertinente pour une division en blocs (ex. le genre ou l'âge pour des réseaux amicaux à l'école primaire) et on teste la partition produite par cette variable. Dans le second, la division en blocs pertinents est connue et le modèle retenu est celui qui s'approche le plus près de la partition désirée.

Cette dernière piste a été explorée notamment par Holland *et al.* [6] ou par Wang et Wong [19]. Ils proposent d'ajouter dans les variables explicatives d'un modèle p_1 celle créant les blocs. Snijders a lui davantage exploré la première piste [17].

Réutilisant les formules relatives à l'équivalence structurale, deux sommets seront qualifiés de stochastiquement équivalents s'ils ont les mêmes probabilités de relations avec les autres sommets.

Le seul logiciel à ma connaissance proposant ce type d'analyse est le module `Blocks` du logiciel Stocnet développé par Boer *et al.* [2]

La plus-value de ces méthodes comparées aux méthodes canoniques du *blockmodel* n'apparaît pas toujours évidente à la lecture d'articles plus ardus les uns que les autres [3]. L'aspect le plus important reste cependant la possibilité de valider par des tests la pertinence de la partition obtenue.

3 *Exponential Random Graph Models*

Si le principe est *grosso modo* le même que celui du modèle p^*, le principal apport des *Exponential random graph models* est qu'ils permettent de comparer la forme d'un réseau empirique des distributions théoriques simulées (voir le numéro spécial de *Social Networks* consacré au sujet et notamment les deux articles introductifs [12] et [14]).

Ces modèles permettent également d'analyser le rôle des attributs des acteurs sur les choix relationnels, le degré ou les triades. Enfin, des méthodes bien documentées et claires permettent de valider la pertinence des modèles obtenus.

Reste que le choix des variables et l'interprétation des résultats nécessitent des hypothèses fortes ainsi qu'une grande prudence [4].

2. http ://www.gmw.rug.nl/ stocnet/StOCNET.htm. La dernière version date de 2007 et le développement du logiciel semble interrompu.

3. Il faut lire dans cette phrase l'aveu de l'incompétence de l'auteur sur ce sujet précis...

4. Il suffit de regarder de temps à autre les messages postés sur la liste de diffusion de `statnet` pour s'en convaincre.

4 Modéliser la dynamique d'un réseau : le modèle Siena

Le modèle Siena [5] développé par Tom Snijders [16] reprend les principes du modèle p^* mais vise à expliquer la dynamique d'un réseau (pour une présentation très claire et non mathématique de ce modèle, voir [3], à compléter par [18]). Le réseau en question doit être d'ordre constant (même nombre d'acteurs), orienté ou non, et, à ma connaissance, binaire. Et il est évidemment nécessaire de connaître l'état du réseau à au moins deux moments différents.

La variable dépendante est le changement observé dans le réseau entre un moment t et un moment $t + n$.

Le modèle est basé sur des chaînes de Markov : l'état à un moment t du graphe détermine de façon probabiliste l'évolution du graphe au moment $t + 1$. Les acteurs contrôlent l'émission de lien (Snijders parle de *actor-based model*), et à chaque moment, un acteur et un seul a la possibilité, en fonction de l'état global du réseau, de créer, de supprimer ou de maintenir ses liens sortants.

Ces caractéristiques entraînent un certain nombre de postulats, parfois peu réalistes :

- les acteurs ont une connaissance complète de l'ensemble du réseau ;
- les acteurs agissent sans coordination ;
- les acteurs n'ont pas de stratégie à long terme.

Le modèle prend en compte simultanément plusieurs types de variables explicatives : des variables relatives aux acteurs (genre, âge...), des variables relative aux paires d'acteurs et enfin des variables relatives aux effets structuraux (réciprocité, transitivité). La densité est par ailleurs systématiquement introduite dans le modèle comme variable de contrôle.

La construction du modèle est une régression logistique pas à pas où les variables sont testées et ajoutées les unes après les autres. Les résultats présentent un paramètre et son écart-type : le paramètre est considéré significatif s'il est égal au moins au double de l'écart-type. Plus le rapport entre les deux est grand, plus la significativité de la variable est forte.

Snijders a récemment proposé des extensions de ce modèle aux réseaux bipartis. Visiter sa page personnelle est recommandé à toute personne désireuse de tester ce modèle [6]. L'auteur précise que ce modèle fonctionne si le nombre de graphes est réduit (jusqu'à 6) et si le nombre de sommets ne dépasse pas quelques centaines.

5. *Simulation Investigation for Empirical Network Analysis.*
6. http ://www.stats.ox.ac.uk/ snijders/

Conclusion

Ce document pourra sembler très incomplet aux spécialistes de ces modèles, notamment en ce qui concerne le volet statistique. L'objectif principal était de fournir une introduction synthétique et claire à des méthodes encore sous utilisées dans les travaux francophones. Si ces modèles peuvent paraître complexes, il n'en reste pas moins qu'ils sont extrêmement stimulants pour plusieurs raisons. Tout d'abord, ils permettent de dépasser le caractère trop souvent descriptif de l'analyse de réseaux. Ensuite, le choix des variables explicatives introduites dans le modèle oblige le chercheur à expliciter ses hypothèses. Enfin, la validation statistique des modèles grâce à des tests constitue un atout indéniable.

Modèles statistiques de réseaux avec R

Les *packages* `stanet` et `RSiena` permettent respectivement de tester des modèles ERGM et des modèles Siena (pour les autres analyses de réseaux possibles avec R, voir le précédent tutoriel du groupe fmr sur Hal-shs). Les deux scripts suivants permettent de connaître les fonctions utiles et commentent brièvement les résultats. Le premier est tiré

Les commandes R et les résultats obtenus sont en `typewritter`. Les commentaires précédés d'un # précisent le rôle des fonctions.

Un modèle ERGM avec le module statnet

Le script suivant est tiré de la session animée par Butts *et al.* aux *INSNA Sunbelt* de février 2011 [7].

```
#chargement du package et des données

library(statnet)
data(florentine)
flomarriage

#visualisation
plot(flomarriage)

#modèle - les liens présents sont dépendants
#de la richesse des acteurs

flomodel <- ergm(flomarriage ~ edges+nodecov('wealth'))
summary(flomodel)
```

7. http ://csde.washington.edu/statnet/Resources/Sunbelt2011/ergm%20tutorial%20sunbelt%202011.pdf

```
#il s'agit de modèles aléatoires
#les résultats peuvent donc varier

Formula:   flomarriage ~ edges + nodecov("wealth")

Newton-Raphson iterations:  4

Maximum Likelihood Results:
               Estimate Std. Error MCMC s.e. p-value
edges          -2.594929   0.536056        NA  <1e-04 ***
nodecov.wealth  0.010546   0.004674        NA  0.0259 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*'

    Null  Deviance: 166.355  on 120  degrees of freedom
 Residual Deviance: 103.109  on 118  degrees of freedom
          Deviance:  63.247  on   2  degrees of freedom

#simuler 30 réseaux ayant les mêmes probabilités
flomodel.sim <- simulate(flomodel, nsim=30)

#vérifier qualité du modèle selon un critère
#ici le degré

flomodel.gof <- gof(flomodel~degree)
plot(flomodel.gof)
```

Les coefficients obtenus se lisent après transformation. Ainsi, 0.01 correspond à $\frac{exp(0.01)}{1+exp(0.01)}$, soit 0.5. La richesse d'un sommet augmente donc la probabilité qu'un lien matrimonial soit présent de 50%.

Un modèle SIENA avec le module RSiena

Le script suivant est une version très synthétisée des programmes du *Manual for SIEA version 4.0* de Ripley, Snijders et Lopez [8].

```
library(Rsiena)

#import des données

friend.data.w1 <- as.matrix(read.table("s50-network1.dat"))
friend.data.w2 <- as.matrix(read.table("s50-network2.dat"))
```

8. http ://www.stats.ox.ac.uk/ snijders/siena/s_man400.pdf

```
friend.data.w3 <- as.matrix(read.table("s50-network3.dat"))

#attributs

drink <- as.matrix(read.table("s50-alcohol.dat"))
smoke <- as.matrix(read.table("s50-smoke.dat"))

#transformation en objet Siena (variable à expliquer)
#empilant les 3 matrices 50*50

friendship <- sienaNet(
    array(c(friend.data.w1, friend.data.w2, friend.data.w3),
    dim = c(50,50,3)))

#variables explicatives

smoke1 <- coCovar(smoke[,1])
alcohol <- varCovar(drink)

#combinaison des variables

mydata <- sienaDataCreate(friendship, smoke1, alcohol)

myeff <- getEffects(mydata)

#tester et sélectionner des effets structuraux

myeff <- includeEffects(myeff, similarity, cycle3)
summary(myeff)

#création du modèle et estimation
#des paramètres

mymodel <- sienaModelCreate(useStdInits = FALSE, projname='s50_3')
ans <- siena07(mymodel, data = mydata, effects = myeff)
summary(ans)
```

Références

[1] Carolyn J. Anderson, Stanley Wasserman et Bradley Crouch : A $p*$ primer : logit models for social networks. *Social Networks*, 21(1):37–66, 1999.

[2] Christophe Boschet et Tina Rambonilaza : Les mécanismes de coordination dans les réseaux sociaux : un cadre analytique de la dynamique territoriale. *Revue d'Économie Régionale et Urbaine*, (3):569–593, 2010.

[3] Ainhoa de Federico de la Rua : L'analyse longitudinale de réseaux sociaux totaux avec Siena. *Bulletin de méthodologie sociologique*, (84):5–39, 2004.

[4] Fabien Éloire : *Les réseaux interorganisationnels dans la restauration lilloise. Une approche néo-structurale du marché et des processus sociaux*. Thèse de doctorat, Lille 1, 2009.

[5] Joseph Galaskiewicz, Stanley Wasserman, Barbara Rauschenbach, Wolfgang Bielefeld et Patti Mullaney : The Influence of Corporate Power, Social Status, and Market Position in Corporate Interlocks in a Regional Network. *Social Forces*, 64(2):403–431, 1985.

[6] Paul W. Holland, K.B. Laskey et Samuel Leinhardt : Stochastic blockmodels : First steps. *Social Networks*, 5(2):109–137, 1983.

[7] Paul W. Holland et Samuel Leinhardt : An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

[8] Emmanuel Lazega, Lise Mounier, Tom Snijders et Paola Tubaro : Norms, status and the dynamics of advice networks : A case study. *Social Networks*, 2009.

[9] Emmanuel Lazega et M. van Duijn : Position in formal structure, personal characteristics and choices of advisors in a law firm : a logistic regression model for dyadic network data. *Social Networks*, 19(3):375–397, 1997.

[10] Claire Lemercier et Paul-André Rosental : The Structure and Dynamics of Migration Patterns in 19th-century Northern France. *Pre-print*. http ://halshs.archives-ouvertes.fr/halshs-00450035.

[11] Philippa Pattison et Stanley Wasserman : Logit models and logistic regressions for social networks : II. Multivariate relations. *British Journal of Mathematical and Statistical Psychology*, 52(2):169–193, 1999.

[12] Garry Robins, Philippa Pattison, Yuval Kalish et Dean Lusher : An introduction to exponential random graph p^* models for social networks. *Social networks*, 29(2):173–191, 2007.

[13] Garry Robins, Philippa Pattison et Stanley Wasserman : Logit models and logistic regressions for social networks : III. Valued relations. *Psychometrika*, 64(3):371–394, 1999.

[14] Garry Robins, Tom A.B. Snijders, Peng Wang, Mark Handcock et Philippa Pattison : Recent developments in exponential random graph p^* models for social networks. *Social Networks*, 29(2):192–215, 2007.

[15] John Skvoretz et Katherine Faust : Logit models for affiliation networks. *Sociological Methodology*, 29(1):253–280, 1999.

[16] Tom A.B. Snijders : The statistical evaluation of social network dynamics. *Sociological methodology*, 31(1):361–395, 2001.

[17] Tom A.B. Snijders et Krzysztof Nowicki : Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14(1):75–100, 1997.

[18] Tom A.B. Snijders, Gerhard G. van de Bunt et Christian E.G. Steglich : Introduction to stochastic actor-based models for network dynamics. *Social networks*, 32(1):44–60, 2010.

[19] Yuchung J. Wang et Georges Y. Wong : Blockmodels for Directed Graphs. *Journal of the American Statistical Association*, 82(397):8–19, 1987.

[20] Stanley Wasserman et Joseph Galaskiewicz : Some generalizations of p_1 : External constraints, interactions and non-binary relations. *Social Networks*, 6(2):177–192.

[21] Stanley Wasserman et Philippa Pattison : Logit model and logistic regression for social networks : I. An introduction to Markov graphs and p^*. *Psychometrika*, 61(3):401–425, 1996.