

# Modèles de familles exponentielles pour les réseaux sociaux

David Hunter

Penn State University USA et Le Studium, CNRS

Recherch soutenu par NIDA Grant DA012831 et NICHD Grant HD041877  
(Martina Morris, University of Washington, investigateur principal)

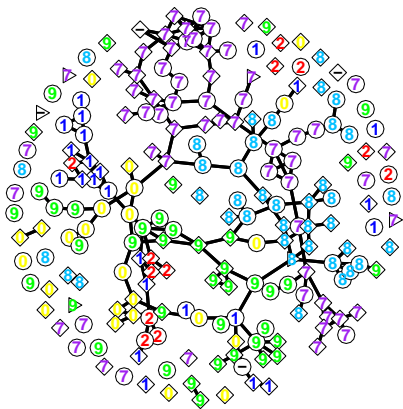
Beaucoup de remerciements à Didier Chauveau, le Studium, Stéphane Cordier, et tous les gens de MAPMO.

Université d'Orléans, 1 Avril 2008 — vraiment

# Exemple d'un réseau social

## Données d'amitié dans un collège/lycée

School 10: 205 Students

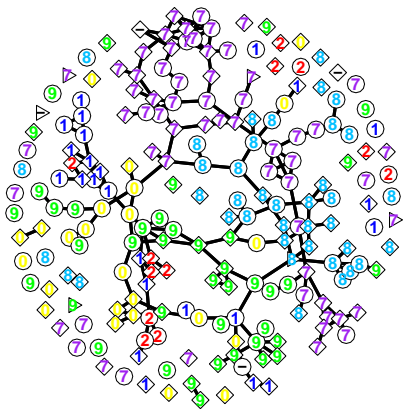


- Arête = amitié mutuelle.
- Etiquettes colorées = "grade" aux états-unis, de 7 à 12.
- Cercle = fille; Carré = garçon
- Positions (dans  $\mathbb{R}^2$ )  $\neq$  importantes

# Exemple d'un réseau social

## Données d'amitié dans un collège/lycée

School 10: 205 Students

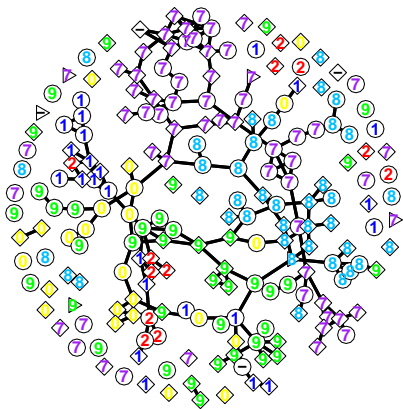


- Arête = amitié mutuelle.
- Etiquettes colorées = "grade" aux états-unis, de 7 à 12.
- Cercle = fille; Carré = garçon
- Positions (dans  $\mathbb{R}^2$ )  $\neq$  importantes
- Anglais: arête="edge"; arc = "arc"; sommet="node" ou "vertex"

# Exemple d'un réseau social

## Données d'amitié dans un collège/lycée

School 10: 205 Students



- Arête = amitié mutuelle.
- Etiquettes colorées = “grade” aux états-unis, de 7 à 12.
- Cercle = fille; Carré = garçon
- Positions (dans  $\mathbb{R}^2$ )  $\neq$  importantes
- Anglais: arête=“edge”; arc = “arc”; sommet=“node” ou “vertex”
- Comment est-ce que les variables explicatives influencent le réseau?
- i.e., comment fait-on la “régression” quand la variable à expliquer est un réseau?

- 1 Les Modèles ERG
- 2 Approximation de l'EMV
- 3 L'estimation du maximum de pseudo-vraisemblance
- 4 L'évaluation de l'adéquation du modèle

## Exponential-family Random Graph Model (ERGM)

$$P_{\theta}(Y = y) \propto \exp\{\theta^t g(y)\}$$

ou

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)},$$

où

- $Y$  = réseau aléatoire avec  $n$  sommets (matrice de 0 et 1)
- $\theta$  = vecteur de paramètres
- $g(y)$  = vecteur donné de statistiques selon le réseau  $y$
- $\kappa(\theta)$  = “constante” de normalisation

## Modèle d'un Réseau Aléatoire de Famille Exponentielle (MRAFE?)

$$P_{\theta}(Y = y) \propto \exp\{\theta^t g(y)\}$$

ou

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)},$$

où

- $Y$  = réseau aléatoire avec  $n$  sommets (matrice de 0 et 1)
- $\theta$  = vecteur de paramètres
- $g(y)$  = vecteur donné de statistiques selon le réseau  $y$
- $\kappa(\theta)$  = “constante” de normalisation

# D'où vient le nom ERGM?

## Famille Exponentielle

Quand la loi d'une variable aléatoire peut être écrite comme

$$f(y) \propto \exp\{\theta^t g(y)\},$$

la famille de toutes ces variables aléatoires (indexée par  $\theta$ ) s'appelle une **famille exponentielle**.

- Parce que les réseaux aléatoires de notre modèle viennent d'une famille exponentielle, ce modèle s'appelle un **Modèle de famille exponentielle d'un réseau aléatoire**, ou en anglais, an **Exponential-family Random Graph Model (ERGM)**.

## Le modèle ERG

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

- Utiliser les données observées (un réseau  $y^{\text{obs}}$ ) pour déterminer le meilleur modèle de la famille.
- C'est à dire, trouver le meilleur  $\theta$ .

## Le modèle ERG

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

## Le modèle ERG

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

- Donc,  $\kappa(\theta)$  est la “constante” de normalisation:

$$\kappa(\theta) = \sum_{\text{tous réseaux } z} \exp\{\theta^t g(z)\}.$$

## Le modèle ERG

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

- Donc,  $\kappa(\theta)$  est la “constante” de normalisation:

$$\kappa(\theta) = \sum_{\text{tous réseaux } z} \exp\{\theta^t g(z)\}.$$

- On peut remplacer  $g(y)$  avec  $[g(y) - g(y^{\text{obs}})]$  sans changer  $P_{\theta}(Y = y)$ .

## Le modèle ERG

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}$$

- Donc,  $\kappa(\theta)$  est la “constante” de normalisation:

$$\kappa(\theta) = \sum_{\text{tous réseaux } z} \exp\{\theta^t g(z)\}.$$

- On peut remplacer  $g(y)$  avec  $[g(y) - g(y^{\text{obs}})]$  sans changer  $P_{\theta}(Y = y)$ .
- Donc, SPDG — sans perte de généralité — on peut “recentrer”  $g(y)$  et puis  $g(y^{\text{obs}}) = 0$ .

## Le modèle ERG

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ où } g(y^{\text{obs}}) = 0$$

- Donc,  $\kappa(\theta)$  est la “constante” de normalisation:

$$\kappa(\theta) = \sum_{\text{tous réseaux } z} \exp\{\theta^t g(z)\}.$$

- On peut remplacer  $g(y)$  avec  $[g(y) - g(y^{\text{obs}})]$  sans changer  $P_{\theta}(Y = y)$ .
- Donc, SPDG — sans perte de généralité — on peut “recentrer”  $g(y)$  et puis  $g(y^{\text{obs}}) = 0$ .

## Le modèle ERG

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ où } g(y^{\text{obs}}) = 0.$$

- Puisque  $g(y^{\text{obs}}) = 0$ , la log-vraisemblance est facilement  $\ell(\theta) = -\log \kappa(\theta)$ .

## Le modèle ERG

$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ où } g(y^{\text{obs}}) = 0.$$

- Puisque  $g(y^{\text{obs}}) = 0$ , la log-vraisemblance est facilement  $\ell(\theta) = -\log \kappa(\theta)$ .
- On veut trouver le  $\hat{\theta}$  qui maximise  $\ell(\theta) = -\log \kappa(\theta)$ .
- Donc,  $\hat{\theta}$  dénote **l'estimateur du maximum de vraisemblance**, ou EMV.

- 1 Les Modèles ERG
- 2** Approximation de l'EMV
- 3 L'estimation du maximum de pseudo-vraisemblance
- 4 L'évaluation de l'adéquation du modèle

## Le modèle ERG

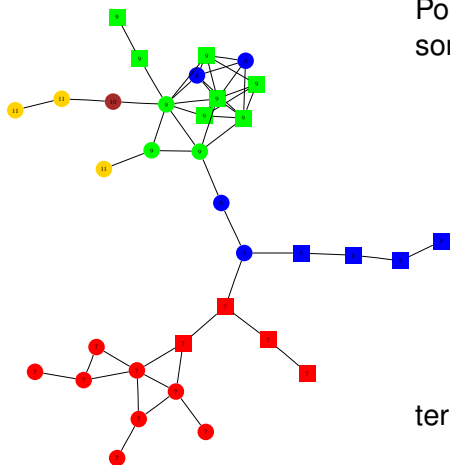
$$P_{\theta}(Y = y) = \frac{\exp\{\theta^t g(y)\}}{\kappa(\theta)}, \text{ where } g(y^{\text{obs}}) = 0$$

- On voudrait maximiser le log-vraisemblance

$$\ell(\theta) = -\log \kappa(\theta) = -\log \sum_{\substack{\text{all possible} \\ \text{graphs } z}} \exp\{\theta^t g(z)\}.$$

- Malheureusement,  $\kappa(\theta)$  est une bête méchante: Juste pour calculer (pas maximiser)  $\ell(\theta)$ , on a besoin de quelques minutes ...

# Combien de temps?

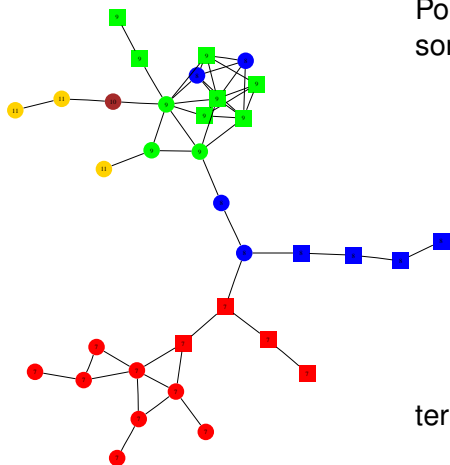


Pour ce réseau-ci, avec juste 34 sommets, il faut ajouter

7,547,924,849,643,082,704,483,  
109,161,976,537,781,833,842,  
440,832,880,856,752,412,600,  
491,248,324,784,297,704,172,  
253,450,355,317,535,082,936,  
750,061,527,689,799,541,169,  
259,849,585,265,122,868,502,  
865,392,087,298,790,653,952

termes.

# Combien de temps?



Pour ce réseau-ci, avec juste 34 sommets, il faut ajouter

7,547,924,849,643,082,704,483,  
109,161,976,537,781,833,842,  
440,832,880,856,752,412,600,  
491,248,324,784,297,704,172,  
253,450,355,317,535,082,936,  
750,061,527,689,799,541,169,  
259,849,585,265,122,868,502,  
865,392,087,298,790,653,952

termes. (i.e.,  $2^{\binom{34}{2}}$  termes)

- Pour  $\theta_0$  connu, on a

$$-\log E_{\theta_0} [\exp \{(\theta - \theta_0)^t g(Y)\}] = \ell(\theta) - \ell(\theta_0),$$

où  $E_{\theta_0}$  dénote l'espérance dans l'ERGM sous  $\theta_0$ .

- Pour  $\theta_0$  connu, on a

$$-\log E_{\theta_0} [\exp \{(\theta - \theta_0)^t g(Y)\}] = \ell(\theta) - \ell(\theta_0),$$

où  $E_{\theta_0}$  dénote l'espérance dans l'ERGM sous  $\theta_0$ .

- Donc, on emploie la LGN:

$$\begin{aligned} \ell(\theta) - \ell(\theta_0) &= -\log E_{\theta_0} (\exp \{(\theta - \theta_0)^t g(Y)\}) \\ &\approx -\log \frac{1}{m} \sum_{i=1}^m \exp \{(\theta - \theta_0)^t g(Y_i)\}, \end{aligned}$$

où  $Y_1, Y_2, \dots, Y_m$  est un échantillon de l'ERGM pour  $\theta_0$ .

# Comment est-ce qu'on choisit $\theta_0$ ?

- Par la LGN, la valeur approximative de  $\ell(\theta) - \ell(\theta_0)$  converge vers la valeur vraie pour  $\theta_0$  arbitraire.
- Mais en fait, la convergence peut être extrêmement lente, surtout si  $\theta_0$  n'est pas près de l'EMV.

# Comment est-ce qu'on choisit $\theta_0$ ?

- Par la LGN, la valeur approximative de  $\ell(\theta) - \ell(\theta_0)$  converge vers la valeur vraie pour  $\theta_0$  arbitraire.
- Mais en fait, la convergence peut être extrêmement lente, surtout si  $\theta_0$  n'est pas près de l'EMV.
- Un choix qui marche parfois est **l'estimateur du maximum de pseudo-vraisemblance**, ou EMPV.

- 1 Les Modèles ERG
- 2 Approximation de l'EMV
- 3 L'estimation du maximum de pseudo-vraisemblance**
- 4 L'évaluation de l'adéquation du modèle

# “Log-odds” conditionnel d’une arête

Notation: Pour un réseau  $y$  et une paire  $(i, j)$  de sommets,

- $y_{ij} = 0$  ou  $1$ , selon qu’il existe une arête
- $y_{ij}^c$  dénote toutes les  $y_{k\ell}$  sauf  $y_{ij}$
- $y_{ij}^+$  dénote le même réseau que  $y$  mais où  $y_{ij} = 1$
- $y_{ij}^-$  dénote le même réseau que  $y$  mais où  $y_{ij} = 0$

# “Log-odds” conditionnel d’une arête

Notation: Pour un réseau  $y$  et une paire  $(i, j)$  de sommets,

- $y_{ij} = 0$  ou  $1$ , selon qu’il existe une arête
- $y_{ij}^c$  dénote toutes les  $y_{k\ell}$  sauf  $y_{ij}$
- $y_{ij}^+$  dénote le même réseau que  $y$  mais où  $y_{ij} = 1$
- $y_{ij}^-$  dénote le même réseau que  $y$  mais où  $y_{ij} = 0$

Lorsque  $Y_{ij}^c = y_{ij}^c$ , le valeur de  $Y$  dépend de  $Y_{ij}$  seulement.

# “Log-odds” conditionnel d’une arête

Notation: Pour un réseau  $y$  et une paire  $(i, j)$  de sommets,

- $y_{ij} = 0$  ou  $1$ , selon qu’il existe une arête
- $y_{ij}^c$  dénote toutes les  $y_{k\ell}$  sauf  $y_{ij}$
- $y_{ij}^+$  dénote le même réseau que  $y$  mais où  $y_{ij} = 1$
- $y_{ij}^-$  dénote le même réseau que  $y$  mais où  $y_{ij} = 0$

Lorsque  $Y_{ij}^c = y_{ij}^c$ , le valeur de  $Y$  dépend de  $Y_{ij}$  seulement.  
Le ratio des deux probabilités  $Y_{ij} = 1$  et  $Y_{ij} = 0$  — appelé “the odds” — dans cette condition est:

Rappel:  $P_{\theta}(Y = y) = \exp\{\theta^t g(y)\} / \kappa(\theta)$ .

# “Log-odds” conditionnel d’une arête

Notation: Pour un réseau  $y$  et une paire  $(i, j)$  de sommets,

- $y_{ij} = 0$  ou  $1$ , selon qu’il existe une arête
- $y_{ij}^c$  dénote toutes les  $y_{k\ell}$  sauf  $y_{ij}$
- $y_{ij}^+$  dénote le même réseau que  $y$  mais où  $y_{ij} = 1$
- $y_{ij}^-$  dénote le même réseau que  $y$  mais où  $y_{ij} = 0$

$$\frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \frac{\exp\{\theta^t g(y_{ij}^+)\}}{\exp\{\theta^t g(y_{ij}^-)\}}$$

Rappel:  $P_\theta(Y = y) = \exp\{\theta^t g(y)\} / \kappa(\theta)$ .

# “Log-odds” conditionnel d’une arête

Notation: Pour un réseau  $y$  et une paire  $(i, j)$  de sommets,

- $y_{ij} = 0$  ou  $1$ , selon qu’il existe une arête
- $y_{ij}^c$  dénote toutes les  $y_{k\ell}$  sauf  $y_{ij}$
- $y_{ij}^+$  dénote le même réseau que  $y$  mais où  $y_{ij} = 1$
- $y_{ij}^-$  dénote le même réseau que  $y$  mais où  $y_{ij} = 0$

$$\frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \exp\{\theta^t [g(y_{ij}^+) - g(y_{ij}^-)]\}$$

Rappel:  $P_\theta(Y = y) = \exp\{\theta^t g(y)\} / \kappa(\theta)$ .

# “Log-odds” conditionnel d’une arête

Notation: Pour un réseau  $y$  et une paire  $(i, j)$  de sommets,

- $y_{ij} = 0$  ou  $1$ , selon qu’il existe une arête
- $y_{ij}^c$  dénote toutes les  $y_{k\ell}$  sauf  $y_{ij}$
- $y_{ij}^+$  dénote le même réseau que  $y$  mais où  $y_{ij} = 1$
- $y_{ij}^-$  dénote le même réseau que  $y$  mais où  $y_{ij} = 0$

$$\log \frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \theta^t [g(y_{ij}^+) - g(y_{ij}^-)]$$

Rappel:  $P_\theta(Y = y) = \exp\{\theta^t g(y)\} / \kappa(\theta)$ .

# “Log-odds” conditionnel d’une arête

Notation: Pour un réseau  $y$  et une paire  $(i, j)$  de sommets,

- $\Delta g(y)_{ij}$  dénote le vecteur de statistiques de changement,

$$\Delta g(y)_{ij} = g(y_{ij}^+) - g(y_{ij}^-).$$

donc  $\Delta g(y)_{ij}$  est le log-odds conditionnel de l’arête  $i \longleftrightarrow j$ .

$$\log \frac{P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)}{P(Y_{ij} = 0 | Y_{ij}^c = y_{ij}^c)} = \theta^t \Delta g(y)_{ij}$$

Rappel:  $P_\theta(Y = y) = \exp\{\theta^t g(y)\} / \kappa(\theta)$ .

# L'intuition de pseudo-vraisemblance

- Supposant qu'il n'y a pas de dépendance (ou seulement de la dépendance faible) parmi les  $Y_{ij}$ .
- C'est à dire, la probabilité marginale  $P(Y_{ij} = 1)$  est exactement ou approximativement la même que la probabilité conditionnelle  $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$ .

# L'intuition de pseudo-vraisemblance

- Supposant qu'il n'y a pas de dépendance (ou seulement de la dépendance faible) parmi les  $Y_{ij}$ .
- C'est à dire, la probabilité marginale  $P(Y_{ij} = 1)$  est exactement ou approximativement la même que la probabilité conditionnelle  $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$ .
- Donc, on a (au moins approximativement)

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^t \Delta g(y^{\text{obs}})_{ij}$$

et on peut estimer  $\theta$  par utilisant la régression logistique.

# L'intuition de pseudo-vraisemblance

- Supposant qu'il n'y a pas de dépendance (ou seulement de la dépendance faible) parmi les  $Y_{ij}$ .
- C'est à dire, la probabilité marginale  $P(Y_{ij} = 1)$  est exactement ou approximativement la même que la probabilité conditionnelle  $P(Y_{ij} = 1 | Y_{ij}^c = y_{ij}^c)$ .
- Donc, on a (au moins approximativement)

$$\log \frac{P(Y_{ij} = 1)}{P(Y_{ij} = 0)} = \theta^t \Delta g(y^{\text{obs}})_{ij}$$

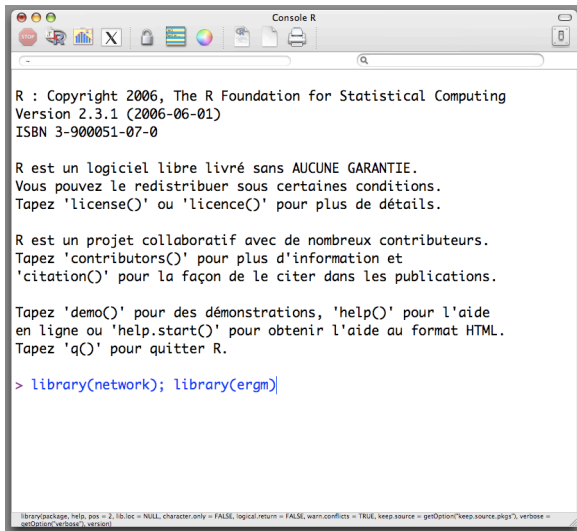
et on peut estimer  $\theta$  par utilisant la régression logistique.

- De cette façon, on obtient l'**estimateur du maximum de pseudo-vraisemblance** (EMPV).
- Au cas où les  $Y_{ij}$  seraient indépendants, l'EMPV = l'EMV.

Malheureusement, on ne sait pas grand chose des EMPVs en général, mais on connaît des situations où ils peuvent être trompeurs.

- Lorsque le modèle est inapproprié, on obtient un EMPV facilement (pas comme un EMV approximatif), qui masque le problème.
- Lorsque le modèle est bon, l'EMPV paraît “près” du EMV; mais “près” peut être trompeur parce que des petites perturbations de  $\theta$  peuvent créer de grosses différences sur le comportement des réseaux aléatoires.

- 1 Les Modèles ERG
- 2 Approximation de l'EMV
- 3 L'estimation du maximum de pseudo-vraisemblance
- 4 L'évaluation de l'adéquation du modèle**



The screenshot shows a window titled "Console R" with a standard macOS-style title bar. The window contains the following text:

```
R : Copyright 2006, The R Foundation for Statistical Computing
Version 2.3.1 (2006-06-01)
ISBN 3-900051-07-0

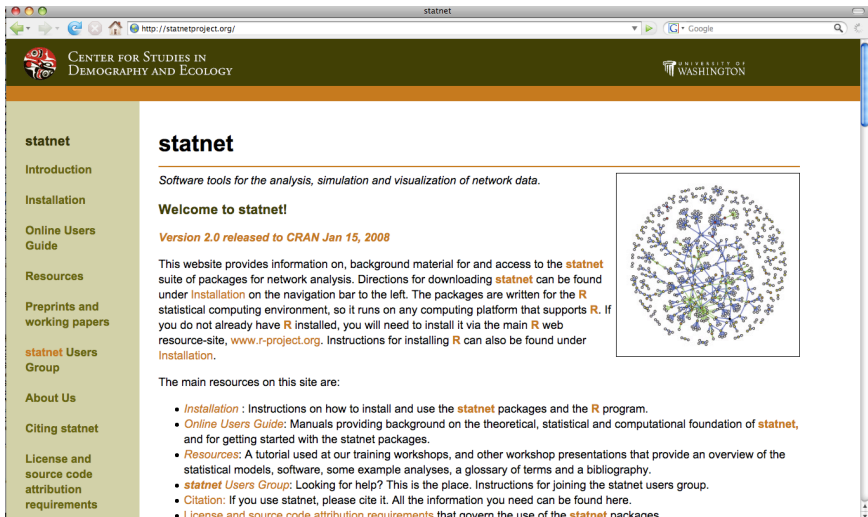
R est un logiciel libre livré sans AUCUNE GARANTIE.
Vous pouvez le redistribuer sous certaines conditions.
Tapez 'license()' ou 'licence()' pour plus de détails.

R est un projet collaboratif avec de nombreux contributeurs.
Tapez 'contributors()' pour plus d'information et
'citation()' pour la façon de le citer dans les publications.

Tapez 'demo()' pour des démonstrations, 'help()' pour l'aide
en ligne ou 'help.start()' pour obtenir l'aide au format HTML.
Tapez 'q()' pour quitter R.

> library(network); library(ergm)
```

At the bottom of the console, there is a small status bar with the following text: `library(package, help, pos = 2, lib.loc = NULL, character.only = FALSE, logical.return = FALSE, warn.conflicts = TRUE, keep.source = getOption("keep.source.pkgs"), verbose = getOption("verbose"), version)`



The screenshot shows a web browser window with the URL <http://statnetproject.org/>. The page header includes the logo of the Center for Studies in Demography and Ecology and the University of Washington. The main content area features a navigation menu on the left and a central section titled "statnet" with a subtitle "Software tools for the analysis, simulation and visualization of network data." Below this, there is a "Welcome to statnet!" message, a note about "Version 2.0 released to CRAN Jan 15, 2008", and a paragraph describing the website's purpose. A circular network graph visualization is shown on the right. The bottom of the page contains a list of resources and a footer with navigation icons.

statnet

Introduction

Installation

Online Users Guide

Resources

Preprints and working papers

statnet Users Group

About Us

Citing statnet

License and source code attribution requirements

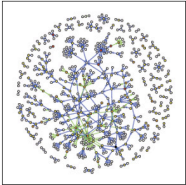
## statnet

Software tools for the analysis, simulation and visualization of network data.

### Welcome to statnet!

Version 2.0 released to CRAN Jan 15, 2008

This website provides information on, background material for and access to the **statnet** suite of packages for network analysis. Directions for downloading **statnet** can be found under **Installation** on the navigation bar to the left. The packages are written for the **R** statistical computing environment, so it runs on any computing platform that supports **R**. If you do not already have **R** installed, you will need to install it via the main **R** web resource-site, [www.r-project.org](http://www.r-project.org). Instructions for installing **R** can also be found under **Installation**.



The main resources on this site are:

- **Installation** : Instructions on how to install and use the **statnet** packages and the **R** program.
- **Online Users Guide**: Manuals providing background on the theoretical, statistical and computational foundation of **statnet**, and for getting started with the statnet packages.
- **Resources**: A tutorial used at our training workshops, and other workshop presentations that provide an overview of the statistical models, software, some example analyses, a glossary of terms and a bibliography.
- **statnet Users Group**: Looking for help? This is the place. Instructions for joining the statnet users group.
- **Citation**: If you use statnet, please cite it. All the information you need can be found here.
- **License and source code attribution requirements** that govern the use of the **statnet** packages.

ERGM

famille

$$\exp\{\theta^t g(y)\}$$

# L'intuition de l'adéquation du modèle

ERGM  
famille  
 $\exp\{\theta^t g(y)\}$

→

↑  
 $y^{\text{obs}}$



# L'intuition de l'adéquation du modèle

ERGM  
famille  
 $\exp\{\theta^t g(y)\}$

→

↑  
 $y^{\text{obs}}$

(approx)  
EMV  
 $\hat{\theta}$



# L'intuition de l'adéquation du modèle

ERGM  
famille  
 $\exp\{\theta^t g(y)\}$



↑  
 $y^{\text{obs}}$

(approx)  
EMV

$\hat{\theta}$



ERGM  
estimé  
 $\exp\{\hat{\theta}^t g(y)\}$



# L'intuition de l'adéquation du modèle

ERGM  
famille  
 $\exp\{\theta^t g(y)\}$

→  
↑  
 $y^{\text{obs}}$

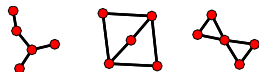


(approx)  
EMV  
 $\hat{\theta}$

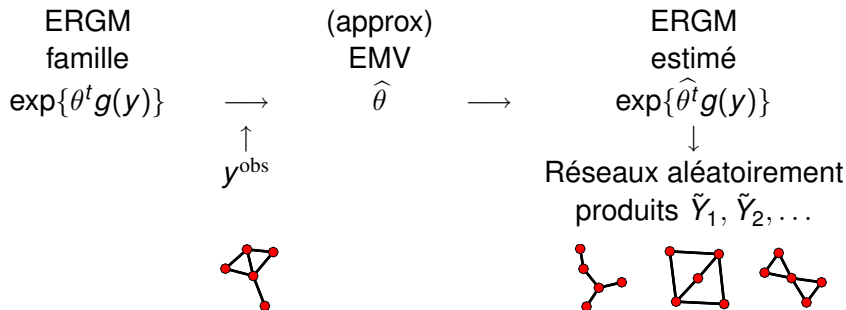
→

ERGM  
estimé  
 $\exp\{\hat{\theta}^t g(y)\}$

↓  
Réseaux aléatoirement  
produits  $\tilde{Y}_1, \tilde{Y}_2, \dots$



# L'intuition de l'adéquation du modèle



- La question: Est-il vraisemblable que  $y^{\text{obs}}$  soit issu de l'échantillon  $\tilde{Y}_1, \tilde{Y}_2, \dots$ ?

# Deux modèles pour comparer

- Modèle A:  $g(y)$  contient les termes:
  - combien d'arêtes
  - Ressemblance d'âge, sexe, et race
  - Les effets principaux d'âge, sexe, et race
  - $\sum_i (0.632)^i EP_i$ , où  $EP_i = \#$  d'arêtes avec  $i$  partenaires communs
  
- Modèle B:  $g(y)$  contient les termes:
  - combien d'arêtes
  - Ressemblance d'âge
  - # de chemins de la longueur 2
  - # de triangles

(NB: On a eu besoin du EMPV pour Modèle B)

- Modèle A:  $g(y)$  contient les termes:
  - combien d'arêtes
  - Ressemblance d'âge, sexe, et race
  - Les effets principaux d'âge, sexe, et race
  - $\sum_i (0.632)^i EP_i$ , où  $EP_i = \#$  d'arêtes avec  $i$  partenaires communs

```
modA <- ergm(data ~ edges
              + nodematch("Age", diff=TRUE)
              + nodematch("Sex", diff=TRUE)
              + nodematch("Race", diff=TRUE)
              + nodefactor("Age")
              + nodefactor("Sex")
              + nodefactor("Race")
              + gwesp(1.0, fixed=TRUE) )
```

# Deux modèles pour comparer

```
modB <- ergm(data ~ edges
              + nodematch("Age", diff=TRUE)
              + twopath
              + triangle )
```

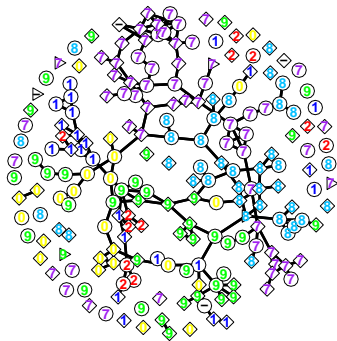
- Modèle B:  $g(y)$  contient les termes:
  - combien d'arêtes
  - Ressemblance d'âge
  - # de chemins de la longueur 2
  - # de triangles

(NB: On a eu besoin du EMPV pour Modèle B)

# Idée très simple: Le test des yeux

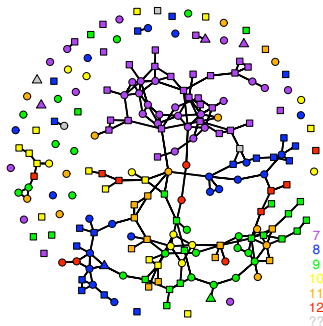
Les données:

School 10: 205 Students



Réseau prélevé,  
modèle A:

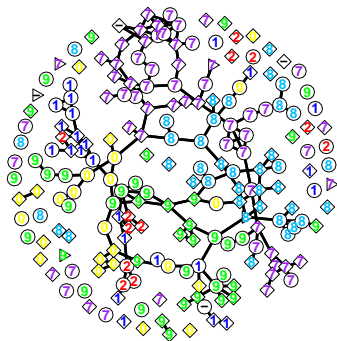
Simulated graph: By grade



# Le test des yeux (continué)

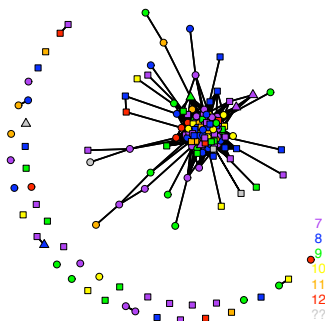
Les données:

School 10: 205 Students



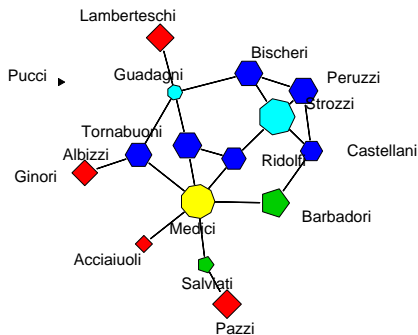
Réseau prélevé,  
modèle B:

Simulated graph: By grade



(Oh la la!)

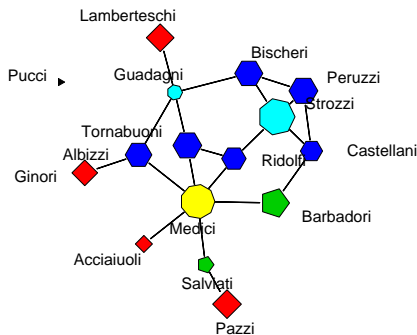
## Un exemple bien connu (arête = mariage entre familles):



```
library(network)
library(ergm)
data(flomarriage)
...
plot(flomarriage,
     displaylabels=TRUE,
     vertex.sides=degrees+3,
     vertex.col=degrees+1,
     vertex.cex=log(wealth))
```

# Des autres tests de l'adéquation du modèle

## Un exemple bien connu (arête = mariage entre familles):



```
library(network)
library(ergm)
data(flomarriage)
...
plot(flomarriage,
     displaylabels=TRUE,
     vertex.sides=degrees+3,
     vertex.col=degrees+1,
     vertex.cex=log(wealth))
```

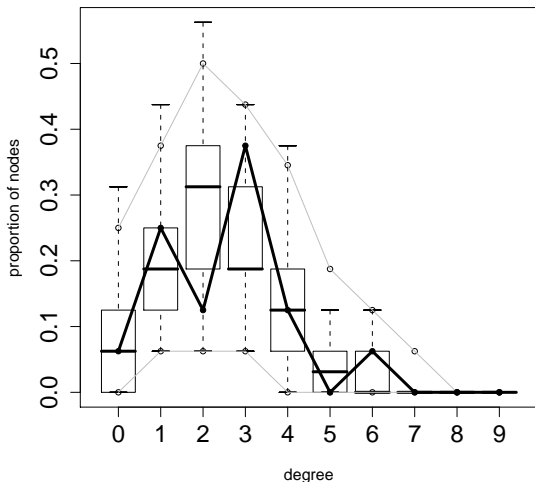
Maintenant, un modèle ERG:

```
mod1 <- ergm(flomarriage ~ edges + kstar(2))
```

# Un test visuel: Distribution des degrés

```
plot(gof(ergm(flomarriage ~ edges + kstar(2))))
```

Goodness-of-fit diagnostics

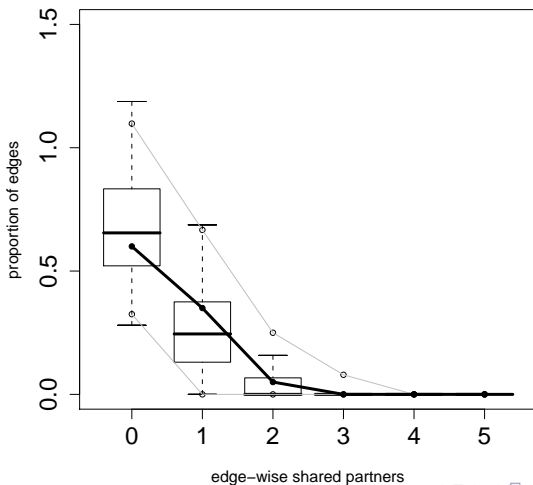


NB: Degré d'un  
sommet = # ses  
partenaires

# Un test visuel: Distribution des partenaires communs

```
plot(gof(ergm(flomarriage ~ edges + kstar(2))))
```

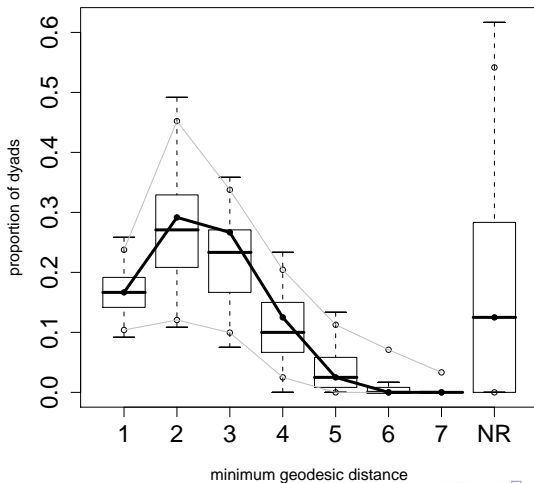
Goodness-of-fit diagnostics



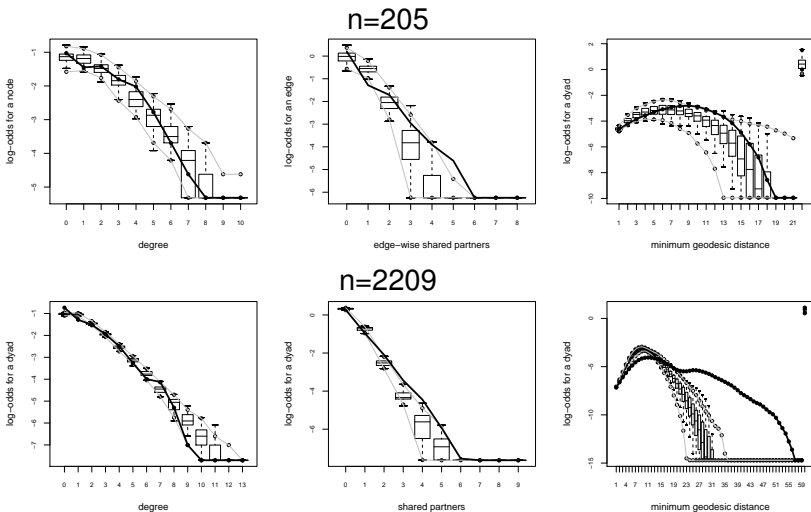
# Distribution des distances les plus courtes

```
plot(gof(ergm(flomarriage ~ edges + kstar(2))))
```

Goodness-of-fit diagnostics



# Quelques exemples plus grands



Hunter, Goodreau, Handcock (2008), *JASA*.

## Quelques références

- Frank, O. and D. Strauss (1986), Markov graphs, *JASA*
- Geyer, C. J. and E. Thompson (1992), Constrained Monte Carlo maximum likelihood for dependent data, *J. Roy. Stat. Soc. B*
- Handcock, M. S. (2003) Assessing degeneracy in statistical models of social networks,  
<http://www.csss.washington.edu/Papers>
- Holland, P. W. and S. Leinhardt (1981), An exponential family of probability distributions for directed graphs, *JASA*
- Snijders, Tom A. B. (2002), Markov chain Monte Carlo estimation of exponential random graph models, *J. Soc. Struct.*
- Strauss, D. and M. Ikeda (1990), Pseudolikelihood estimation for social networks, *JASA*
- Wasserman, S. and P. Pattison (1996), Logit models and logistic regression for social networks: I. An introduction to Markov graphs and  $p^*$ , *Psychometrika*