L'analyse de données relationnelles longitudinales par une approche réseau

Les modèles SAOM

Timothée Chabot
Post-doctorant INED – UR6

Présentation au Service des Méthodes Statistiques
Novembre 2022

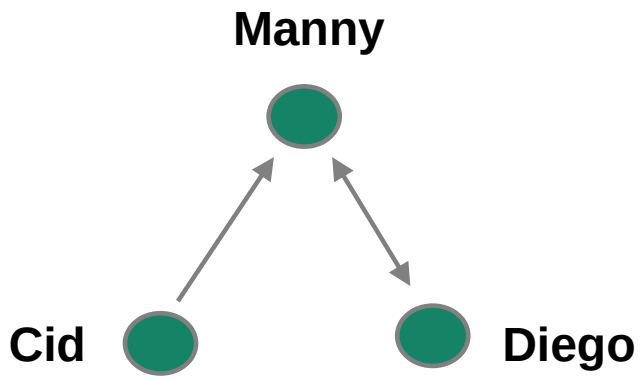
Notions de base

- Réseau = ensemble de noeuds (individus) et de liens entre ces noeuds (relations)

  Réseau *complet* = on connaît (presque) tous les liens entre individus dans une population délimitée (e.g. une école).

- Souvent représenté *via* une *matrice de nominations* :

	Manny	**Diego**	**Cid**
Manny	0	1	0
Diego	1	0	0
Cid	1	0	0



Programme des hostilités

I. Philosophie générale du modèle

II. Présentation du modèle (sans maths!)

1. SAOM comme régression logistique au niveau du lien
2. SAOM comme modèle génératif basé sur des simulations
3. SAOM comme modèle *actor-oriented* et en temps continu

III. Exemple d'application : l'homophilie sociale au collège

I. Philosophie générale du modèle

Stochastic Actor-Oriented Model (SAOM)

- Une classe de modèles créée par Tom Snijders pour étudier l'évolution de réseaux complets (Snijders 2017)

  → des applications en multi-niveaux

  → possibilité d'étudier la co-évolution d'un réseau et d'un comportement (études des phénomènes de sélection et d'influence : les fumeurs deviennent-ils amis avec des fumeurs ou les amis de fumeurs se mettent-ils à fumer?)

- Implémentation par le programme SIENA *(Simulation Investigation for Empirical Network Analysis), via* le paquet 'Rsiena' du logiciel R

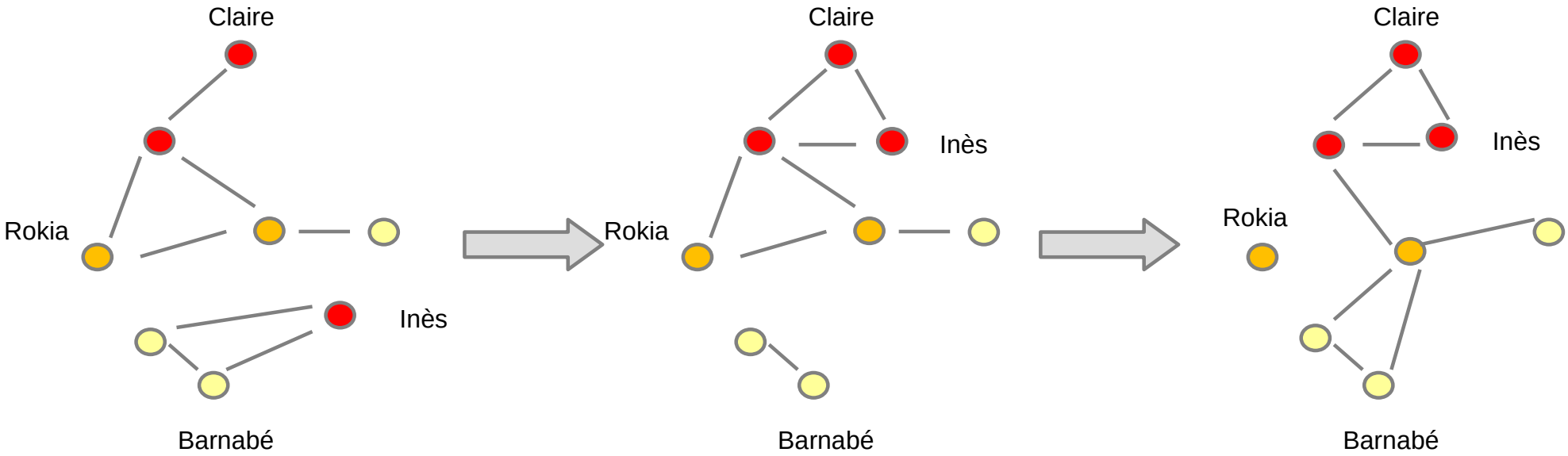
- De nombreuses ressources très bien conçues :

  → site internet : https://www.stats.ox.ac.uk/~snijders/siena/siena.html

  → '*users' manual'* de Rsiena (depuis le site)

  → la mailing list du groupe d'utilisateurs : https://groups.io/g/RSiena

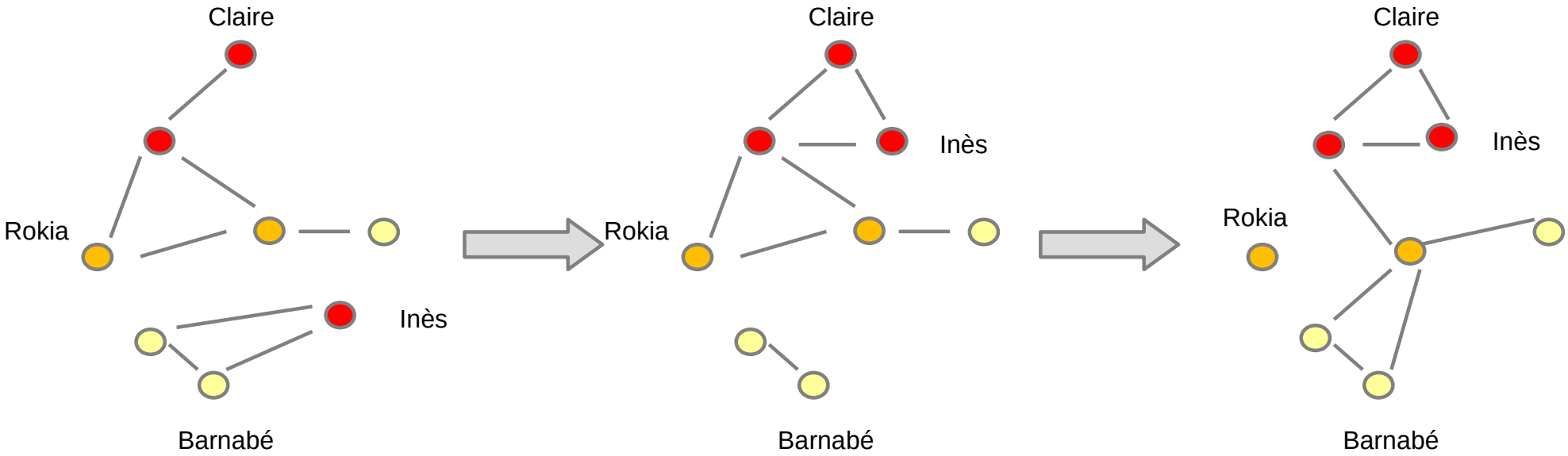
Les réseaux évoluent dans le temps

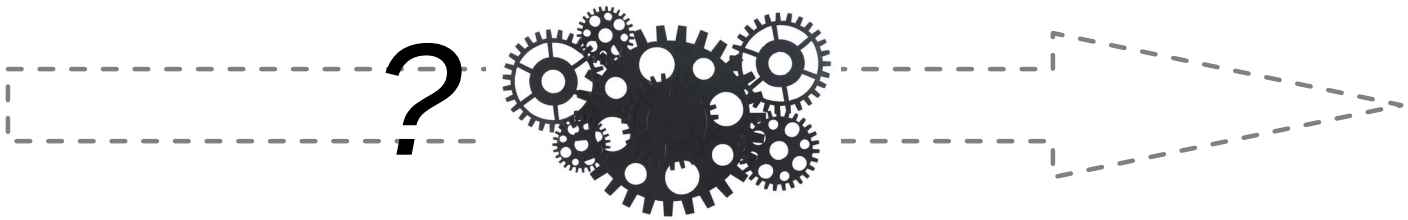


Deux approches envisageables :

- Étudier la séquence des états successifs (~analyse de séquence)
- **Étudier les mécanismes de transitions entre états** (~régression de panel)

L’objectif : inférer un processus continu inobservé d’évolution du réseau





Expliquer l’apparition d’une relation

- Le but de SAOM est de trouver des règles locales et conditionnelles d’apparition, de persistance ou de disparition d’un lien entre deux transitions

  → une amitié a-t-elle plus de chances d’apparaître (ou de persister) entre deux personnes de même genre ?

  → une amitié a-t-elle plus de chances d’apparaître (ou de persister) lorsqu’elle s’inscrit dans un triangle d’amis ?

- Ces règles (des log-odds, en fait) peuvent être ‘atemporelles’, i.e. s’appliquer de façon uniforme sur toutes les transitions.

- SAOM est un modèle ‘*micro-to-macro*’ : on veut trouver la combinaison de **processus locaux** qui rend le mieux compte de la **structure agrégée** du réseau.

II. Présentation du modèle

(1) SAOM comme régression logistique au niveau de la dyade

- Idée de base : on cherche à prédire l'existence (1) ou l'absence (0) d'un lien, à partir d'un modèle *logit*.
- Les unités d'observation ne sont pas les individus, mais les paires (dirigées) d'individus pris deux-à-deux (i.e. les cellules de la matrice de nominations).
- Les variables indépendantes (*covariates*) sont aussi situées au niveau de la dyade.

  → définies *via* les attributs individuels : par ex. effet de *matching,* d'émission et de réception pour un attribut catégoriel (comme le genre).

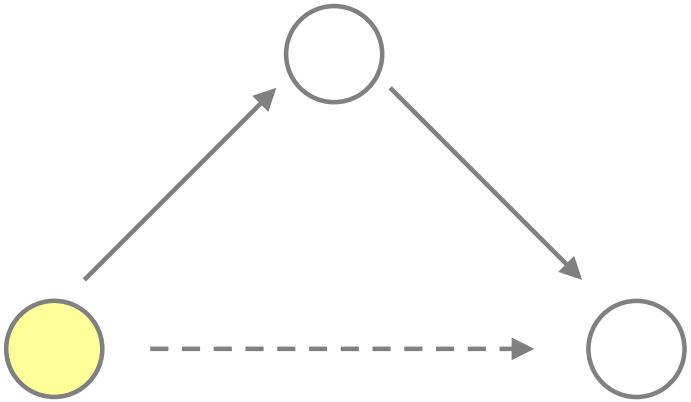
  → définies *via* les configurations de réseaux : par ex. la transitivité

Exemples d’effets (variables indépendantes)

- Homophilie de genre : *ego* et *alter* ont le même genre

- Emission de genre : les filles émettent plus que les garçons

- Transitivité : “les amis de mes amis sont mes amis”



(2) SAOM comme modèle génératif basé sur des simulations

- Problème classique de l'analyse de réseaux : la dépendance entre observations rend les méthodes usuelles d'inférences caduques

  → la valeur des variables indépendantes pour une observation dépend de la valeur des variables dépendantes pour d'autres observations (e.g. transitivité).

- Pour contourner ce problème, on utilise une procédure générative basée sur des simulations.

(2) SAOM comme modèle génératif basé sur des simulations

a) L'utilisateur spécifie un set d'effets qui sont définis par leur *statistiques*, i.e. des propriétés macros du réseau (e.g. nb de triplets fermés ; nb de liens entre élèves de même genre)

b) Chaque effet a aussi un *paramètre* associé, i.e. le changement du log-odd d'occurrence d'un lien lorsqu'il participe à la configuration locale comptée par la statistique (e.g. odds d'un lien qui ferme un triplet transitif; d'un lien entre élèves de même genre)

c) Un algorithme d'optimisation trouve le set de paramètres qui prédit la bonne valeur des statistiques (simulations de réseaux fictifs et ajustement itératif des paramètres)

(2) SAOM comme modèle génératif basé sur des simulations

a) L'utilisateur spécifie un set d'effets qui sont définis par leur *statistiques*, i.e. des propriétés macros du réseau (e.g. nb de triplets fermés ; nb de liens entre élèves de même genre)

b) Chaque effet a aussi un *paramètre* associé, i.e. le changement du log-odd d'occurrence d'un lien lorsqu'il participe à la configuration locale comptée par la statistique (e.g. odds d'un lien qui ferme un triplet transitif; d'un lien entre élèves de même genre)

c) Un algorithme d'optimisation trouve le set de paramètres qui prédit la bonne valeur des statistiques (simulations de réseaux fictifs et ajustement itératif des paramètres)

(Notez qu'une régression logistique classique peut être vue comme une réponse au même genre de problème)

(3) SAOM comme modèle *actor-oriented* et en temps continu

- Ce qu'on vient de voir s'applique à d'autres classes de modèles, notamment transversaux (ERGM).
- SAOM s'intéresse en fait aux probabilités d'apparition ou de disparition des liens entre deux états du réseau (transition).
- Deux spécificités de SAOM :
  - *actor-oriented* = les effets sont toujours envisagés du point de vue d'un acteur focal *ego*
  - modèle en temps continu : on simule un grand nombre de "*mini-steps*" fictifs entre deux observations du réseau, et le modèle *logit* qui détermine la probabilité d'un lien à partir de la valeur des paramètres opère à chacun de ces "*mini-steps*".

(3) Comment SAOM simule-t-il des réseaux à partir d'un set de paramètres ?

- La 1ère observation du réseau (t1) est prise pour acquise. La modélisation cherche à prédire la structure de t2 sachant t1, celle de t3 sachant t2, etc.
- Entre deux observations, le modèle :
  a) Postule un grand nombre de "*mini-steps*" (le nombre est déterminé par une *rate function*).
  b) À chaque *mini-step,* un noeud (ego) est sélectionné au hasard, et a la possibilité d'effectuer 3 actions : émettre un nouveau lien vers un autre noeud (alter), détruire un lien existant, ou ne rien faire.
  c) La probabilité de chaque action est déterminée par un modèle *logit* multinomial, qui utilise la valeur des paramètres (sous forme de coefficients log-odds) (c'est l'*objective function*).
- On obtient des réseaux simulés (sauf pour t1), dont on compare les *statistiques* à celles des vrais réseaux pour ajuster la valeur des paramètres.

La double interprétation des effets

- Interprétation minimaliste : les paramètres estimés capturent simplement une sur- ou une sous-représentation de certaines configurations relationnelles locales. Raisonnement “toutes choses égales par ailleurs” proche d’une régression standard.
- Interprétation réaliste (“*micro-level interpretation*” selon Block et al. 2018) : les paramètres estimés sont interprétés comme de bonnes approximations de **processus sociaux réels** (comportements des individus).
- SAOM a plutôt été conçu pour permettre le second type d’interprétation, mais le modèle est suffisamment flexible pour laisser l’utilisateur libre de ce point de vue.

  \+ il est possible d’interpréter différemment différents effets au sein du même modèle

Mon modèle est-il bon ?

- Comment savoir si le *set* d'effets inclus dans le modèle constitue une représentation suffisante des processus qui nous intéressent ?

  → essentiel pour passer de l'interprétation "minimaliste" à l'interprétation "réaliste" : si je ne mets que quelques effets très simples dans mon modèle, SAOM convergera sans problème et me donnera les paramètres associés, mais l'information sociologique sera pauvre.

- Pas de solution miracle. On se base sur (1) des considérations théoriques et la littérature et (2) des *goodness-of-fit*, pour savoir si le modèle représente correctement certaines propriétés des réseaux **non-inclues sous forme de statistiques dans le modèle**.

III. Exemple d’application : l’homophilie sociale au collège

Données et questions de recherche

- Réseaux d'amitié entre des élèves de collège observés à 4 reprises (tous les 6 mois, entre la 6e à la 4e).

  Deux établissements différents (Paris 1 et Savoie 1).

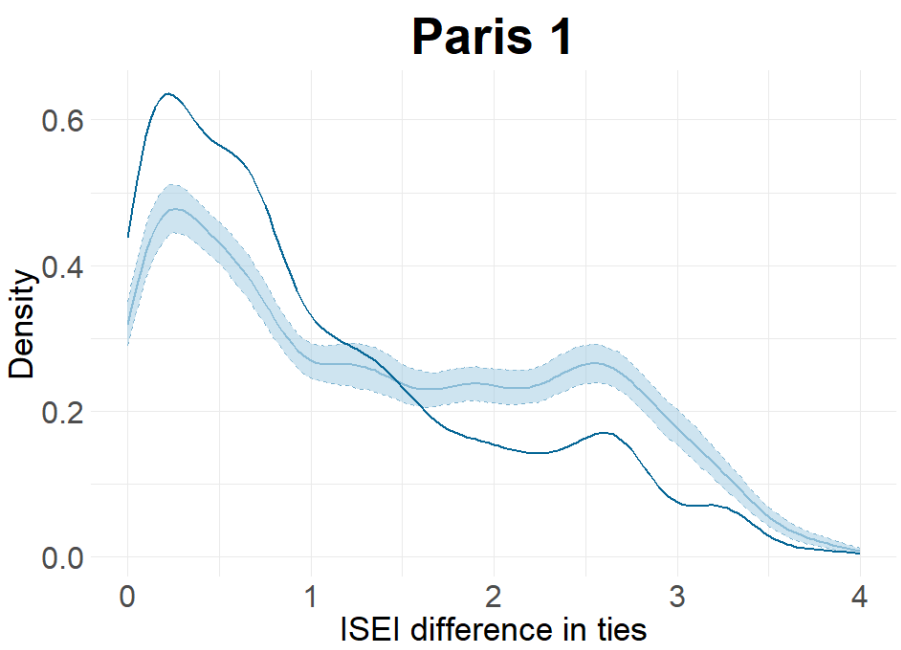
- Homophilie sociale = tendance à avoir des amis socialement similaires (origine sociale mesurée par la profession des parents).

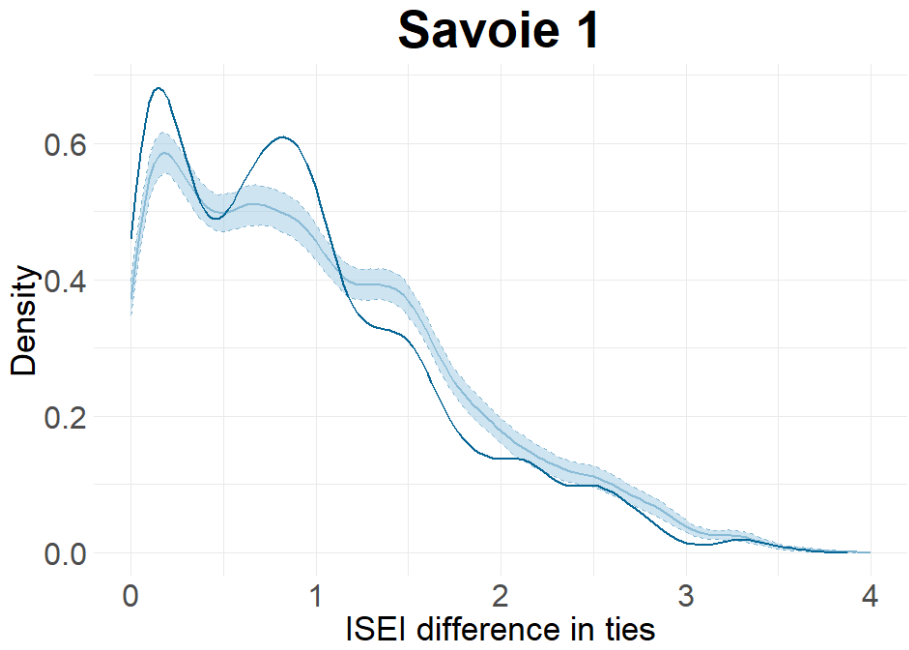
- Questions descriptives : y a-t-il de l'homophilie ? Evolue-t-elle dans le temps ?

- Questions explicatives : quels processus relationnels expliquent l'homophilie ? Les élèves ont-ils une préférence pour les amis socialement similaires ? Ces processus changent-ils dans le temps ?

Il y a de l’homophilie sociale au collège

Tests de permutations – Distribution de la distance sociale entre amis dans les réseaux permutés (bleu clair) et observés (bleu foncé)
(t=1 à t=4)



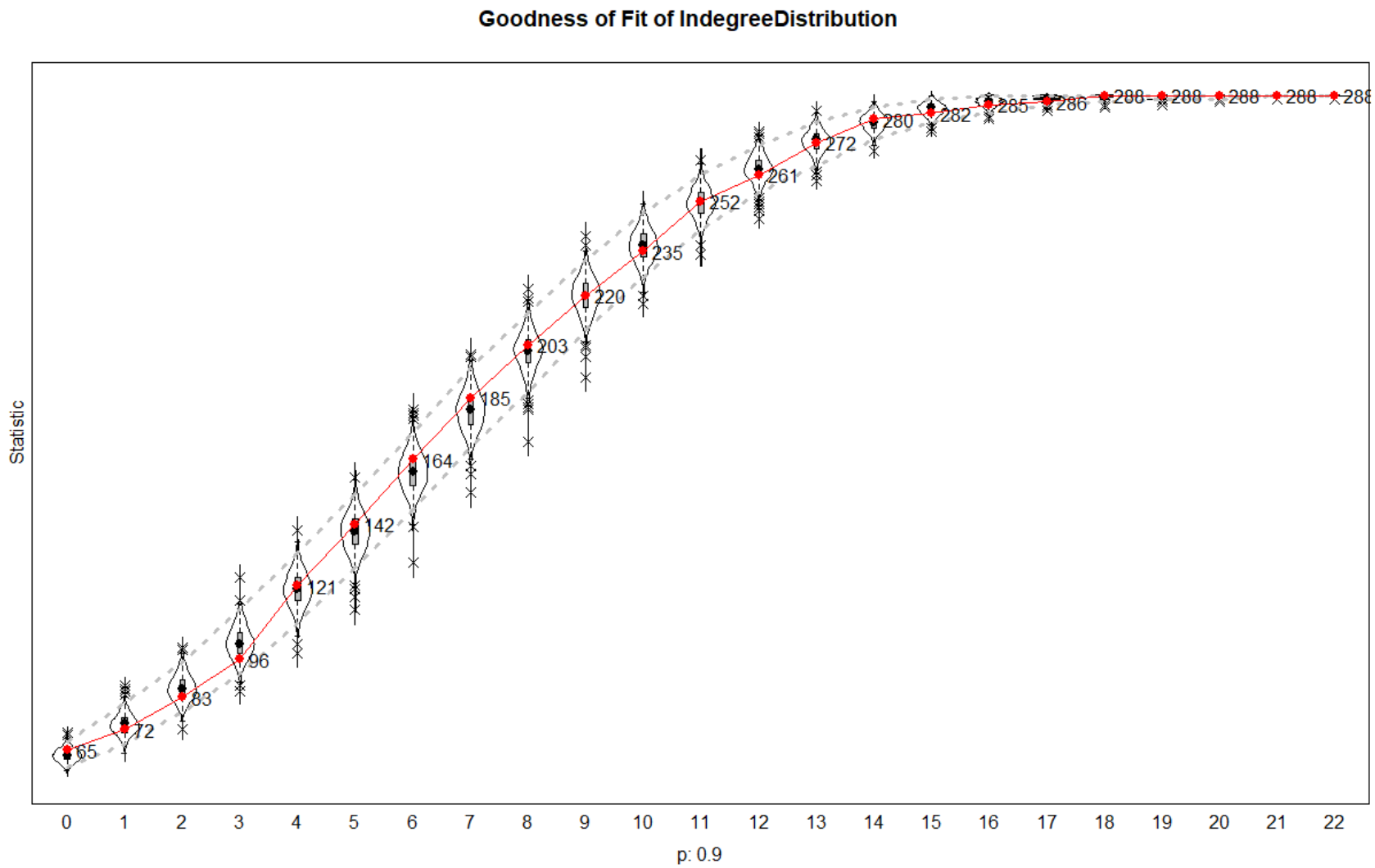


SAOMs estimés

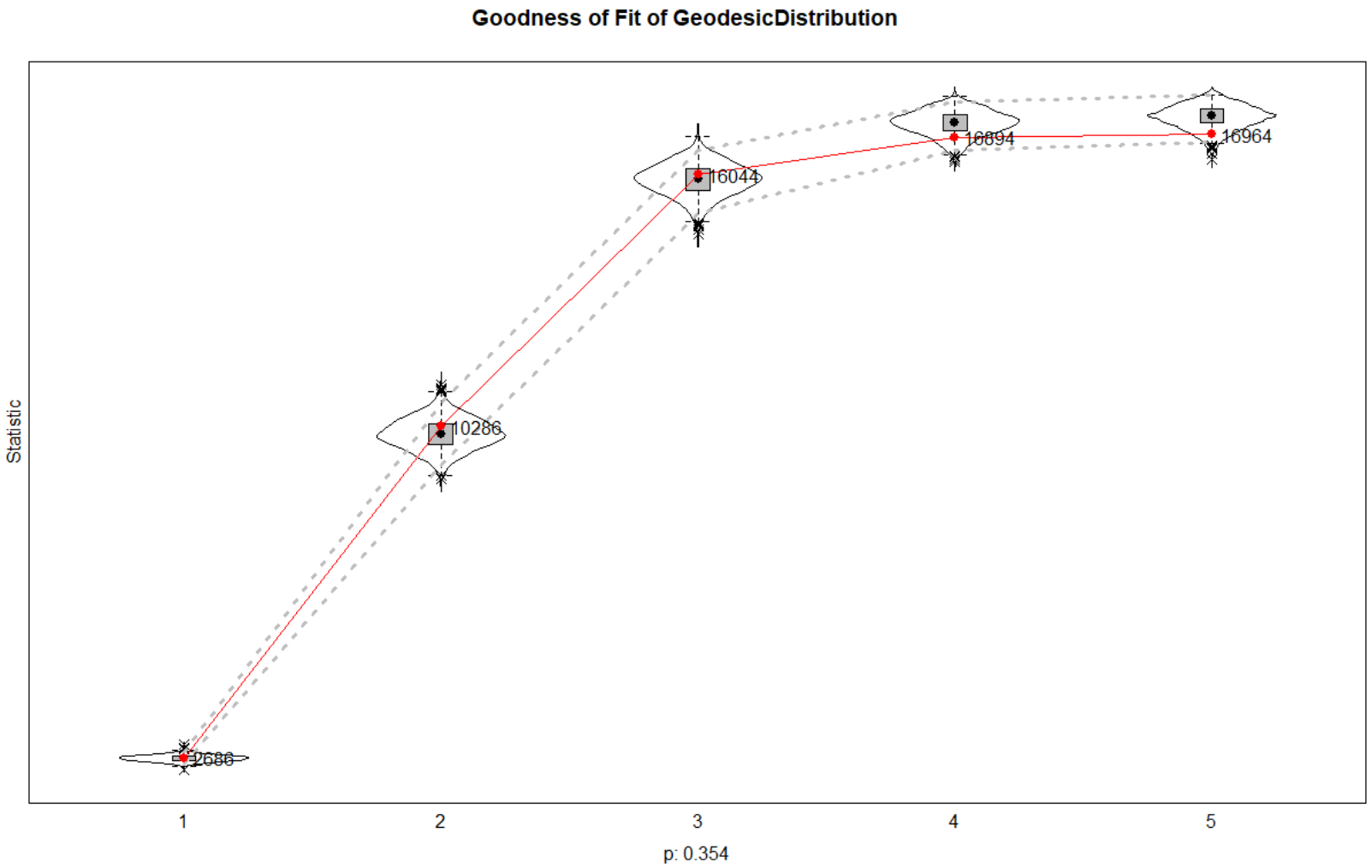
(nb: seule une petite partie des effets inclus est présentée)

	Paris 1			Savoie 1		
Effect Name	est.	sgn.	s.e.	est.	sgn.	s.e.
outdegré (densité)	-2.88	***	(0.14)	-2.84	***	(0.07)
reciprocité	2.41	***	(0.19)	2.42	***	(0.12)
gwespFF (transitivité)	0.89	***	(0.12)	0.66	***	(0.11)
gwespFB (transitivité)	0.33	***	(0.13)	0.56	***	(0.14)
gwespFF * réciprocité	-0.59	***	(0.10)	-0.67	***	(0.09)
Similarité - ISEI	0.21	*	(0.12)	-0.04		(0.08)
Matching - Ethnicité	0.15	**	(0.06)	0.18	***	(0.05)
Similarité – Résultats scolaires	0.53	***	(0.17)	0.20	*	(0.11)
Matching - Genre	0.46	***	(0.06)	0.28	***	(0.03)
Matching – Classe d'école (actuelle)	0.48	***	(0.07)	0.71	***	(0.04)
Matching – Classe d'école (an passé)	0.17	**	(0.06)	0.19	***	(0.03)
Matching – section internationale	0.17		(0.17)	-0.15		(0.16)
Matching – section normale	0.56	***	(0.13)	0.41	***	(0.12)
Matching – ancienne école primaire	0.04		(0.06)	0.26	***	(0.04)
Temps de marche entre domiciles	-0.29		(0.26)	-0.06	**	(0.02)

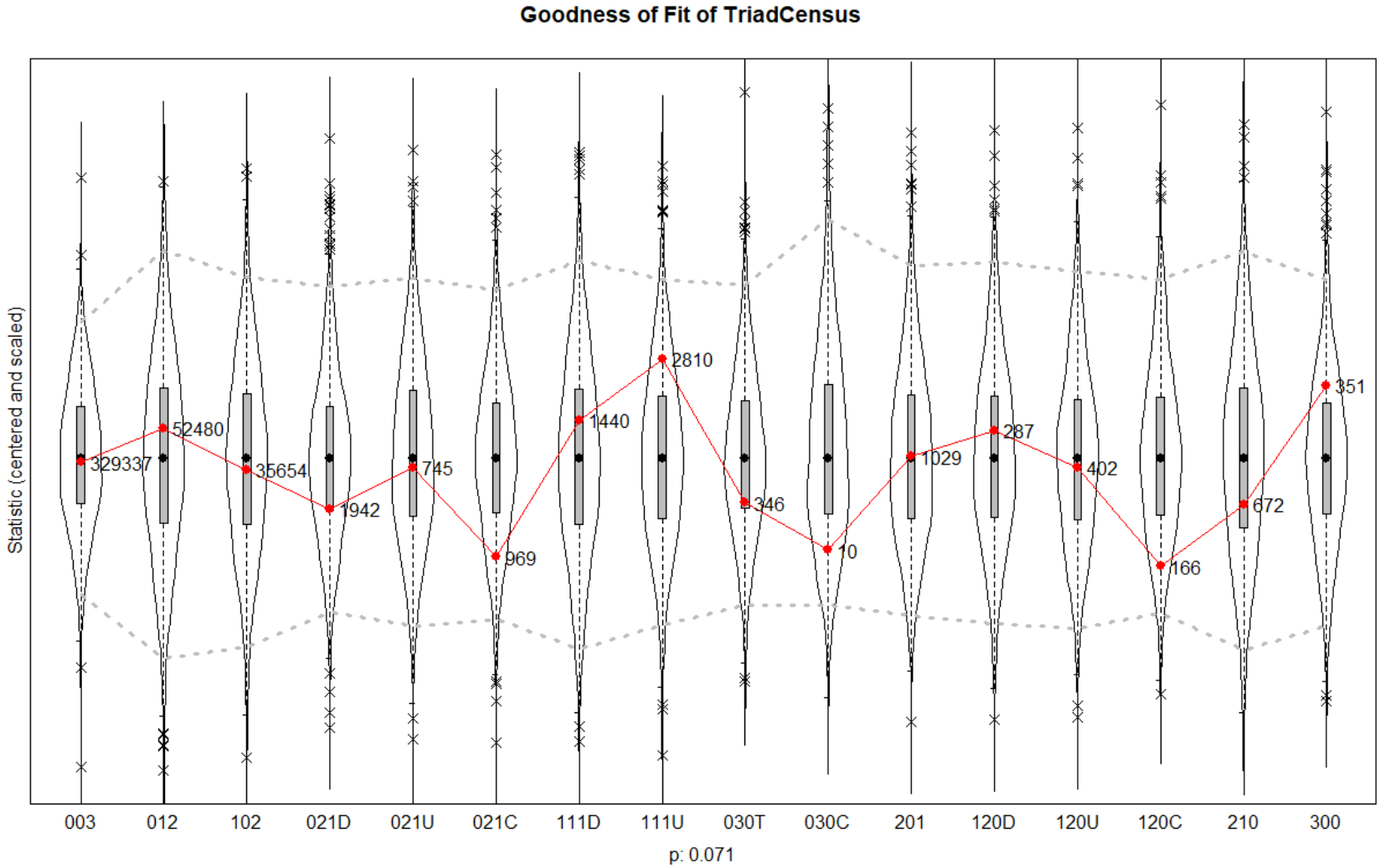
Goodness-of-fit (Paris 1)



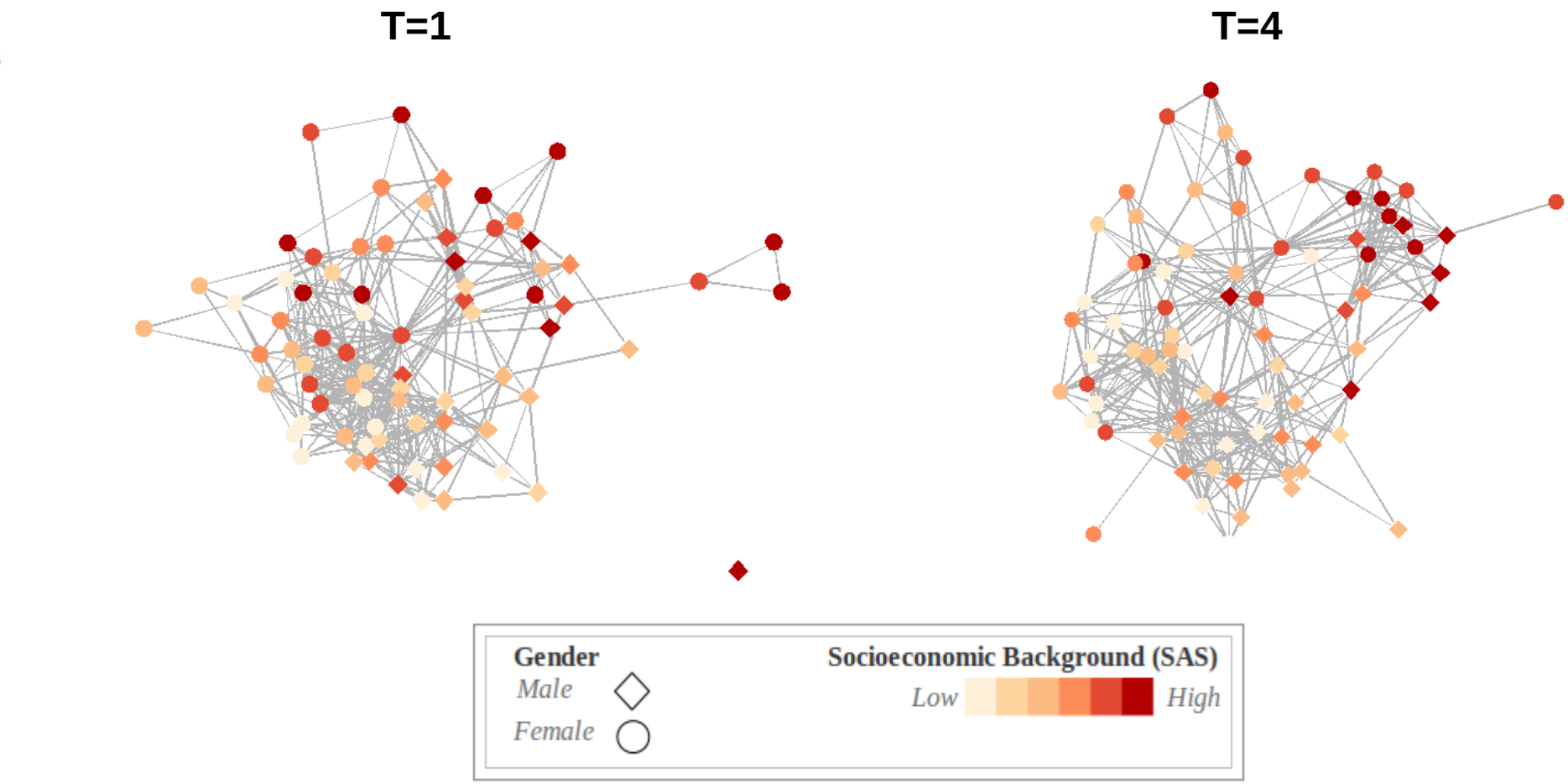
Goodness-of-fit (Paris 1)



Goodness-of-fit (Paris 1)



Dans un des deux collèges, l'homophilie augmente dans le temps



Test d'hétérogénéité temporelle des paramètres

	Paris 1		Savoie 1	
	chi-sq.	*p-value*	*chi-sq.*	*p-value*
reciprocité	0,2	0,91	15,11	0
gwespFF (transitivité)	0,46	0,8	0,29	0,87
gwespFB (transitivité)	0,16	0,92	0,42	0,81
gwespFF * réciprocité	0,47	0,79	11,46	0
Similarité - ISEI	2,27	0,32	1,92	0,38
Matching - Ethnicité	2,06	0,36	1,37	0,5
Similarité – Résultats scolaires	0,07	0,97	5,78	0,06
Matching - Genre	1,56	0,46	12,33	0
Matching – Classe d'école (actuelle)	0,1	0,95	1,71	0,43
Matching – section internationale	0,15	0,93	0,14	0,93
Matching – section normale	0,88	0,64	1,47	0,48
Temps de marche entre domiciles	0,17	0,92	20,69	0
Matching – ancienne école primaire	0,07	0,97	0,05	0,98

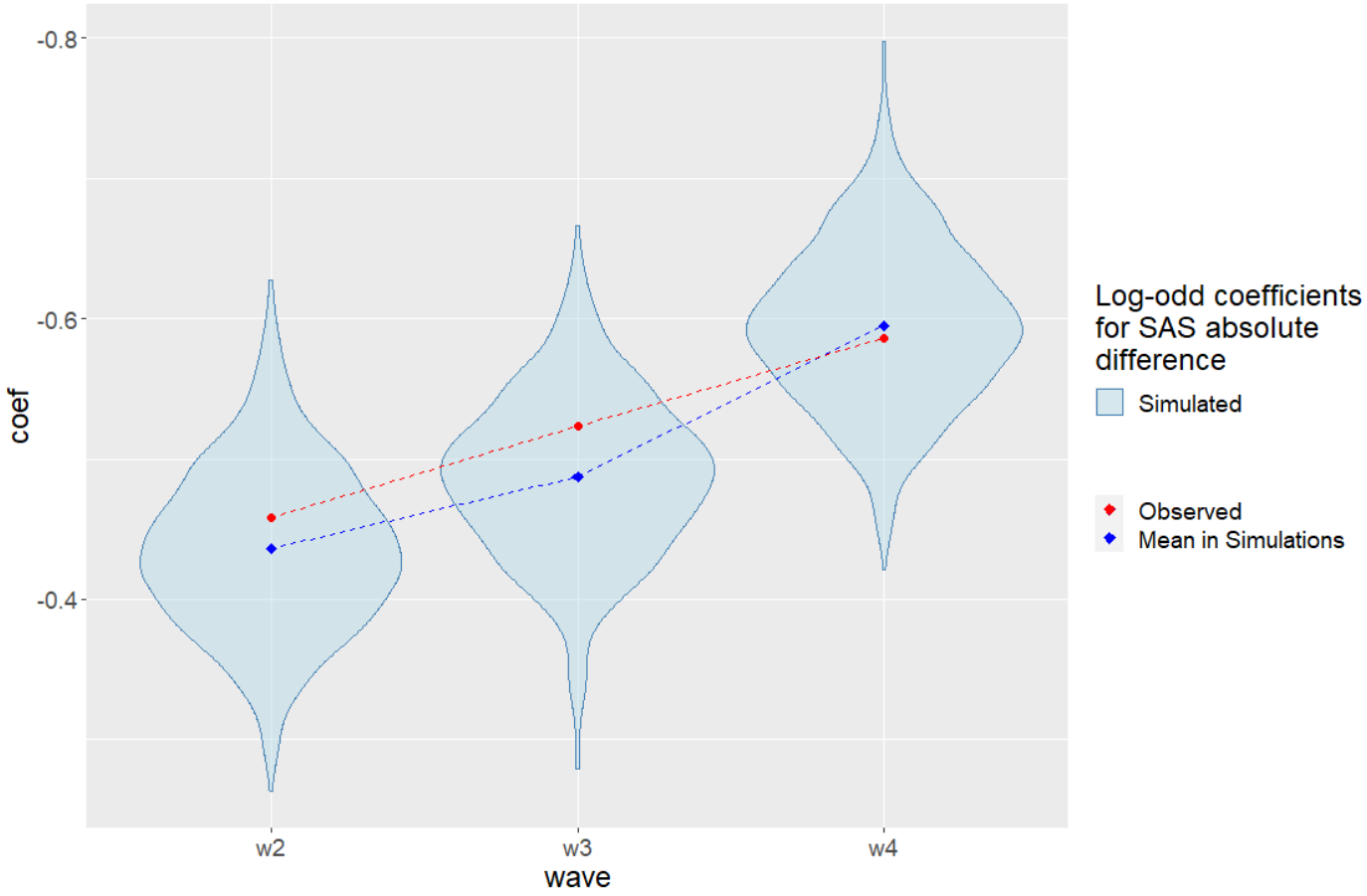
"Time heterogeneity tests" sous RSiena - voir Lospinoso et al. 2011

Test d’hétérogénéité temporelle des paramètres

- Si certains paramètres présentent de l’hétérogénéité, cela veut dire que le modèle **prédit trop d’une statistique à certaines vagues et pas assez à d’autres.**

- On utilise alors des “*time dummies*”, i.e. un effet d’interaction entre un effet et le temps d’observation (= la valeur du paramètre ne sera pas la même pour la transition t1/t2 que pour la transition t2/t3)
- En l’occurrence, pas nécessaire : les effets ne présentent pas d’hétérogénéité (cf. slide précédente).

Des processus constants prédisent bien l'accroissement dans le temps (Paris 1)



Références

- Snijders T.A.B., 2017, « Stochastic Actor-Oriented Models for Network Dynamics », Annual Review of Statistics and Its Application, 4, 1, p. 343-363.
- Block P., Koskinen J., Hollway J., Steglich C., Stadtfeld C., 2018, « Change we can believe in: Comparing longitudinal network models on consistency, interpretability and predictive power », Social Networks, 52, p. 180-191.