Inférence de dates d'activité à partir d'un réseau d'interactions datées

Fabrice Rossi & Pierre Latouche

SAMM EA 4543

JDS 2013

General setting

Decorated interaction networks

- interaction between “actors”
- each interaction is described by some characteristics
- multiple interactions between the same actors

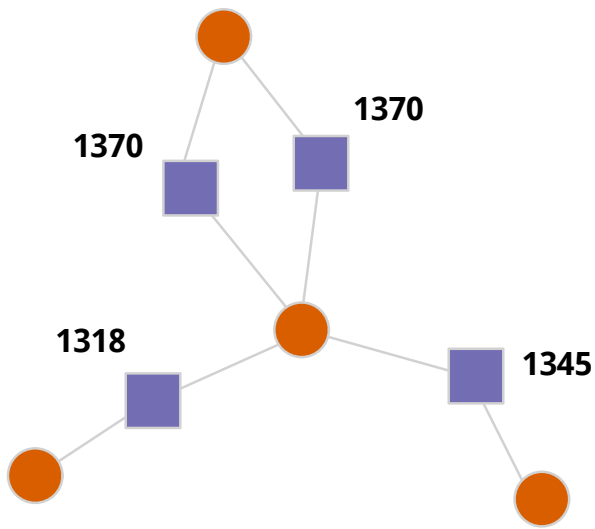
General setting

Decorated interaction networks

- interaction between “actors”
- each interaction is described by some characteristics
- multiple interactions between the same actors

Ancient Notarial Acts

- very precise recording of transactions about long lasting goods (lands, houses, etc.)
- not so precise description of the persons involved in the transactions (e.g., only first names)



Goal

Inference about actors

- propagate information associated to interactions to actors
- for instance with notarial acts:
  - dates of acts $\Rightarrow$ living period
  - geographical position of the goods $\Rightarrow$ living area
  - status in unbalanced interactions $\Rightarrow$ social status

Goal

Inference about actors

- propagate information associated to interactions to actors
- for instance with notarial acts:
  - dates of acts $\Rightarrow$ living period
  - geographical position of the goods $\Rightarrow$ living area
  - status in unbalanced interactions $\Rightarrow$ social status

Timestamped Interaction Network

- temporal decoration: a time stamp is associated to each interaction
- the network may outlives the actors (notarial acts)
- estimate a central date of activity for each actor, based on the time stamps of its interactions
- an activity interval can be estimated in some situations

Local solution

Simple local solution

- “propagate” interaction associated characteristics to the actors
- summarize the data (if needed)

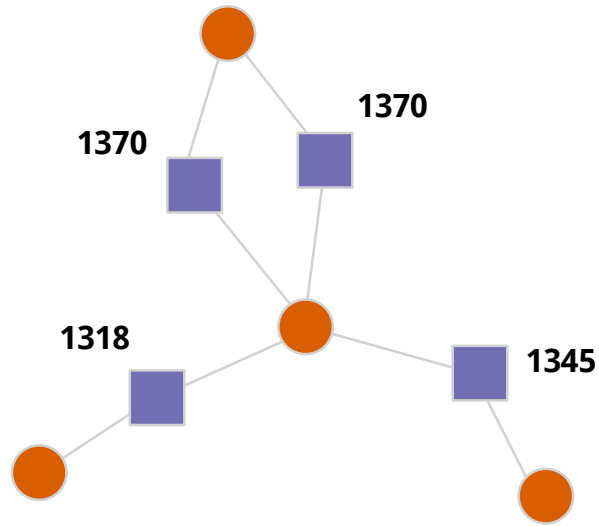
Local solution

Simple local solution

- “propagate” interaction associated characteristics to the actors
- summarize the data (if needed)

Activity date

- central actor : 1318, 1345, 1370, 1370, with an average of $\sim$ 1351
- other actors : their unique (or repeated) date



Drawbacks

- based only on local interactions not at all on non interaction
- summarizes the characteristics but not the network

Global solution

Consistency hypotheses

- interaction characteristics are close to actors characteristics
- interactions happen preferably between actors who share similar characteristics

Global solution

Consistency hypotheses

- interaction characteristics are close to actors characteristics
- interactions happen preferably between actors who share similar characteristics

Generative approach

- actor i has characteristics $Z_i \in \mathcal{Z}$ (dissimilarity space)
- $i \leftrightarrow j$ with some probability decreasing with $d(Z_i, Z_j)$
- if $i \leftrightarrow j$, then the decoration is generated
  - "around" Z_i and Z_j (same space $\mathcal{Z}$)
  - or at least in a way "consistent" with Z_i and Z_j (possible in another space)

Technicalities (1/2)

General Model (single interaction)

- data: A adjacency matrix, D decoration table
- parameters: $(Z_i)_{1 \leq i \leq N}$, θ
- likelihood:

$$p(A, D|Z, \theta) = \prod_{i \neq j, A_{ij}=0} P(A_{ij} = 0|Z_i, Z_j, \theta) \times \prod_{i \neq j, A_{ij}=1} P(A_{ij} = 1|Z_i, Z_j, \theta) p(D_{ij}|A_{ij} = 1, Z_i, Z_j, \theta).$$

Technicalities (1/2)

General Model (single interaction)

- data: A adjacency matrix, D decoration table
- parameters: $(Z_i)_{1 \leq i \leq N}$, θ
- likelihood:

$$p(A, D|Z, \theta) = \prod_{i \neq j, A_{ij}=0} P(A_{ij} = 0|Z_i, Z_j, \theta) \times \prod_{i \neq j, A_{ij}=1} P(A_{ij} = 1|Z_i, Z_j, \theta) p(D_{ij}|A_{ij} = 1, Z_i, Z_j, \theta).$$

Numerical decorations

- logistic connection model (related to Hoff et al., 2002):

$$\log \frac{P(A_{ij} = 1|Z_i, Z_j, \alpha, \beta)}{P(A_{ij} = 0|Z_i, Z_j, \alpha, \beta)} = \alpha - \beta \|Z_i - Z_j\|^2,$$

- Gaussian decoration: $D_{ij}|Z_i, Z_j, \Sigma \sim \mathcal{N}\left(\frac{Z_i + Z_j}{2}, \Sigma\right)$.

Technicalities (2/2)

Logistic connection model

- connection probability: $P(A_{ij} = 1 | Z_i, Z_j, \alpha, \beta) = \dfrac{1}{1 + e^{\beta \|Z_i - Z_j\|^2 - \alpha}}$
- $\frac{1}{1+e^{-\alpha}}$: maximal density of the interaction network
- $\frac{1}{\beta}$: interaction “radius”

Technicalities (2/2)

Logistic connection model

- connection probability: $P(A_{ij} = 1 | Z_i, Z_j, \alpha, \beta) = \dfrac{1}{1 + e^{\beta \|Z_i - Z_j\|^2 - \alpha}}$
- $\frac{1}{1+e^{-\alpha}}$: maximal density of the interaction network
- $\frac{1}{\beta}$: interaction “radius”

Timestamps

- $Z_i \in \mathbb{R}$: (central) activity date, $D_{ij} \sim \mathcal{N}\left(\frac{Z_i + Z_j}{2}, \sigma^2\right)$
- $\frac{1}{\beta}$ and σ: lifespan of actors

Technicalities (2/2)

Logistic connection model

- connection probability: $P(A_{ij} = 1 | Z_i, Z_j, \alpha, \beta) = \dfrac{1}{1 + e^{\beta \|Z_i - Z_j\|^2 - \alpha}}$
- $\frac{1}{1+e^{-\alpha}}$: maximal density of the interaction network
- $\frac{1}{\beta}$: interaction "radius"

Timestamps

- $Z_i \in \mathbb{R}$: (central) activity date, $D_{ij} \sim \mathcal{N}\left(\frac{Z_i+Z_j}{2}, \sigma^2\right)$
- $\frac{1}{\beta}$ and σ: lifespan of actors

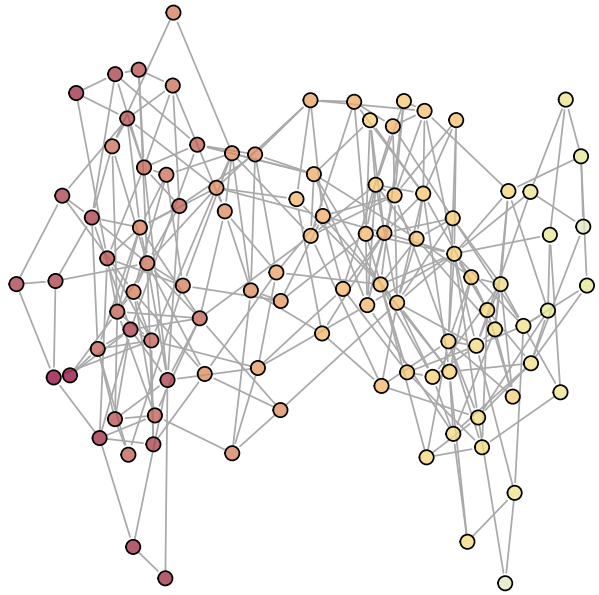
Estimation

- here by maximum likelihood: non convex/concave optimization problem, solved by standard techniques
- other techniques could be used

Experiments

Validation of the model

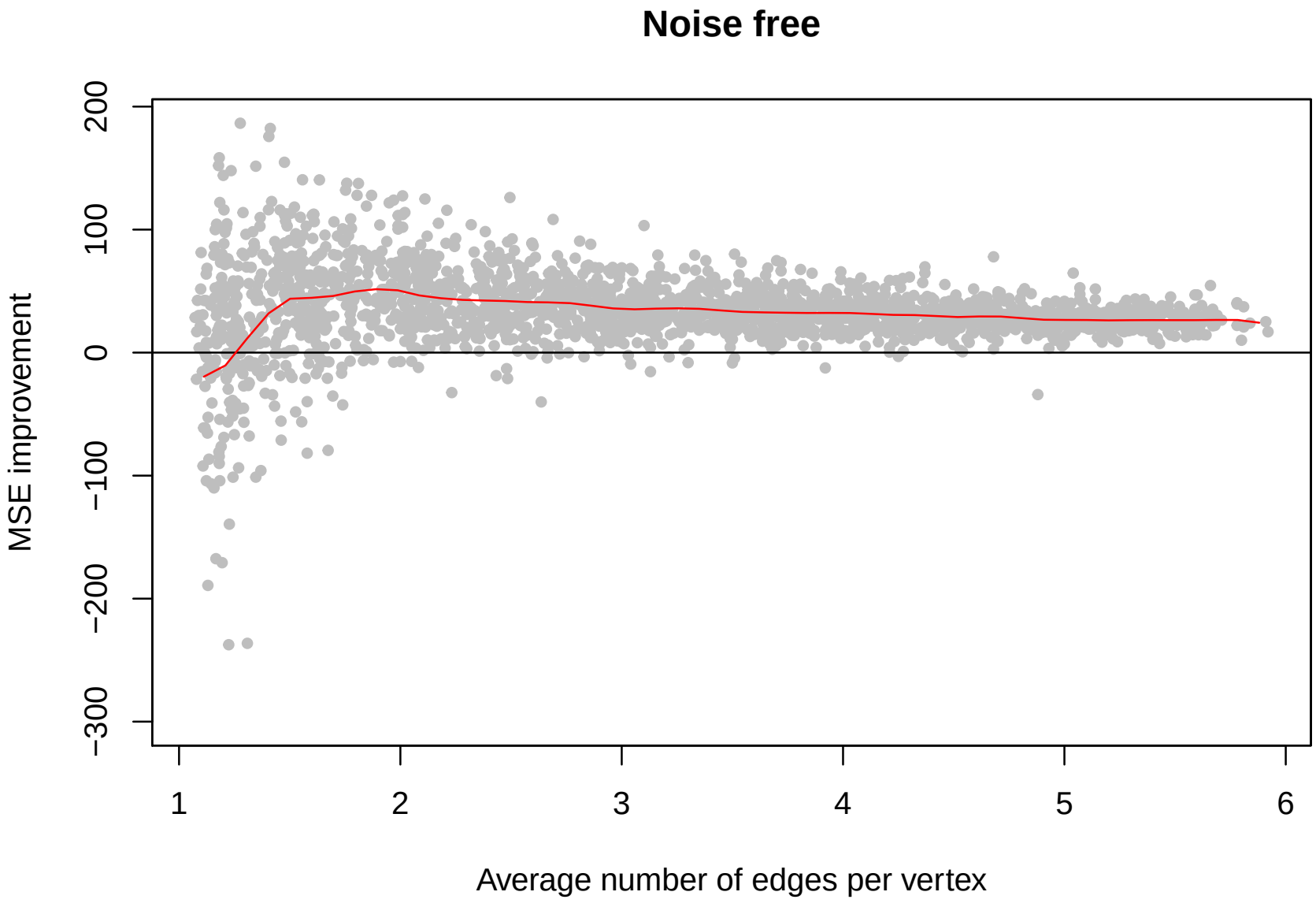
- data generated according to the model
- realistic values for β and $\sigma = 20$ (lifespan ~ 80)
- α varies to simulate different densities
- the Z_i are uniformly distributed in $[1200, 1400]$ (small size networks with 100 agents)



Quality criterion

- mean square error (MSE) between true Z_i and estimated one
- baseline: local average
- quality: reduction in MSE with respect to the baseline

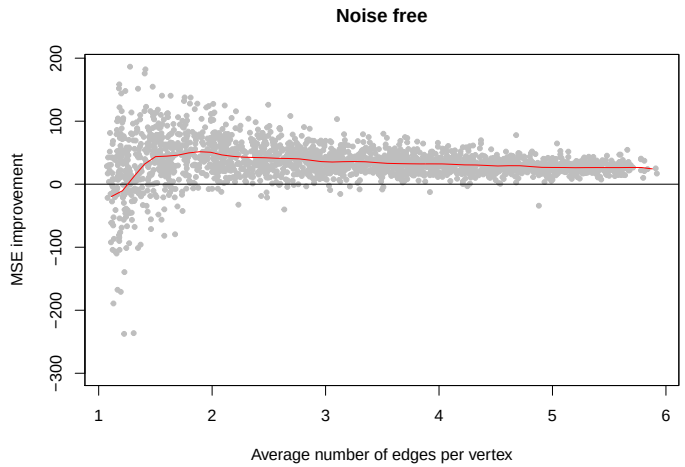
Results


Noise free
MSE improvement
200
100
0
−100
−200
−300
1
2
3
4
5
6
Average number of edges per vertex

Results

Summary

- roughly 2200 networks generated
- break even at $\sim$ 1.3 interaction per actor
- (almost) systematic improvement after 2 interactions per actor
- some convergence issues (easy to spot)



Robustness

- very bad for low density network: below 1.1 interaction per actor, Z_i estimations are frequently very bad
- good with respect to misspecification of the date distribution, e.g. using a uniform date distribution rather than a Gaussian one (see the paper)

Noisy networks (1/2)

Imperfect data sets

- decorations are assumed to be exact or at least precise
- but they can be attached to a wrong pair of actors

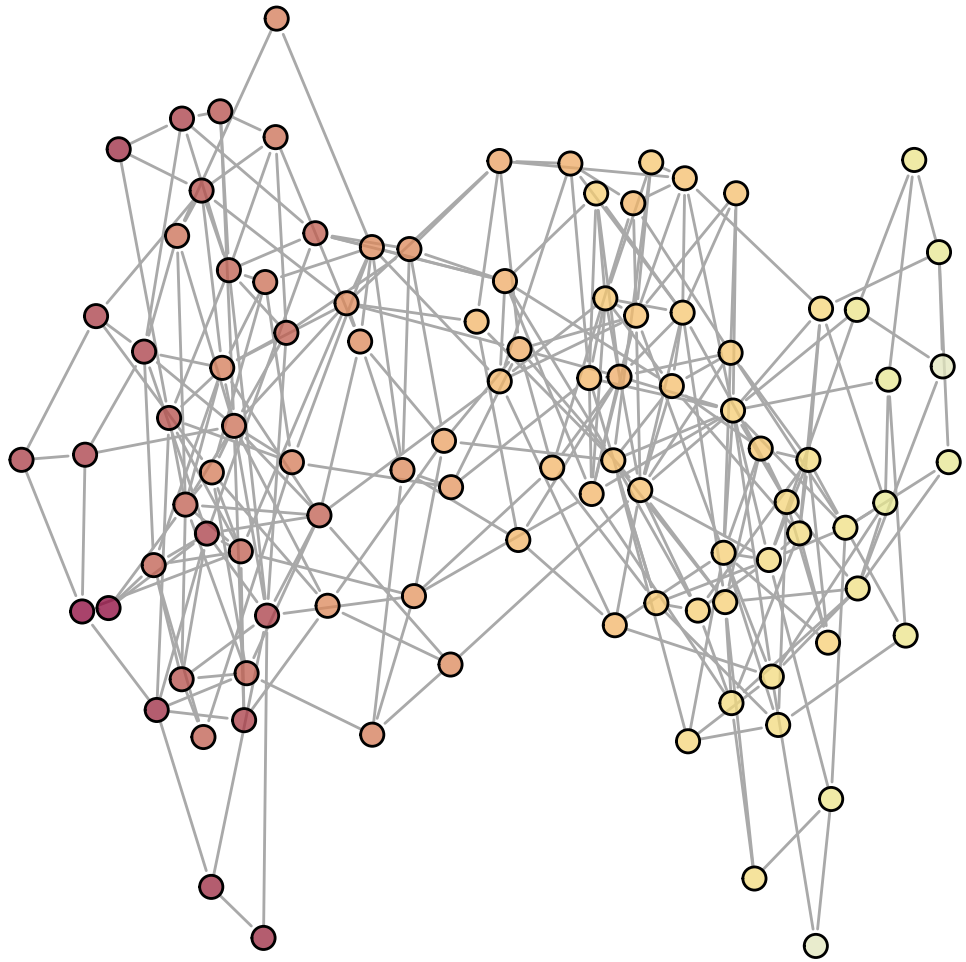
Motivation

- notarial acts were exact at their redaction time
- but we miss accurate registry of the persons, in particular, many persons share the same name, which are the unique identifiers in the acts
- this leads to ambiguous assignment of persons to acts

Noisy networks (2/2)

Simulated by random rewiring

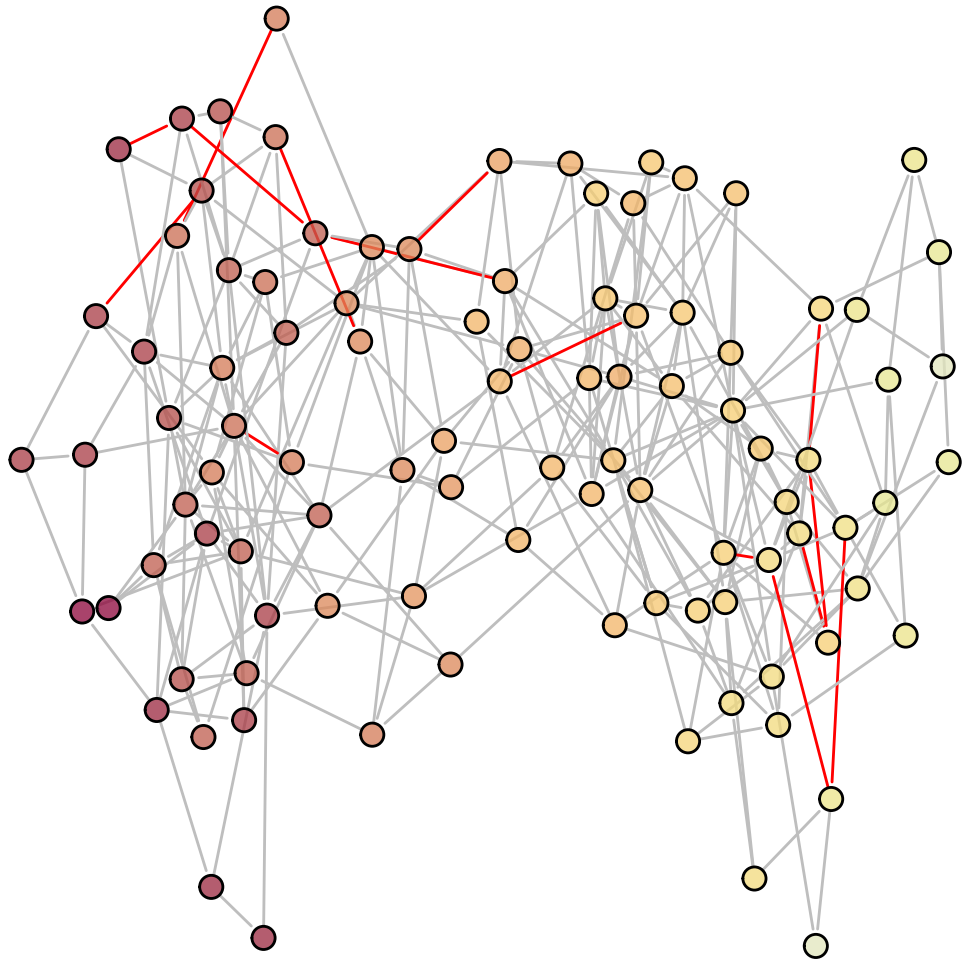
- generate a network



Noisy networks (2/2)

Simulated by random rewiring

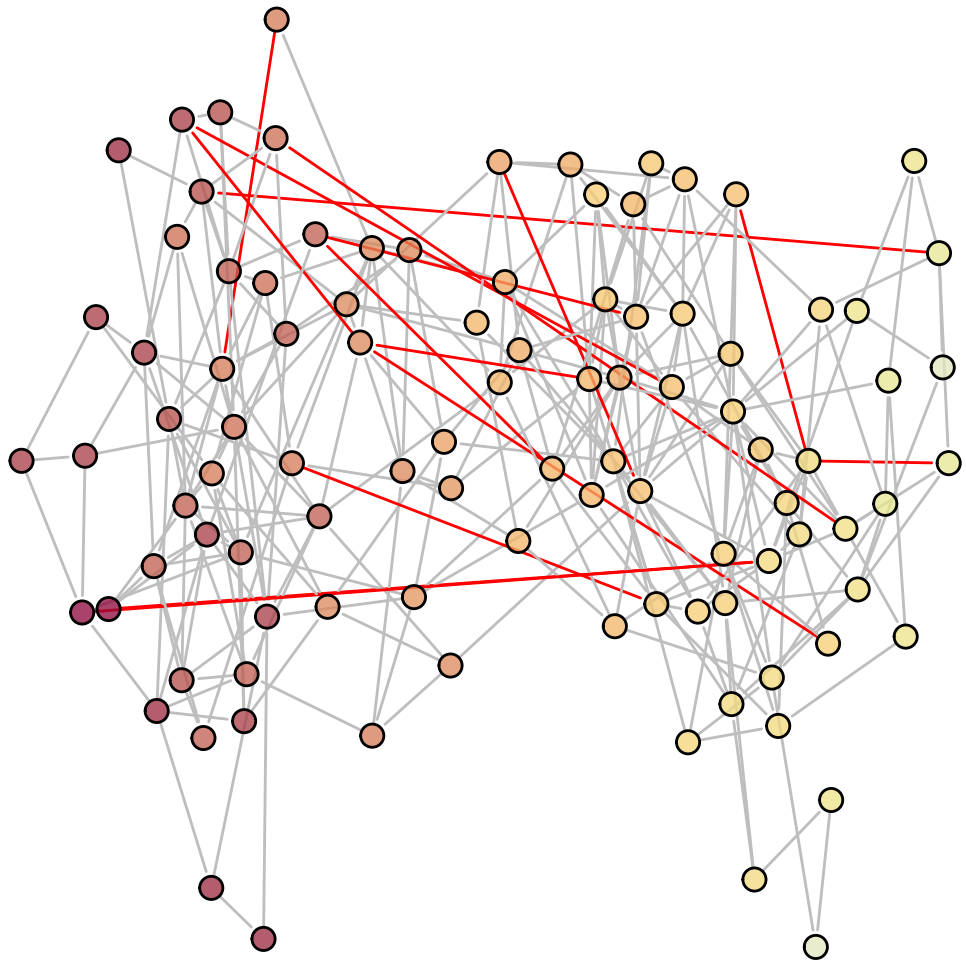
- generate a network
- select (randomly) an edge to rewire



Noisy networks (2/2)

Simulated by random rewiring

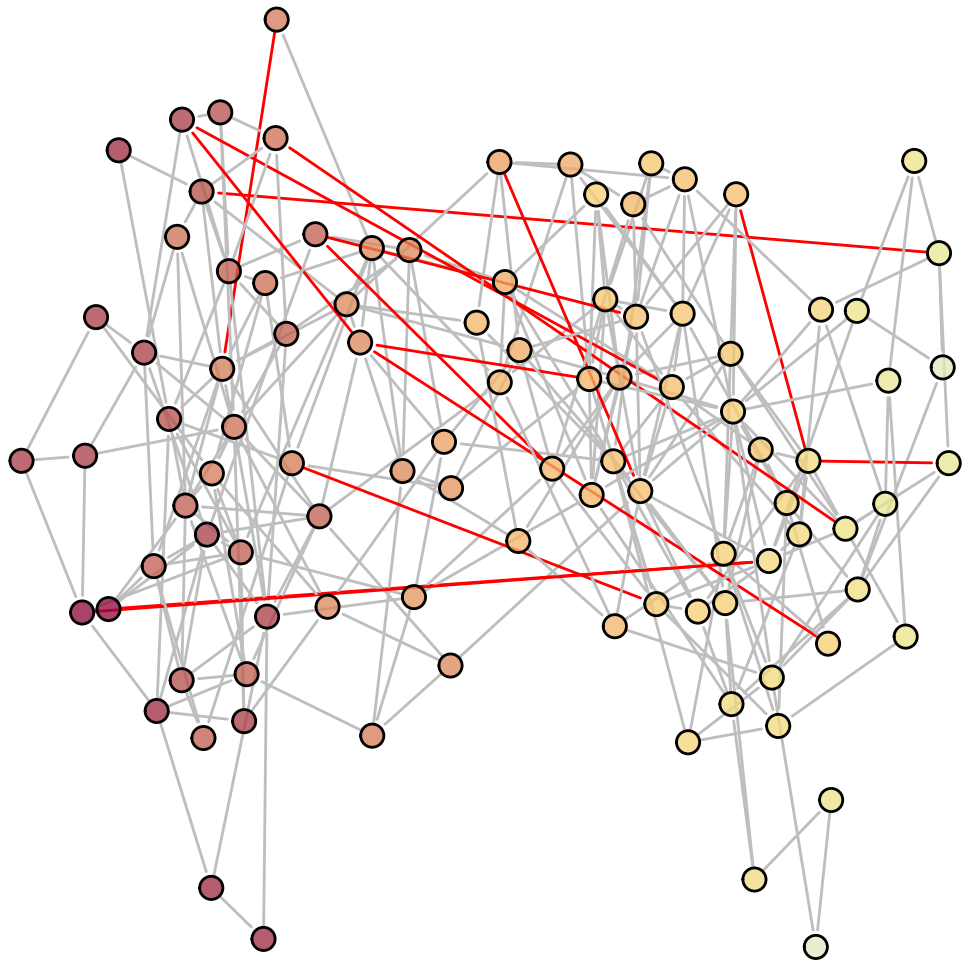
- generate a network
- select (randomly) an edge to rewire
- chose (randomly) a new “ending” object



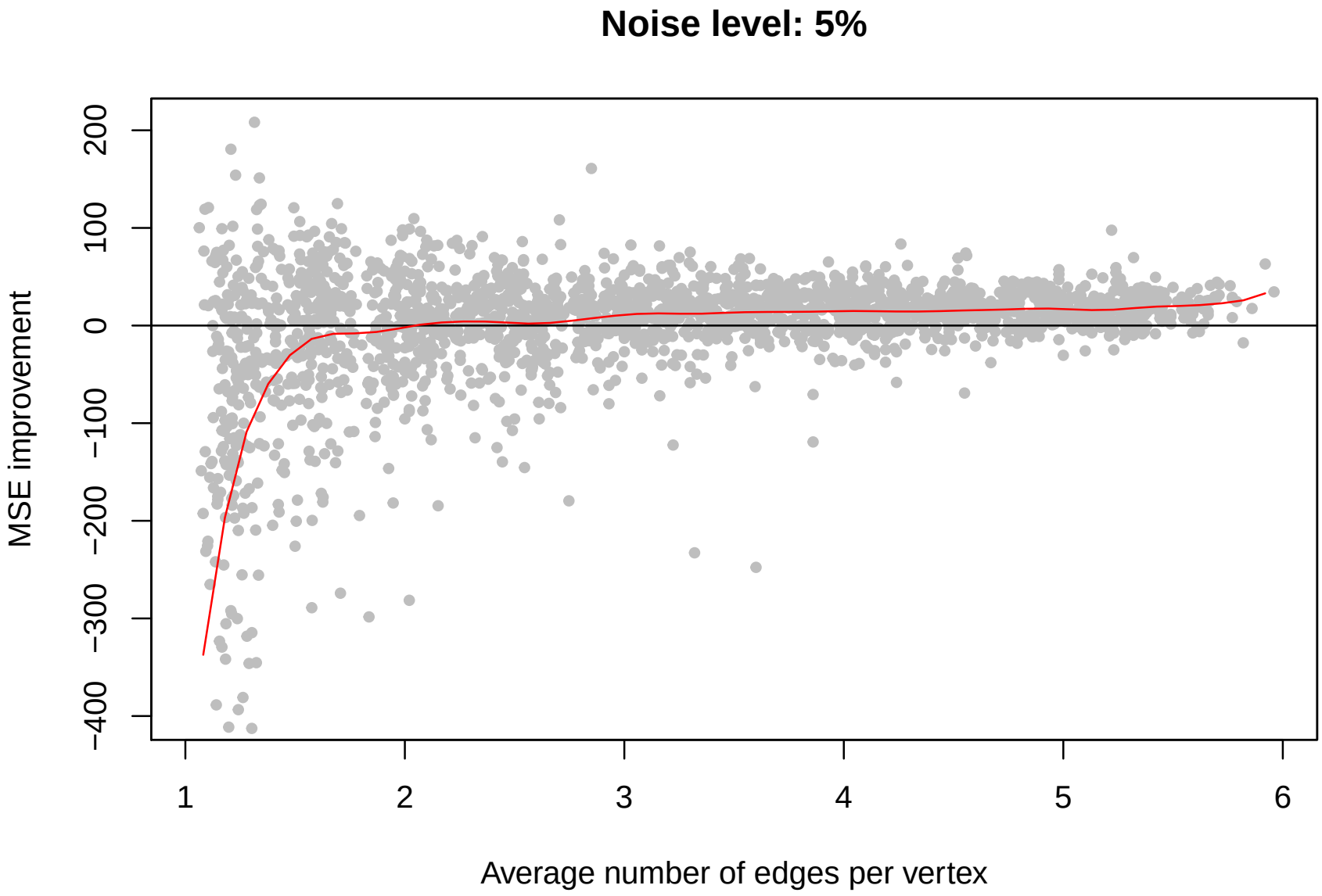
Noisy networks (2/2)

Simulated by random rewiring

- generate a network
- select (randomly) an edge to rewire
- chose (randomly) a new “ending” object
- keep the original date!



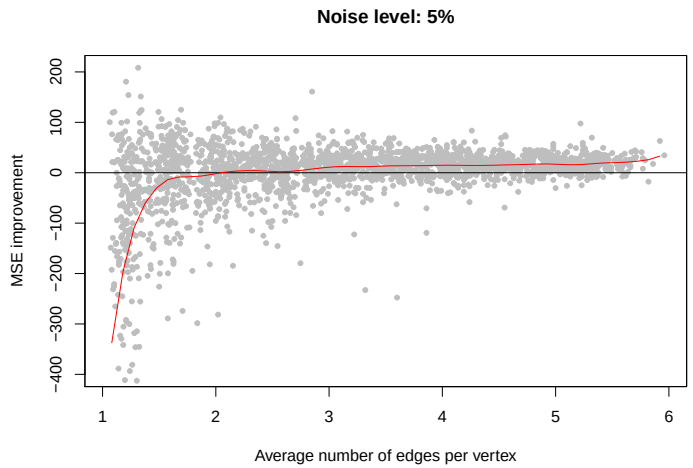
Results


Noise level: 5%
MSE improvement
200
100
0
−100
−200
−300
−400
1
2
3
4
5
6
Average number of edges per vertex

Results

Summary

- roughly 2200 networks generated, 5 % of edge rewiring
- break even at $\sim$ 2.1 interaction per actor
- good behavior after 3 interactions per actor
- more convergence issues (easy to spot)



Robustness

- a low level of noise (e.g. 1 %) has almost no effect on the estimation
- a high level of noise (10 %) has strong adverse effects

Summary and conclusion

A generative model for decorated graphs

- introduces a way to “push” edges decorations to agents
- estimate characteristics that explain both the network and the decorations
- exhibit some robustness to misspecification

Future work

- real world data
- mixture model: generative model + a noise component (ongoing work)
- more complex model: explains the network with the characteristics but also with some structural properties (e.g., block model like)