Morpho-statistical description of networks through graph modelling and Bayesian inference

Quentin Laporte-Chabasse, Marianne Clausel, Radu Stoica, François Charoy, Gérald Oster

Morpho-statistical description of networks through graph modelling and Bayesian inference

Quentin Laporte-Chabasse
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
Marianne Clausel and Radu S. Stoica
Université de Lorraine, IECL, CNRS, F-54000 Nancy, France
François Charoy and Gérald Oster
Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

Summary.
Collaborative graphs are relevant sources of information to understand behavioural tendencies of groups of individuals. Exponential Random Graph Models (ERGMs) are commonly used to analyze such social processes including dependencies between members of the group. Our approach considers a modified version of ERGMs, modeling the problem as an edge labelling one. The main difficulty is inference since the normalising constant involved in classical Markov Chain Monte Carlo (MCMC) approaches is not available in an analytic closed form.
The main contribution is to use the recent ABC Shadow algorithm. This algorithm is built to sample from posterior distributions while avoiding the previously mentioned drawback. The proposed method is illustrated on real data sets provided by the Hal platform and provides new insights on self-organised collaborations among researchers.

1. Introduction

Networks are widely studied mathematical objects (Bollobás, 2013, chapter 5). They describe molecular interactions, relationships between individuals in a social application, collaborative links among organisations, etc.

For example, when different organisations collaborate to produce new scientific results, a part of these results are presented through scientific papers. The publication process induces a network describing interactions among the organisations involved in this process. Here, the network is made of the co-authorship relation of researchers belonging to the different organisations Glänzel (2001).

Let us consider the set of scientific publications produced by LORIA†, during the year 2018. The laboratory is organised in 28 scientific teams. The data was gathered from the open publication archive Hal (`data.archives-ouvertes.fr`). We collected all the publications submitted during 2018 with at least one author member of Loria. The co-authorship network is represented by a graph structure, as shown in the Figure 1. The nodes of the graph are the researchers. An edge of the graph represents the link between two researchers who collaborated in 2018. Nodes are coloured according to researchers affiliation. LORIA members are coloured in yellow. All the others institutions have their own dedicated colour.

†The equivalents in French for "Lorraine Research Laboratory in Computer Science and its Applications" `http://www.loria.fr`

Several connected components are visible. This tends to reflect the team oriented activity developed by the Lab. Looking at a single connected component or at a single research team raises several questions:

- What determines the occurrence of a collaborative link ? The link between two researchers is not a random connection phenomenon in a social network. The resulting graph components may look more "clustered" or more "repulsive" than in a purely random network.
- How cooperation relation between individuals can be characterised ? Inside a research team, people cooperate with members from the same team or from other institutes. Some of the researchers are able to maintain both types of cooperation. We call them "hubs".
- How to characterize the cooperative patterns of a research team ? The structure and the type of interactions, the presence of hubs are characteristics describing the activity of a research team.

The aim of this paper is to propose a "morpho-statistical" methodology approach for network description that will allow to answer these questions. To this end, we will rely on Markov random graph modelling, Monte Carlo simulation and Bayesian inference.

The structure of the paper is as follows. Section 2 presents the modelling of the network as a line graph, obtained by transforming the nodes of the initial graph into edges, and the previous edges into nodes. For the considered application, this allows to see the network as a graph with edges given by the researchers and the nodes given by the co-authorship link. This underscores the collaboration over the people. Networks seen as labelled graphs are complex system. They induce an extremely high number of configurations. Stochastic modelling allows to deal with this situation. The approach we propose considers an appropriate version of exponential graph models to represent collaborations initiated by a communities of researchers. The model presented in Section 2.2.1 is inspired by Potts or Ising like models. The model distribution exhibits a normalising constant that is not available in an analytic closed form. Therefore, we use Monte Carlo methods to perform statistical inference. We provide at the end of Section 2 a presentation of the simulation algorithm, the Metropolis-Hastings (MH) dynamics. Next, in Section 3, we describe the ABC Shadow algorithm (Stoica et al., 2017) used to build posterior based inference.

The remainder of the paper (Section 5) is dedicated to the practical application based on real data analysis. The case study handles the structures of scientific collaborations of research teams from the LORIA laboratory. The ABC Shadow algorithm is applied to this dataset providing the whole a posteriori distribution of the model. Thereafter, the output results are used to perform parameter estimation, statistical tests and classification procedures, in order to analyse and characterize the collaboration patterns within this institution.

Finally, in Section 6, conclusions and perspectives are depicted. Source code, notebooks and instructions used for this paper are provided in a GIT repository `https://github.com/quentinl-c/ABCShadow_article_assets`.

2. Modelling social networks

Graphs have been used to model social network since the beginning of sociology (Scott, 1988). We propose to understand intra and inter relation between organisations based on participant collaboration network. From this collaborative graphs, we will associate a more relevant one, considering the relation as the object of primary interested taking into account that collaborative links can be internal of external. This is what we explain in the next subsection.

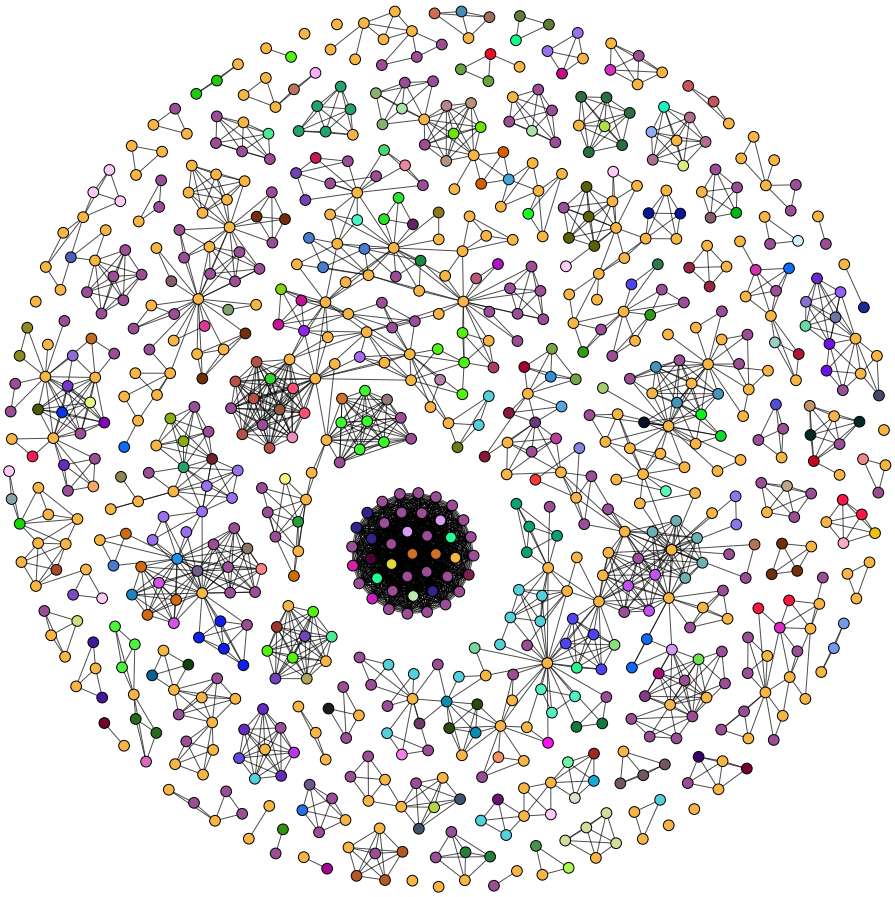


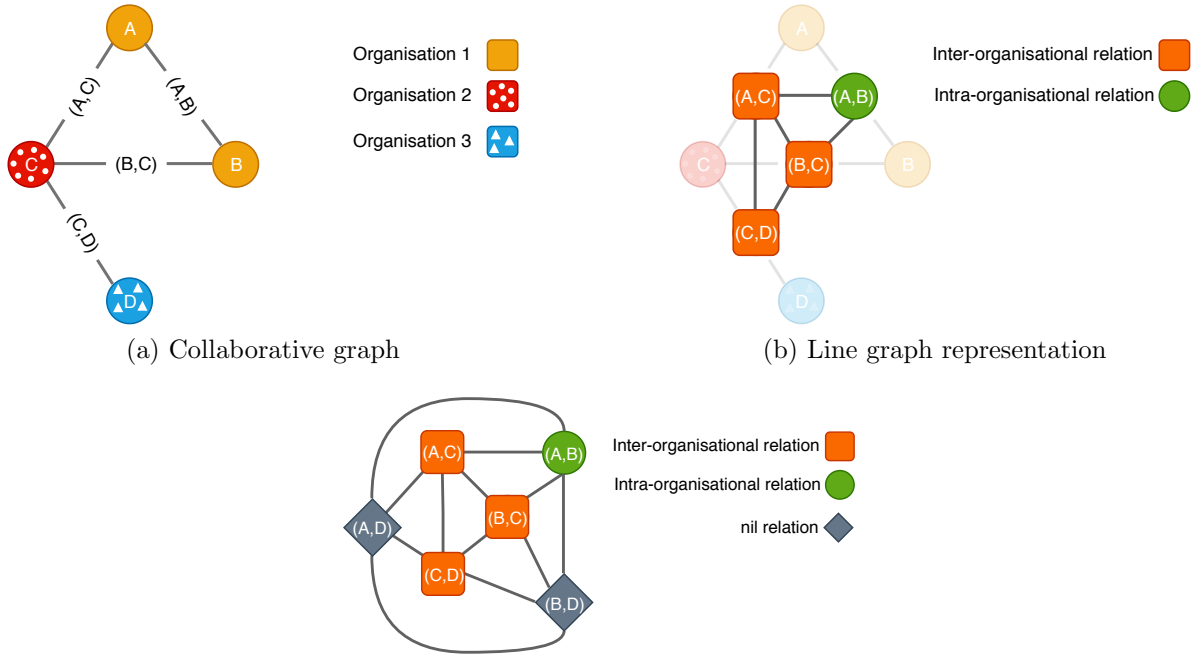
Fig. 1. Collaborations among researchers within the Loria laboratory (2018) – Each node represents a researcher, the edges are collaborative links and nodes' color represent the affiliation to a laboratory. For example, all Loria members are coloured in yellow, while the members of the other labs are differently coloured.

2.1. Network representation through line graphs

Usually, social structure studies are conducted on graphs whose vertices are individuals and links represent social ties, as in Figure 1. Here we use a representation relying on the dual graph of the network, the so-called *line graph*. This graph is obtained from the initial graph by transforming edges into nodes, and nodes into edges, as in (Frank and Strauss, 1986). This principle is illustrated in Figures 2a and 2b. The first example in Figure 2a shows the dual transformation of a graph towards its dual. The second example shows that the dual graph is not necessarily a complete graph. The line graph provides a representation of the network that emphasizes relationships over people and allows to reason on these relationships and the structure they propose.

Throughout this paper, we evaluate to which extend inter and intra organisational links occur. In the example presented in Figure 2a, A and B represent researchers working in the organisation of interest -in our case LORIA- while C and D are researchers working at other institutes. The augmented line graph in Figure 2c describes the structure of the type of interactions as follows. The green (circle) node is an intra-organisational relation, the orange (square) ones are inter-organisational

relations, while the grey (diamond) ones represent a nil relation. This last type of relation represents two researchers potentially connected that do not work together at all.



(a) Collaborative graph

(b) Line graph representation

(c) Augmented line graph representation *taking nil edges into account*

Fig. 2. An example of collaborative graph and its line graph representation

2.2. Markov Random Fields on graphs.

The example of Figure 2c illustrates our two main questions. The first one is the morpho-statistical description of different interactions in a social network. The second one is the description of the labeling distribution of the nodes in a graph. To each vertex of the line graph is associated a label, depending on the kind of link of the corresponding edge of the social graph *nil, intra_organisational, inter_organisational*.

The uncertain and dynamic nature of individuals behaviour recommends stochastic modelling of social interactions (Rezvanian and Meybodi, 2016). Within this context we propose a random graph model whose parameters provide a meaningful description of the social network of interest. Markov Random Fields (MRFs) are maybe the mathematical framework the most used to deal with this type of problems Besag (1974, 1972). They are also known in the literature related to social networks modelling under the name of Exponential Random Graph models Wasserman and Pattison (1996); Snijders et al. (2006).

Let $\mathcal{G}$ be the considered line graph, with $\mathcal{V} = \{1, \ldots, n\}$ the vertices set, $\mathcal{E} = \{e_{ij} | i \sim j, \forall i, j \in \mathcal{V}\}$ the set of its edges and $\mathcal{L} = \{(\ell_1, \cdots, \ell_n\}$ the set of the possible labels. The structure of $\mathcal{L}$ was chosen discrete for the sake of the simplicity and for the purpose on the application on hand. Its description using more general measurable spaces is perfectly possible. Following (Besag, 1974), a random field Y

is associated to $\mathcal{G}$, via the labels in a phase space that we denote $\mathcal{L}$ that have been attached to each vertex. A realization of the random field Y is denoted by y.

The set $\mathcal{L}^{\mathcal{V}}$ of all possible label configurations is denoted Ω and called the state space.

In Section 2.2.1, general notions on MRFs applied to social network analysis are given. The related simulation and inference procedure are given in Section 2.3.2. For a tourough and rigorous presentation of MRFs we recomend and the references within Winkler (2013).

2.2.1. Markov Random Fields models and social networks analysis

The MRFs were applied for social networks analysis by Frank and Strauss (1986); Wasserman and Pattison (1996); Snijders et al. (2006). This class of models enables to take into account dependencies between vertices assuming *local* interactions associated to the graph nodes. This class of models was already considered with applications to image analysis and to random graph modelling in whole generality (Besag, 1974, 1972).

In order to specify a MRF we need a neighbourhood relation. Here, two vertices i and j are neighbours, $i \sim j$, if there is a direct edge linking them. Following (Besag, 1974), the probability function of a MRF Y is described by a Gibbs distribution of the form:

$$p(Y = y|\theta) = \frac{exp(U(y|\theta))}{\kappa(\theta)} = \frac{exp(\langle\theta \cdot t(y)\rangle)}{\kappa(\theta)}, \tag{1}$$

where:

- $\theta = [\theta_0, \ldots, \theta_n]$ is the vector of parameters
- $t(\cdot)$ is the sufficient statistics vector
- $U(\cdot|\cdot)$ is the energy function
- $\kappa(\theta)$ the normalising constant.

The difficulty with this class of model is that $\kappa(\theta)$, the normalising constant is not directly available under an analytic closed form. This requires special procedures for simulation and inference. Still, their advantage is that through local specifications they allow the modelling of complex systems.

2.2.2. A Potts-like model for characterising interactions on social networks

For the problem on hand, the aim is to characterize interactions between researchers. Let us consider the following MRF model:

$$\begin{aligned} p(Y = y|\theta) = \frac{1}{\kappa(\theta)} \exp\Big[&\theta_{11} \sum_{i\sim j} \mathbb{1}\{y_i = 1, y_j = 1\} \\ &+ \theta_{12} \sum_{i\sim j} \mathbb{1}\{y_i = 1, y_j = 2\} \\ &+ \theta_{22} \sum_{i\sim j} \mathbb{1}\{y_2 = 2, y_j = 2\}\Big]. \end{aligned} \tag{2}$$

where y is the realisation of the graph representation given by the labels $\{0, 1, 2\}$ associated to each node. They correspond respectively to *nil*, *intra-organisational* and *inter-organisational* links. The sufficient statistics vector is given by

$$\begin{aligned} t(y) &= [t_{11}(y), t_{12}(y), t_{22}(y)] \\ &= \left[\sum_{i \sim j} \mathbb{1}\{y_i = 1, y_j = 1\}, \sum_{i \sim j} \mathbb{1}\{y_i = 1, y_j = 2\}, \sum_{i \sim j} \mathbb{1}\{y_i = 2, y_j = 2\} \right]. \end{aligned} \tag{3}$$

The condition in $\mathbb{1}\{y_1 = 1, y_j = 1\}$ is verified whenever a researcher cooperates with two members of his team. It means that the statistic t_{11} indicates how the researchers interact within their own team. The condition $\mathbb{1}\{y_i = 1, y_j = 2\}$ is checked whenever a researcher cooperates with a member from his own team and a member from a different team. The statistic t_{12} indicates how the researchers exhibit a hub behaviour, since they interact with both kind of teams, their own and different ones. Finally, $\mathbb{1}\{y_i = 2, y_j = 2\}$ is checked whenever a researcher cooperates with two members not belonging to his own team. Then, the statistic t_{22} indicates how the researchers interact with other teams. To sum up, the vector $\theta = [\theta_{11}, \theta_{12}, \theta_{22}]$ controls the "weight" of the previous statistics. If $\theta_{ij} > 0$ then the model tends to favour configurations with a high value for the statistic t_{ij}.

This model colours the line graph associated to a network in a similar manner as the Potts model does it. If important patches of $(1, 1)$ appear this means that there is an important tendency that the researchers on the network cooperate within their teams. Similar interpretation can be given, for the patches $(1, 2)$ and $(2, 2)$. The weight, the importance of these patches, hence of the general behaviour of the members of the network is given by the model parameters.

2.3. Simulation and inference procedures.

In this section, we review the state-of-art of inference of random graphs, beginning with simulation in Section 2.3.1, since it is a key part of the inference process presented in Section 2.3.2.

The presence of $\kappa(\theta)$ in (2) imposes special strategies for the sampling of the model, Markov chains Monte Carlo methods. The most known sampling algorithms are the MH and the Gibbs sampler. Here, we briefly recall the MH algorithm (Hastings, 1970), due to its role within the inference procedures used through this paper.

2.3.1. MCMC simulation

The purpose is to sample distributions which are not analytically tractable such as (2). The MH algorithm provides a solution for this problem. The algorithm works by iterating a two steps procedure. The first step of the procedure is the following: being in an initial state y a new candidate y' is generated according to the proposal density $q(y \to y')$. The second one is to accept the candidate with the probability given by

$$\alpha_{y \to y'} = min \left[1, \frac{p(y'|\theta) q(y' \to y)}{p(y|\theta) q(y \to y')} \right]. \tag{4}$$

This dynamic reproduces the iteration of a transition kernel of a Markov chain with equilibrium distribution, the probability distribution one wants to simulate. Reasonable conditions are required for the proposal q to ensure convergence of the algorithm. The proposal should allow the simulated

chain to be irreducible, recurrent and ergodic. In our situation, simple choices for the proposal, such as a uniform distribution over the set of labels, guarantee all the needed convergence properties. Furthermore, the computation of (4) does not require the knowledge of the normalising constants $\kappa(\theta)$. Still, the price to pay for this naive choice is an important correlation of the samples and a high level of rejection of the proposed samples. Some strategies to overcome these drawbacks will be mentioned within the application and conclusion parts of the paper.

EXPLIQUER PLUS TARD : GIBBS OK pour simuler des graphs, mais attention ABC Shadow est construit sur MH car pas de dependence locale entre les parametres ...
– Radu

2.3.2. Inference procedures

Parameter estimation of MRFs (2) is not trivial due to the normalising constant:

$$\kappa(\theta) = \sum_{y \in \Omega} exp(\langle \theta \cdot t(y) \rangle).$$

where $\cdot$ represents the scalar product between the parameters and sufficient vectors, respectively.

The classical way of dealing with this problem is to use Monte Carlo Maximum Likelihood estimation (Geyer and Thompson, 1992; Geyer, 1999; Handcock et al., 2003). Let y_{obs} be an observed graph and let us consider θ_0 a given parameter value. The log-likelihood function can be written as:

$$l_{\theta_0}(\theta) = \langle (\theta - \theta_0) \cdot t(y_{obs}) \rangle - log \left[\frac{\kappa(\theta)}{\kappa(\theta_0)} \right]. \tag{5}$$

It can be shown that the ratio of the normalising constants is

$$\frac{\kappa(\theta)}{\kappa(\theta_0)} = \mathbb{E}_{\theta_0} exp(\langle (\theta - \theta_0) \cdot t(Y) \rangle) \tag{6}$$

which give for its Monte Carlo counterpart :

$$\frac{\kappa(\theta)}{\kappa(\theta_0)} \approx \frac{1}{n} \sum_{i=0}^{n-1} exp(\langle (\theta - \theta_0) \cdot t(y_i) \rangle) \tag{7}$$

where the $\{y_i\}_{0 \leq i < n}$ are realizations of $\{Y_i\}_{0 \leq i < n}$ i.i.d. sampled from $p(y|\theta_0)$.

If the sampling algorithm satisfies convenient assumptions, an almost sure convergence result allows the practical use of this approximation. In fact (7) is plugged into (5) and the Monte Carlo likelihood is obtained

$$l_{n,\theta_0}(\theta) = \langle (\theta - \theta_0) \cdot t(y_{obs}) \rangle - log \left[\frac{1}{n} \sum_{i=0}^{n-1} exp(\langle (\theta - \theta_0) \cdot t(y_i) \rangle) \right]. \tag{8}$$

For the exponential family models, the log-likelihood is concave Geyer (1999); Monfort (1997). This motivates to compute the gradient and the Hessian of (8). The approximated gradient and Hessian can be easily computed via importance sampling. These quantities are consistent estimators of their exact counterparts, respectively, that are computed from the original log-likelihood. Finally, using these quantities a Monte Carlo Newton Raphson (MCNR) local optimisation method can be implemented.

This method exhibits convergence results and two asymptotics explaining the estimation error can be computed. The first error is the Monte Carlo Standard Error that approximates the difference between the true model parameters and the Maximum Likelihood Estimate, that are both unknown. The second error is the Monte Carlo Maximum Likelihood Error that approximates the difference between the Maximum Likelihood Estimate (which is unknown) and the Monte Carlo Maximum Likelihood Estimate, the result given by the MCNR method.

The drawback of the MCNR method is that it requires θ_0 to be close to the final estimate. This is due to the fact that the computation of the importance sampling weights needed in the evaluation of the gradient and the Hessian are not stable from a numerical point of view. Several strategies are available. Among them, the most robust is to resample the model $p(y|\theta)$ whenever the difference between the current value of the parameters and θ_0 exceeds a given threshold. Due to the concavity of the log-likelihood function, this strategy leads towards a convergent method but with a high computational cost. This question is still an open problem (Geyer and Thompson, 1992; Geyer, 1994, 1999).

3. Posterior based inference

According to the Bayes' theorem, with $p(\theta)$ the prior knowledge on parameter distribution, the posterior distribution $p(\theta|y)$ is:

$$p(\theta|y) \propto p(y|\theta) \cdot p(\theta) = \frac{\exp\langle t(y), \theta\rangle}{\kappa(\theta)} p(\theta). \tag{9}$$

The difference between maximum likelihood that we described in Section 2.3.2 and posterior inference is the following. In the first case, under the assumption of a parametric model and with no prior knowledge regarding these parameters, the most probable model is proposed as an explanation of the observed data. The posterior based inference assumes also a parametric model and it uses prior knowledge with respect to these parameters. But, each model belonging to the family may explain the data. The quality of this explanation is given by the posterior distribution that weights each model within the considered parametric family. Posterior based inference is much more informative. It can be seen also as a generalisation of the maximum likelihood approach. Whenever $p(\theta)$ is the uniform distribution of the parameter space Θ, both inference paradigm, posterior and likelihood, are strictly equivalent.

Despite the interest in performing posterior based inference, this is not done often, since sampling the posterior or the likelihood is far from being a trivial task. A straightforward application of Monte Carlo sampling strategies such as MH or Gibbs dynamics requires the computation of the normalising constants ratio (6).

The authors in Møller et al. (2006) give a very elegant solution to this problem. They propose a MH dynamics based on auxiliary variables. The use of the auxiliary variables requires appropriate proposal distributions. The proposal distributions can be tailored to cancel the computation of the normalising constants within the acceptance ratio of the MH algorithm. The authors indicate themselves that their rigorous mathematical solution cannot prevent the resulting chain from poor mixing.

Approximate Bayesian Computation (ABC) algorithms are methods used to approximately sample from the posterior distributions of the models that cannot be expressed entirely under analytic closed form. They are easy to implement, but they require adapted strategies in order to obtain samples whose distribution is close from the posterior distribution. The ABC methods then need to control the distance between the observations and the output of the algorithm (Atchadé et al., 2013; Beaumont et al., 2009; Grelaud et al., 2009; Marin et al., 2012).

The ABC Shadow method proposed by Stoica et al. (2017) is directly inspired from the previous two ideas, while trying to solve some of their drawbacks. The ABC Shadow is an approximate sampling method for posterior distribution, exhibiting better numerical properties than the auxiliary variable method and offering a more robust control than the ABC classical framework. Recent work Stoica et al. (2019) builds a convergent simulated annealing process based on a ABC Shadow dynamics.

The ABC Shadow algorithm is presented in Algorithm 1. For all the technical details and mathematical proofs the reader has to refer to Stoica et al. (2017). The method is general in the sense that it can be applied to sample posterior distributions, assuming only their continuous differentiability with respect to the model parameters. The algorithm needs for initialisation the observed graph y_{obs}, the initial value θ_0 of θ, Δ an error control parameter and m the number of steps the algorithm runs. The Δ parameter supports the proposal distribution whose form is given line 4 of Algorithm 1. All theoretical details about the construction of the proposal can be found in Stoica et al. (2017). First the algorithm samples an auxiliary graph x according to the chosen model. Then for each step in the loop it proposes a new parameter value θ' that is accepted with the probability α (see line 7 of Algorithm 1). If this new state is not accepted, the algorithm remains in its previous state. The distribution of the output of the algorithm follows approximately $p(\theta|y_obs)$ with an error limits controlled by m and Δ. The value of Δ has to be tuned in a fine way, since there an acceptable compromise to reach between quality of approximation and good mixing properties of the chain. If the number of steps m is too large, the algorithm goes away from the posterior of interest whereas if m is too small the mixing property is negatively impacted. Hence, a reasonable value for these two parameters is needed. In Stoica et al. (2017) is proved that for any fixed value m there exists a positive value Δ so that the outputs of the ABC Shadow algorithm are distributed as close as desired from the posterior distribution of interest. If more than one sample from the posterior is needed, this can be obtained by iterating the ABC Shadow algorithm as described in Algorithm 2.

Algorithm 1 ABC Shadow algorithm

1: **function** ABC_SHADOW(θ_0, y_{obs}, m, Δ) ▷ Where θ_0 - initial parameters, y_{obs} - observation
2: $x \sim p(x|\theta_0)$
3: **for** i = 1 to m **do**
4: $\theta' \sim \mathcal{U}_\Delta(\theta_{i-1} \to \theta')$
5: $\alpha \leftarrow min\left\{1, \exp[(t(y_{obs}) - t(x))(\theta' - \theta_{i-1})]\frac{p(\theta')}{p(\theta_{i-1})}\right\}$
6: $accepted \leftarrow \mathcal{U}(0,1)$
7: **if** $\alpha > accepted$ **then**
8: $\theta_i \leftarrow \theta'$
9: **else**
10: $\theta_i \leftarrow \theta_{i-1}$
11: **end if**
12: **end for**
13: Return θ_m
14: **end function**

Algorithm 2 Main Routine

function MAIN(θ_{prior}, y_{obs}, m, Δ, $iters$) ▷ Where $iters$ is the number of samples
  $samples \leftarrow [\theta_{prior}]$
  for $\_ \in [0 \dots iters - 1]$ **do**
    $\theta_{last} \leftarrow samples.last()$
    $\theta_{res} \leftarrow ABC\_SHADOW(\theta_{last}, y_{obs}, m, \Delta)$
    $samples.append(\theta_{res})$
  end for
  Return $samples$
end function

4. ABC shadow in practice : illustration on synthetic data

The use of ABC Shadow algorithm requires the set-up of its parameters. Regarding the auxiliary variable sampling, this is perfectly possible using exact simulation methods Huber (2016). Here, for numerical purposes and due also to the rather weak requirements regarding the auxiliary variable, a Metropolis-Hastings dynamics whose parameters are given below was chosen. In order to chose m and Δ the ABC Shadow algorithm was run on known models, with controllable expected results. Whenever it was possible, the outputs of the ABC Shadow algorithm were compared with a classical Monte Carlo sampler of the posterior, the MH algorithm.

4.1. Binomial distribution

Let y be generated by a Binomial distribution of parameters n and p. This may correspond to the independent random labelling, following a Bernoulli distribution with parameter p, of a bi-coloured graph of size of n. We know the parameter n and we want to estimate p. Within this context the likelihood reads :

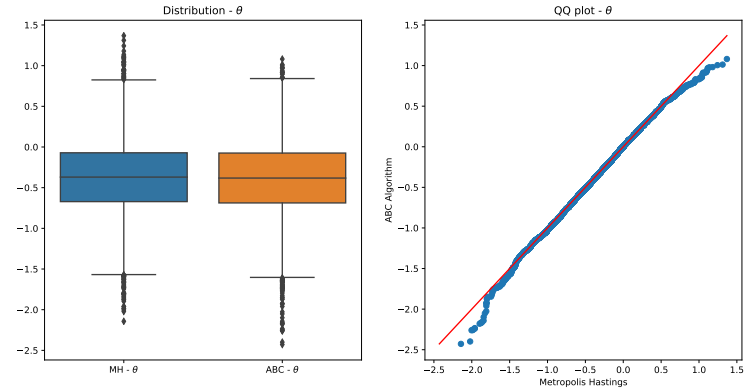
$$p(y|\theta) = \binom{n}{y} p^y (1-p)^{(N-y)} = exp\left[y\theta - n\log(1+e^\theta) + log\binom{n}{y}\right] \tag{10}$$

with $\theta = log(p/(1-p))$. For our experiment $n = 20$, $p = 0.4$ ($\theta = -0.405$) and $m = 100$. The observed Binomial variable obtained with these values was $y = 8$. The MH algorithm is set up to sample from the distribution (10). The proposal distribution $p(\theta)$ is uniform over the interval $[-100, 100]$ of width $\Delta = 0.005$ centred on the current value. This procedure was executed to sample 10^6 posteriors. The first 2×10^3 samples was cut off and a subsampling kept every 100 samples. This resulted in a chain $(\theta^{(t)})_{t=1,\cdots,T}$ of 9800 samples.

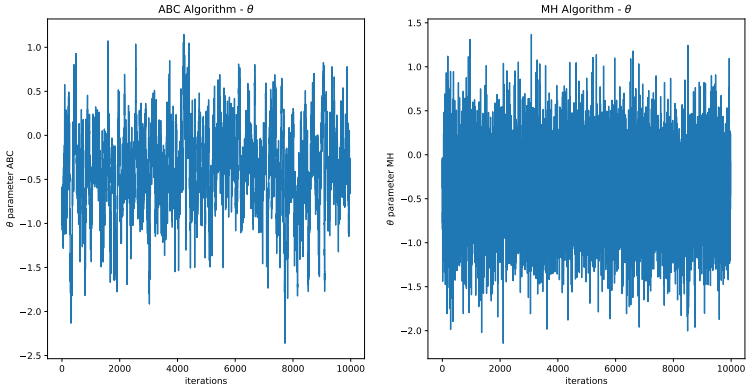
For the ABC Shadow, the proposal distribution is the same as the one of the MH algorithm. The auxiliary variable is simulated from 100 samples following (10). The procedure described in Algorithm 2, is implemented and applied to our simulated data with $m = 100$ and $iters = 10^6$. Likely to the MH, the output of Algorithm 2 is a chain $(\theta^{(t)})_{t=1,\cdots,T}$ that we subsample, keeping only every 100 values. It improves the mixing properties of the chain. In addition, we skipped the first 2×10^3 samples of the chain $(\theta^{(t)})_{t=1,\cdots,T}$. To illustrate the robustness of these two algorithms, the initial value of the chain of samples $\theta^{(0)}$ is chosen far from the true value of θ. We set $\theta^{(0)} = 1$.

Figure 3a represents the distributions respectively obtained with the MH algorithm which is a perfect simulation algorithm and the ABC algorithm which is an approximated one. According to the

box plot and the quantile-quantile plot schema, both distributions are very close to each other showing how accurate the approximated ABC algorithm is. It is worth to notice that the two algorithms (especially ABC) converge toward the true parameter value $\theta = -0.405$, although the initial value of the chain ($\theta^{(0)} = 1$) is quite far from the truth. Statistics, Maximum A Posteriori (MAP) and errors of both distributions are summarized in Table 1. We emphasize that both methods provide not only an estimation of the parameter of the model but the whole a posteriori distribution, yielding notably error metrics and confidence bounds for θ.



(a) Posterior distribution



(b) Posterior trace

Fig. 3. Posterior sampling of a Bernoulli distribution using Metropolis Hasting and ABC Shadow

4.2. Posterior sampling on the Potts Model

We now consider the Potts model involved in the description of our application context. Due to the normalising constant, the Potts model (described in Section 2.2.2) is not directly tractable with the traditional MH algorithm as previously performed in Section 4. To circumvent this problem and

Table 1. Statistics on the posterior of Bernoulli distribution

	Q_{10}	Q_{25}	Q_{50}	$mean$	Q_{75}	Q_{95}	MAP	$\hat{\sigma}_\theta$	$\hat{\sigma}_\theta^{MC}$
ABC (θ)	-0.986	-0.688	-0.382	-0.39	-0.074	0.346	-0.408	0.212	4.5×10^{-4}
MH (θ)	-0.961	-0.672	-0.37	-0.377	-0.071	0.353	-0.3718	0.211	4.4×10^{-4}

following the strategy in Stoica et al. (2017), we tested the Potts model by comparing the maximum of the approximated a posteriori distribution with the true parameter of the model previously simulated.

In a first experiments, all interaction parameters are fixed to 0 : $\theta_{11} = \theta_{12} = \theta_{22} = 0$, so that interaction effects are annihilated. Since we have three type of patterns, this leads to a Bernoulli graph model with a occurrence probability for each pattern equal to $\frac{1}{3}$. The observation was generated from $N = 10^3$ samples with a $size = 20$ using a MH sampler. By averaging sufficient statistics we obtain $\bar{t}(y) = [764.54, 759.80, 755.75]$ (see (3)) from the ABC Algorithm. In the ABC algorithm, the prior distribution $p(\theta)$ was a uniform distribution on the interval $[-4, 4] \times [-4, 4] \times [-4, 4]$. The parameters n and Δ were respectively set to $n = 200$ and $\Delta = [0.01, 0.01, 0.01]$. As in Section 4.1, the ABC Shadow was executed to yield $iters = 10^6$ samples. We subsample keeping every 10^3 value. At each iteration the auxiliary variable x was updated using 200 steps of a MH sampler.

Figure 4a represents the histograms of the posterior distributions provided by the ABC Shadow of each parameter as well as two-dimensional posterior distributions for each couple of parameters. Blue lines mark the MAP for each parameter's distribution computed by taking the maximum of the kernel density estimation: $\hat{\theta} = [-0.0262, 0.036, -0.0436]$.

We now consider a model with repulsion effects. To that end, we set $\theta_{11} = -0.5$, $\theta_{12} = 0.2$, $\theta_{22} = 0.3$ and we simulate $N = 10^3$ samples with a $size = 20$ using a MH sampler as we did before. The generated observation yielded the following averaged sufficient statistics: $\bar{t}(y) = [572.54, 1190.15, 509.66]$. Figure 4b represents the resulting posterior distribution. Blue lines representing the MAP are aligned on $\hat{\theta} = [-0.5306, 0.2189, 0.3004]$ which is close to the true parameter $\theta = [1.5, 2, 1]$ represented with the green lines. The dashed lines are respectively the first quartile, the median and the third quartile. The mean and the median of the posterior estimates are respectively: $[-0.702, 0.294, 0.263]$ and $[-0.64, 0.268, 0.277]$.

dtype: float64

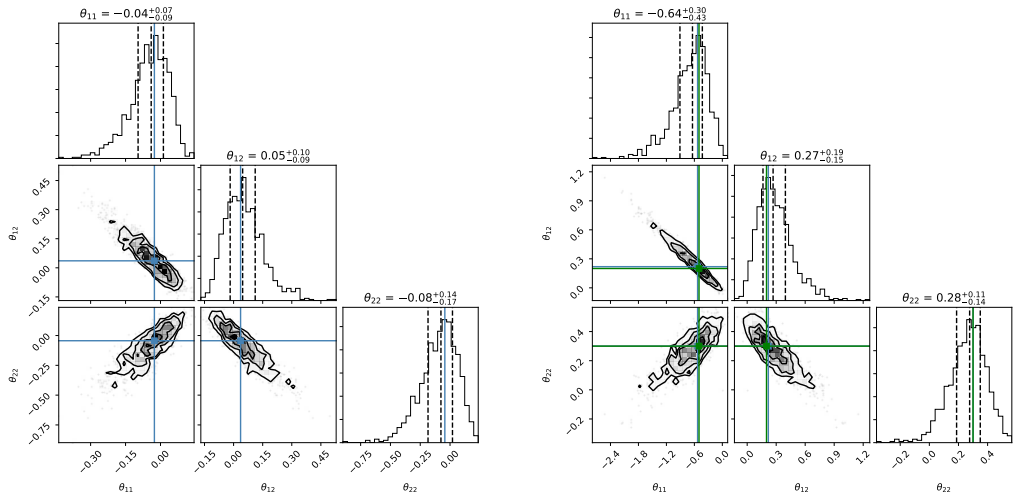
As a result, we showed that ABC Shadow approximate accurately a posterior distribution of untractable model as the one we plan to study. We can now confidently go further and apply it to real data.

5. Application

A collaboration network is obtained using the HAL publication database. A node represents a researcher. Two researchers are connected if they have at least a common publication during the year 2018. We collected metadata of publications deposited by the members of LORIA in 2018. The dataset is available at (Laporte-Chabasse et al., 2019).

The aim of the study is to fit the model defined in Section 2.2.2 to the graph associated with each team. Comparing the structural aspects of those graphs through posterior analysis enables the identification of patterns characteristic of these scientific collaborations.

For each team, the graph is constructed as follows. Figure 5 exhibits the different steps of the processing. First, two kinds of nodes were distinguished, the members of LORIA and the other



(a) (b)

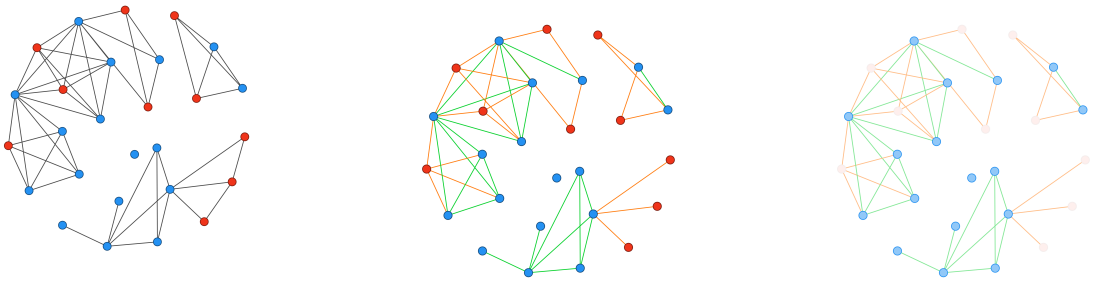
Fig. 4. Corner plots of marginal distributions of posterior sampling for the Potts model using an ABC algorithm. *(Blue lines mark the MAP of each parameter, green lines correspond to the true parameter values)*

researchers who had no affiliation with LORIA, the external stakeholders (Figure 5a). We took the point of view of each team and studied the way they collaborate with internal and external stakeholders. This means that only edges linking at least one member of LORIA are considered (Figure 5b). In addition, we only took into account interactions between edges linked by a member of LORIA (Figure 5c). Following the framework of Section 2.1, the line graph representation encodes the different types of research collaboration. An inter-organisational link connects one member of LORIA with an external collaborators whereas intra-organisational links connect two collaborators who are affiliated to LORIA. Under the hypothesis of the model, the sufficient statistics were computed. The obtained values are presented in Table 2.

The number of authors from LORIA as well as external stakeholders are different according to each team. Statistics of both quantities are given in Table 3. Compared to the means, the standard deviations are important. This indicates that the sizes of the different collaborations are distributed on an important range. It is important to bear in mind that the number of referenced authors on HAL is neither a representative picture of the actual size of the team nor a quantifier of the research activity.

We identified 22 teams with a sufficiently large number of submissions on the HAL platform. The ABC Shadow algorithm was launched with the same initial conditions for every team. The ABC Shadow algorithm was setup to generate $iters = 10^6$ samples, the number of iterations of the shadow chain and the volume bound were set respectively to $m = 200$ and $\Delta = [0.01, 0.01, 0.01]$. The auxiliary variable x was sampled with 300 iterations of the MH procedure. A subsampling procedure kept every 10^3 value of each chain yielded by the ABC Shadow. Consequently, for each team the size of the corresponding chain was 1000 samples.

Figure 6 shows the box plots of posteriors sampled by the ABC Shadow for the parameters $\theta = [\theta_{11}, \theta_{12}, \theta_{22}]$ for each team. In complement to Figure 6, Tables 4, 5 in Appendix A present the mean,



(a) *step 1:* Distinguish two kinds of nodes, members of LORIA (in blue) and external stakeholders (in red).

(b) *step 2:* Keep edges involving at least one member of LORIA.

(c) *step 3:* Consider only interactions between edges linked by a member of LORIA.

Fig. 5. Example of a pre-processing performed on a team's collaborative graph. The graph in Figure 5a illustrated co-authoring links involving the members of the team COAST. Blue nodes represents the members of LORIA, whereas red nodes are external collaborators. The inter-organisational links are coloured in orange, while the intra-organisational links are in green.

the median and the estimated MAP of the posterior distribution distribution of θ for each team. At first sight, we note that the value ranges of parameters are near to zero, even slightly lower than zero for the majority of teams. Relatively to all possible connections, this reflects a weak global tendency for a researcher to co-author with all other researchers whether he belongs to the same lab or not. This means, at the scale of teams that the collaboration graph is sparse. Putting this observation in the context of publication activities, this corroborates the intuition that every researcher does not co-author with everyone else. Co-authoring a paper implies that all stakeholders are involved in the same scientific work. These are demanding tasks which consequently restrict the number of publications and the underlying potential collaborations a researcher is able to undertake.

Regarding the tables 4 and 5 in Appendix A, note that both the median and the mean are close to the estimated MAP. Similarly to Van Lieshout and Stoica (2003); Stoica et al. (2017), the asymptotic standard deviation and the Monte Carlo standard deviation were computed. The results are given by Table 6 (in Appendix A). To that end, for each estimated model a simulation providing 10,000 samples was performed. Given the Monte Carlo standard deviation, we can determinate the 95% confidence interval reported by Table 7 (in Appendix A).

The closeness of values ranges to 0 raises the question of their significance. In other words, do the three studied patterns are more or less likely to occur in the collaboration than pure randomness ? To answer this question we applied for each parameter a t-test to determine if the expectations of the posteriors equal 0. The null hypothesis and the alternative hypothesis are written as follows for each parameter:

$$\mathcal{H}_0 : \mathbb{E}[\theta] = 0,$$
$$\mathcal{H}_1 : \mathbb{E}[\theta] \neq 0.$$

The results are shown in Table 8. For most of the teams, the parameters are significantly non zero since the associated p-values of the t-test are very small. There is only one team for which the p-value is greater than the usual 5% level of significance. In this case, the rejection of the null hypothesis is not relevant. As a conclusion, for almost all teams (except the latter mentioned), the likelihood of

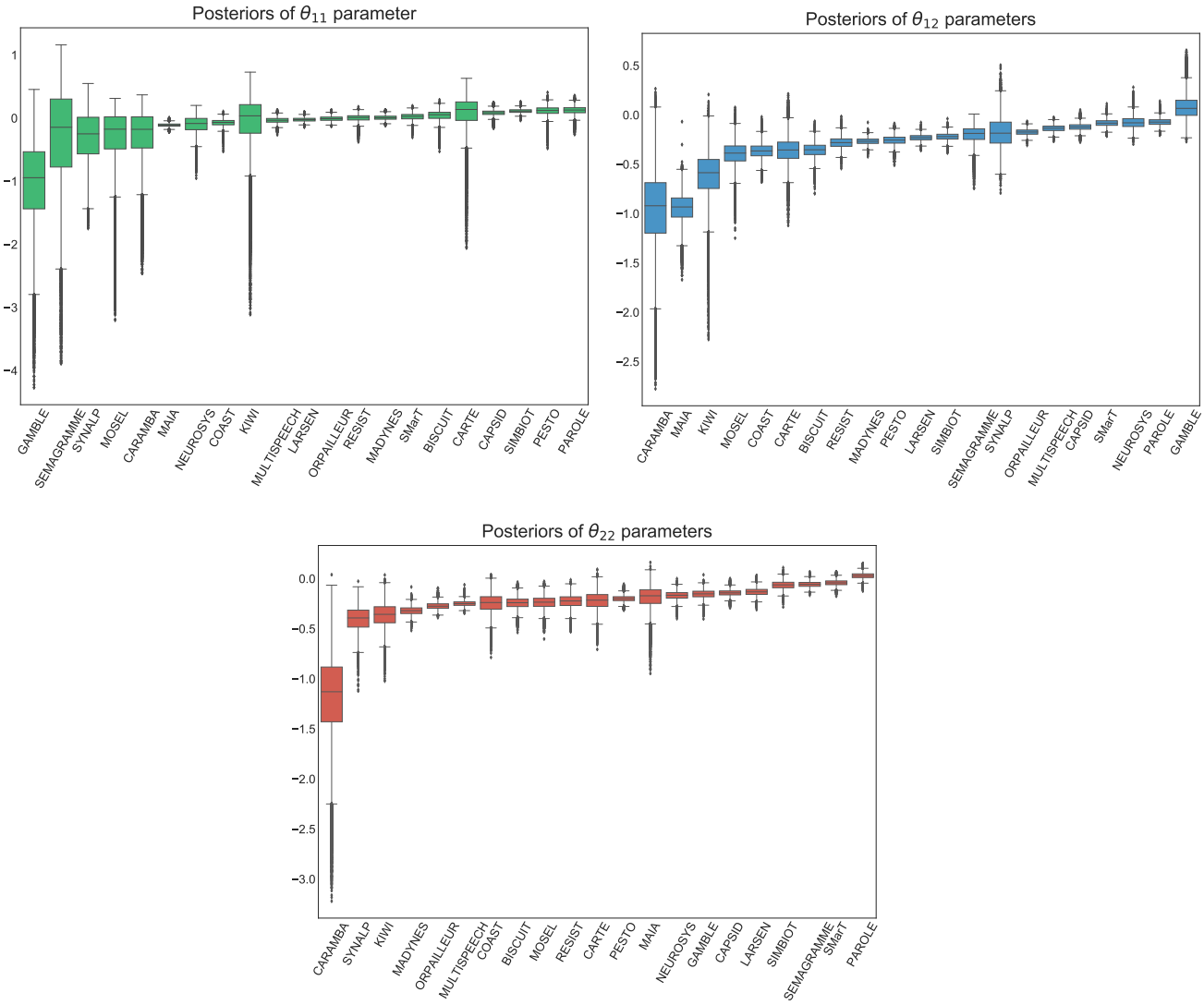


Fig. 6. Box plots of the posterior distributions for the parameters of the Potts model estimated from the collaboration graphs of each team (ordered by ascending mean values)

link creation is not merely due to the chance.

Figure 7 presents the three 2d projections of the Potts model parameters. Each team is associated with a colour. Grey dashed lines set limits between positive and negative trends. For instance, considering (θ_{11}, θ_{12}), the vertical line delimits the positive and negative tendencies that a pattern linking two intra-organisational ties occurs, while the horizontal line is about the occurrence of inter-organisational links. Depending on projections we have an overview of trends followed by teams.

The major part of the estimated MAPs is concentrated in the same region. For the parameter θ_{11} the MAPs are distributed closely around 0. For θ_{12} and θ_{22}, they are mostly negative. This observation refines our analysis. The latter two, show that hub patterns and collaborative links with the outside

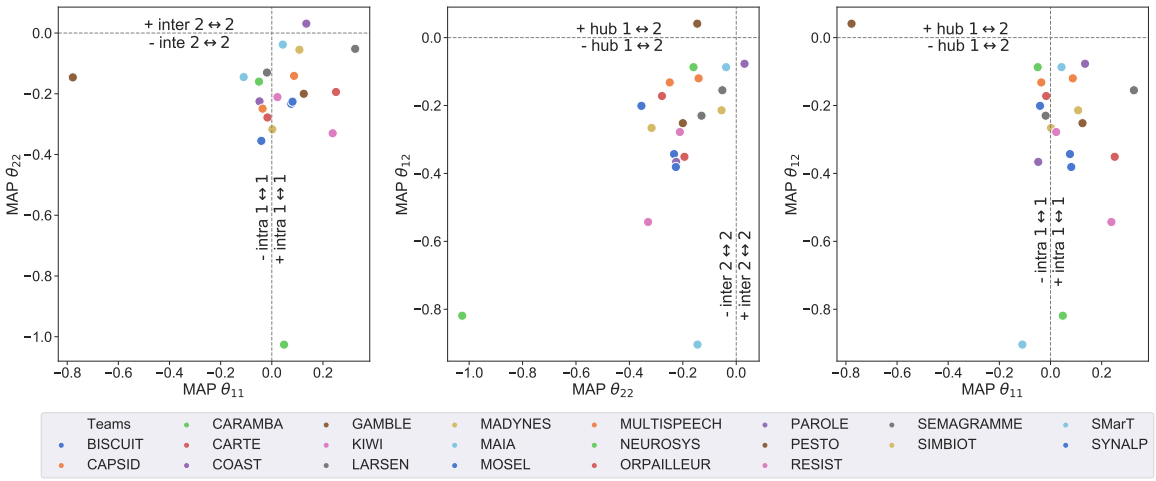


Fig. 7. Scatter plots of estimated MAPs representing the positioning of teams with respect to the different collaborative tendencies controlled by θ_{01}, θ_{02}, θ_{12}

are less likely to occur in collaboration graphs. This strengthens the prior idea that collaborations with external teams are complex to set up and maintain. The weak presence of hubs in the collaboration means that only few researchers are connected at the same time with members of their team and researchers from other labs. If a hub leaves, the ties between the corresponding organisations break. This is a serious concern that should be carefully addressed in the design of collaborative applications to ensure the availability of the collaboration against the churn.

Some outliers presenting different structural features are identifiable. In particular, two points, the purple point (GAMBLE), located at the top right-hand side of the first plot and the brown point (PAROLE) located above the horizontal dashed line of the two later plots. Both teams represented by these points, exhibits external collaboration structural patterns. Regarding the team GAMBLE, it is noticeable that the number of external researchers is high compared to the number of referenced authors from LORIA (Table 2). This fosters the emergence of inter-organisational links to the detriment of intra-organisational links. The team PAROLE, on the other hand, shows a prevalence of hub patterns. Figure 8 illustrates its co-authorship graph. According to Table 2, the ratio between external and internal authors is more balanced. In Figure 8, some actors from LORIA playing a key role are well marked, they are represented by nodes in a higher size. In that team, a major part of LORIA researchers acts as a bridge between their counterparts and external stakeholders. This is a special configuration no met in other co-authorship graphs.

Figure 7 shows some overlapping points or points very close to each other. This suggests that some teams share with each other similar structural characteristics. By relying not only on the MAPs but on the whole posterior distributions, we aim to verify these observations.

An unsupervised hierarchical classification was performed, from the Kolmogorov-Smirnov distance computed between all posterior distributions of the three parameters: θ_{11}, θ_{12} and θ_{22}. The results in figure 9 are shown in the form of dendrograms. The branches' height of the dendrogram gives indications about the proximity of the sub-clusters : shallower is a branch, closer are the sub-clusters and vice-versa. The few identified clusters correspond to the coloured branches. They merely correspond to groups of overlapped points in Figure 7. It is worth noting that the clustered teams don't

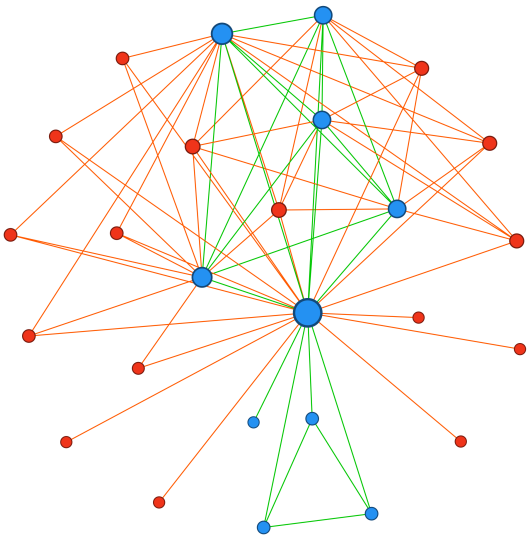


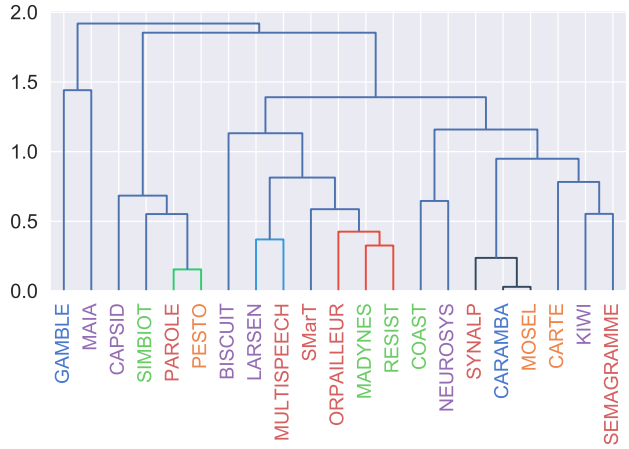
Fig. 8. Co-authorship graph of the team PAROLE. Blue nodes represents the members of LORIA, whereas red nodes are external collaborators. Size of nodes is proportional to their degree. The inter-organisational links are coloured in orange, while the intra-organisational links are in green.

necessarily work on the same topic. Consequently, studied structural patterns not a feature specific to the research thematic. However we also noticed that the closeness between two teams can be related to the fact that one originates from the other one. In practice, it is not unusual that a researcher keep signing with the old affiliation a long time after the creation of the new team. Besides, when a team split in two new teams, members of both teams keep working together. This means that both teams keep intrinsic collaboration links affecting their collaboration networks. This requires to pay particular attention to the real-life context in particular to the teams' life cycle : birth, split, death. More broadly, the classification throughout the Kolmogorov-Smirnov distance of posteriors provides information about shared tendencies that is not possible to infer by only observing the posterior distributions independently or comparing the MAPs.

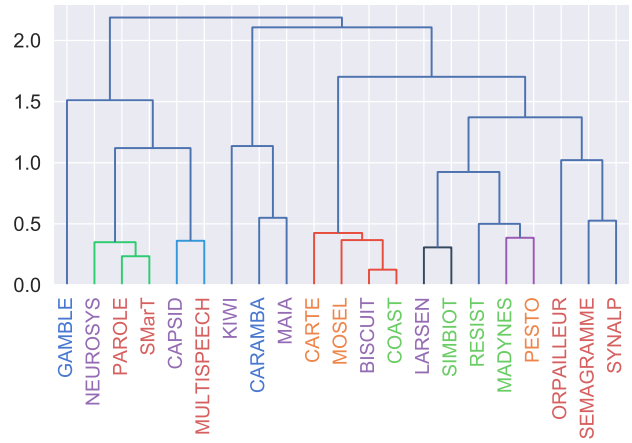
6. Conclusion

In this paper, we proposed a method to make inference on structural aspects of collaboration networks. We focused on the inter-organisational collaborations yet sparsely addressed by the state of the art. An illustration of such kind of collaboration is the one initiated by researchers to write publications. We drew the collaboration network among researchers by considering the co-authorship of publications deposited on the French open-archive HAL as collaboration links between authors.

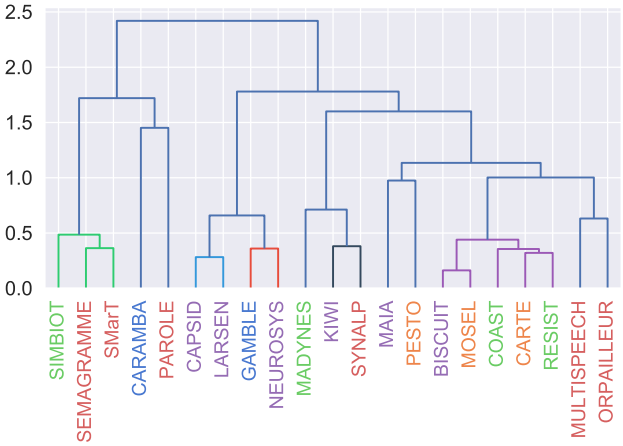
The first part of our approach was on the representation of the collaboration graph. Instead of taking the collaboration network directly as the observation of our study, we relied on the line graph. Considering this alternate representation as fixed random field, we considered link creation in the collaboration as a labelling issue respecting Markov's properties. In that way, we were able to better encompass structural interactions not between individuals but among relations themselves. We used a



(a) According to the parameter θ_{11}



(b) According to the parameter θ_{12}



(c) According to the parameter θ_{22}

Fig. 9. Hierarchical classification throughout the Kolmogorov-Smirnov distance of posteriors. The label are coloured according to the research thematic addressed by each team : Algorithms, Computation, Image & Geometry, Formal methods, Networks, Systems and Services, Natural Language Processing & Knowledge Discovery, Complex Systems, Artificial Intelligence and Robotics.

generalisation of the Ising model, the Potts model, to describe the interactions between relations. As for all exponential models, the inference remains difficult due to the intractable normalising constant. To that end, we used a Bayesian tool, the ABC Shadow algorithm which was firstly tuned on tractable model and simulated data. It was applied on collaboration networks of different observed research teams. The main aim was to characterise and classify collaborations among researchers in their publication activities. First of all, we observed that links formation between collaborators are not mere coincidence but the result of tendencies for almost all teams. From the posterior distributions provided by the ABC Shadow we showed that some actors play a key role since they connect collaborators of

their organisation toward the outside. Given the posteriors, we also demonstrate how to classify the way the different teams collaborate.

Hub in the collaboration are points of failure who can endanger the inter-organisational collaboration if they leave. This is a concern that must be addressed in the design of collaborative applications, to better support inter-organisational scenarios. Study of the dynamic of the network (Guyon and Hardouin, 2002) enables the assessment of the churn and the detection of breaks over time.

Finally, the complete approach we proposed here can be applied in different contexts and is not only related to collaborations between researchers. Extending this study to other collaborative contexts is required to acquire a comprehensive understanding of features inherent to inter-organisational collaborations.

References

Atchadé, Y. F., Lartillot, N. and Robert, C. (2013) Bayesian computation for statistical models with intractable normalizing constants. *Brazilian Journal of Probability and Statistics*, **27**, 416–436. URL: `https://projecteuclid.org/euclid.bjps/1378729981`.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. and Robert, C. P. (2009) Adaptive approximate Bayesian computation. *Biometrika*, **96**, 983–990. URL: `https://academic.oup.com/biomet/article/96/4/983/220502`.

Besag, J. (1974) Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**, 192–236. URL: `https://www.jstor.org/stable/2984812`.

Besag, J. E. (1972) Nearest-Neighbour Systems and the Auto-Logistic Model for Binary Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 75–83. URL: `https://www.jstor.org/stable/2985051`.

Bollobás, B. (2013) *Modern graph theory*, vol. 184. Springer Science & Business Media.

Frank, O. and Strauss, D. (1986) Markov graphs. *Journal of the american Statistical association*, **81**, 832–842.

Geyer, C. J. (1994) On the convergence of monte carlo maximum likelihood calculations. *Journal of Royal Statistical Society, Series B*, **54**, 261–274.

— (1999) *Likelihood inference for spatial point processes. In 0. E. Barndorff-Nielsen, WS Kendall, and MNM van Lieshout, editors, Stochas-tic Geometry: Likelihood and Computation*. Chapman and Hall.

Geyer, C. J. and Thompson, E. A. (1992) Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 657–699.

Glänzel, W. (2001) National characteristics in international scientific co-authorship relations. *Scientometrics*, **51**, 69–115. URL: `https://doi.org/10.1023/A:1010512628145`.

Grelaud, A., Robert, C. P., Marin, J.-M., Rodolphe, F. and Taly, J.-F. (2009) ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, **4**, 317–335. URL: `https://projecteuclid.org/euclid.ba/1340370280`.

Guyon, X. and Hardouin, C. (2002) Markov chain markov field dynamics: Models and statistics. *Statistics: A Journal of Theoretical and Applied Statistics*, **36**, 339–363.

Handcock, M. S., Robins, G., Snijders, T., Moody, J. and Besag, J. (2003) Assessing degeneracy in statistical models of social networks. *Journal of the American Statistical Association*, **76**, 33–50.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109. URL: https://doi.org/10.1093/biomet/57.1.97.

Huber, M. L. (2016) *Perfect Simulation*. CRC Press. Google-Books-ID: xD5qCwAAQBAJ.

Laporte-Chabasse, Q., Stoica, R. S., Clausel, M., Charoy, F. and Oster, G. (2019) Co-authoring graphs of research teams in a laboratory in computer science. URL: https://doi.org/10.5281/zenodo.3570831.

Marin, J.-M., Pudlo, P., Robert, C. P. and Ryder, R. J. (2012) Approximate Bayesian computational methods. *Statistics and Computing*, **22**, 1167–1180. URL: https://doi.org/10.1007/s11222-011-9288-2.

Møller, J., Pettitt, A. N., Reeves, R. and Berthelsen, K. K. (2006) An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, **93**, 451–458.

Monfort, A. (1997) *Cours de statistique mathématique*. Economica.

Rezvanian, A. and Meybodi, M. R. (2016) Stochastic graph as a model for social networks. *Computers in Human Behavior*, **64**, 621 – 640. URL: http://www.sciencedirect.com/science/article/pii/S0747563216305222.

Scott, J. (1988) Social network analysis. *Sociology*, **22**, 109–127. URL: https://doi.org/10.1177/0038038588022001007.

Snijders, T. A., Pattison, P. E., Robins, G. L. and Handcock, M. S. (2006) New specifications for exponential random graph models. *Sociological methodology*, **36**, 99–153.

Stoica, R., Deaconu, M., Philippe, A. and Hurtado-Gil, L. (2019) Shadow Simulated Annealing algorithm: a new tool for global optimisation and statistical inference. URL: https://hal.archives-ouvertes.fr/hal-02183506. Working paper or preprint.

Stoica, R. S., Philippe, A., Gregori, P. and Mateu, J. (2017) Abc shadow algorithm: a tool for statistical analysis of spatial patterns. *Statistics and computing*, **27**, 1225–1238.

Van Lieshout, M. and Stoica, R. S. (2003) The candy model: properties and inference. *Statistica Neerlandica*, **57**, 177–206.

Wasserman, S. and Pattison, P. (1996) Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. *Psychometrika*, **61**, 401–425.

Winkler, G. (2013) *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction*. Springer Science & Business Media.

A. Teams summary

Table 2. Sufficient statistics of each observed collaboration graph

	t_{11}	t_{12}	t_{22}	LORIA members	External collaborators
BISCUIT	18.0	36.0	86.0	12	17
CAPSID	104.0	273.0	268.0	13	30
CARAMBA	3.0	4.0	5.0	7	9
CARTE	6.0	16.0	38.0	7	12
COAST	55.0	70.0	39.0	17	10
GAMBLE	7.0	95.0	204.0	8	26
KIWI	3.0	6.0	27.0	7	12
LARSEN	212.0	270.0	172.0	25	17
MADYNES	153.0	195.0	175.0	23	33
MAIA	234.0	18.0	17.0	25	8
MOSEL	3.0	16.0	78.0	10	16
MULTISPEECH	153.0	651.0	866.0	30	82
NEUROSYS	43.0	192.0	202.0	13	23
ORPAILLEUR	324.0	591.0	453.0	31	54
PAROLE	95.0	309.0	299.0	10	16
PESTO	22.0	59.0	318.0	14	38
RESIST	47.0	81.0	66.0	14	14
SEMAGRAMME	1.0	15.0	801.0	8	45
SIMBIOT	123.0	116.0	102.0	12	14
SMarT	108.0	349.0	325.0	14	19
SYNALP	5.0	24.0	29.0	7	14

Table 3. Statistics on the number of internal and external stakeholders accounted for each team

	Mean	Median	Standard deviation
Number of LORIA's members	14.82	13.0	7.38
Number of external collaborators	24.91	17.0	17.51

Table 4. Summary of estimates obtained from the collaboration networks of teams for the parameter θ_{11} and θ_{12}. (ordered by ascending θ_{11} median values)

	mean θ_{11}	Q50 θ_{11}	MAP θ_{11}	mean θ_{12}	Q50 θ_{12}	MAP θ_{12}
GAMBLE	-1.024	-0.943	-0.778	0.079	0.067	0.041
SYNALP	-0.312	-0.249	-0.041	-0.175	-0.184	-0.201
CARAMBA	-0.280	-0.177	0.048	-0.965	-0.920	-0.819
MOSEL	-0.300	-0.174	0.081	-0.394	-0.386	-0.381
SEMAGRAMME	-0.334	-0.144	0.326	-0.202	-0.187	-0.155
MAIA	-0.110	-0.110	-0.110	-0.944	-0.932	-0.904
NEUROSYS	-0.104	-0.084	-0.050	-0.074	-0.080	-0.087
COAST	-0.075	-0.067	-0.048	-0.364	-0.364	-0.366
MULTISPEECH	-0.035	-0.033	-0.036	-0.136	-0.134	-0.132
LARSEN	-0.023	-0.024	-0.019	-0.230	-0.230	-0.230
ORPAILLEUR	-0.007	-0.008	-0.017	-0.174	-0.173	-0.172
MADYNES	0.008	0.007	0.002	-0.266	-0.266	-0.266
RESIST	0.000	0.010	0.022	-0.279	-0.280	-0.278
SMarT	0.021	0.027	0.043	-0.081	-0.083	-0.087
KIWI	-0.066	0.037	0.238	-0.614	-0.585	-0.543
BISCUIT	0.036	0.054	0.076	-0.356	-0.353	-0.343
CAPSID	0.084	0.086	0.087	-0.122	-0.121	-0.120
SIMBIOT	0.113	0.112	0.108	-0.220	-0.219	-0.214
PESTO	0.115	0.122	0.125	-0.257	-0.255	-0.252
PAROLE	0.123	0.129	0.135	-0.071	-0.072	-0.077
CARTE	0.081	0.138	0.251	-0.360	-0.354	-0.351

Table 5. Summary of estimates obtained from the collaboration networks of teams for the parameter θ_{22}. (ordered by ascending median values)

	mean θ_{22}	Q50 θ_{22}	MAP θ_{22}
CARAMBA	-1.185	-1.129	-1.026
SYNALP	-0.405	-0.392	-0.355
KIWI	-0.366	-0.356	-0.330
MADYNES	-0.320	-0.319	-0.317
ORPAILLEUR	-0.272	-0.275	-0.278
MULTISPEECH	-0.248	-0.250	-0.249
BISCUIT	-0.243	-0.240	-0.233
COAST	-0.246	-0.238	-0.225
MOSEL	-0.238	-0.234	-0.226
RESIST	-0.227	-0.222	-0.211
CARTE	-0.221	-0.215	-0.194
PESTO	-0.199	-0.199	-0.200
MAIA	-0.186	-0.171	-0.145
NEUROSYS	-0.168	-0.165	-0.160
GAMBLE	-0.154	-0.152	-0.146
CAPSID	-0.142	-0.142	-0.141
LARSEN	-0.132	-0.132	-0.130
SIMBIOT	-0.064	-0.062	-0.055
SEMAGRAMME	-0.056	-0.056	-0.052
SMarT	-0.042	-0.041	-0.038
PAROLE	0.029	0.030	0.031

Table 6. Error of the estimations : Asymptotic standard deviation and Monte Carlo Standard deviation

	$\hat{\sigma}_{\theta_{01}}$	$\hat{\sigma}_{\theta_{02}}$	$\hat{\sigma}_{\theta_{12}}$	$\hat{\sigma}^{MC}_{\theta_{01}}$	$\hat{\sigma}^{MC}_{\theta_{02}}$	$\hat{\sigma}^{MC}_{\theta_{12}}$
BISCUIT	2.028e-01	8.423e-02	5.27e-02	9.914e-06	6.770e-07	1.040e-07
CAPSID	7.532e-02	3.816e-02	2.788e-02	8.905e-08	1.501e-08	5.652e-09
CARAMBA	2.882e-01	3.231e-01	3.947e-01	4.865e-05	8.172e-05	2.097e-04
CARTE	2.062e-01	1.172e-01	8.299e-02	9.468e-06	1.293e-06	4.131e-07
COAST	5.289e-02	7.463e-02	9.841e-02	7.6e-08	2.484e-07	6.988e-07
GAMBLE	7.287e-01	1.183e-01	3.963e-02	2.174e-03	4.45e-05	2.096e-06
KIWI	3.501e-01	1.93e-01	1.125e-01	1.145e-04	1.135e-05	1.655e-06
LARSEN	3.060e-02	3.363e-02	3.473e-02	3.824e-09	7.184e-09	8.535e-09
MADYNES	6.213e-02	4.128e-02	3.593e-02	4.262e-08	1.168e-08	1.075e-08
MAIA	1.857e-02	1.334e-01	1.228e-01	2.556e-08	2.989e-06	2.006e-06
MOSEL	5.164e-01	1.208e-01	5.702e-02	5.532e-04	1.18e-05	3.384e-07
MULTISPEECH	7.395e-02	2.736e-02	1.663e-02	1.335e-07	1.29e-08	2.021e-09
NEUROSYS	1.434e-01	6.373e-02	4.112e-02	1.688e-06	2.306e-07	5.947e-08
ORPAILLEUR	4.354e-02	2.803e-02	2.371e-02	7.301e-09	1.935e-09	1.867e-09
PAROLE	7.777e-02	3.871e-02	2.722e-02	9.523e-08	1.686e-08	6.413e-09
PESTO	5.22e-01	6.130e-02	2.306e-02	6.427e-04	4.628e-06	2.514e-08
RESIST	7.470e-02	6.591e-02	6.56e-02	1.053e-07	8.614e-08	1.194e-07
SEMAGRAMME	3.859e+00	6.954e-02	9.888e-03	2.192e+00	2.11e-04	1.262e-08
SIMBIOT	5.737e-02	4.300e-02	3.452e-02	3.625e-08	1.431e-08	8.900e-09
SMarT	6.678e-02	3.733e-02	2.731e-02	4.905e-08	1.012e-08	4.741e-09
SYNALP	4.247e-01	1.601e-01	1.215e-01	1.790e-04	1.36e-05	3.603e-06

Table 7. Ranges of confidence intervals 95% for estimated MAPs computed from the MC variance of the table 6

	CI 95% θ_{11}	CI 95% θ_{12}	CI 95% θ_{22}
BISCUIT	0.076 ± 1.983e-05	-0.343 ± 1.354e-06	-0.233 ± 2.080e-07
CAPSID	0.087 ± 1.781e-07	-0.12 ± 3.002e-08	-0.141 ± 1.130e-08
CARAMBA	0.048 ± 9.730e-05	-0.819 ± 1.634e-04	-1.026 ± 4.193e-04
CARTE	0.251 ± 1.894e-05	-0.351 ± 2.585e-06	-0.194 ± 8.262e-07
COAST	-0.048 ± 1.52e-07	-0.366 ± 4.968e-07	-0.225 ± 1.398e-06
GAMBLE	-0.778 ± 4.347e-03	0.041 ± 8.899e-05	-0.146 ± 4.191e-06
KIWI	0.238 ± 2.29e-04	-0.543 ± 2.270e-05	-0.33 ± 3.31e-06
LARSEN	-0.019 ± 7.649e-09	-0.23 ± 1.437e-08	-0.13 ± 1.707e-08
MADYNES	0.002 ± 8.523e-08	-0.266 ± 2.335e-08	-0.317 ± 2.150e-08
MAIA	-0.11 ± 5.112e-08	-0.904 ± 5.978e-06	-0.145 ± 4.012e-06
MOSEL	0.081 ± 1.106e-03	-0.381 ± 2.36e-05	-0.226 ± 6.768e-07
MULTISPEECH	-0.036 ± 2.671e-07	-0.132 ± 2.58e-08	-0.249 ± 4.042e-09
NEUROSYS	-0.05 ± 3.376e-06	-0.087 ± 4.611e-07	-0.16 ± 1.189e-07
ORPAILLEUR	-0.017 ± 1.460e-08	-0.172 ± 3.870e-09	-0.278 ± 3.734e-09
PAROLE	0.135 ± 1.905e-07	-0.077 ± 3.372e-08	0.031 ± 1.283e-08
PESTO	0.125 ± 1.285e-03	-0.252 ± 9.256e-06	-0.2 ± 5.028e-08
RESIST	0.022 ± 2.107e-07	-0.278 ± 1.723e-07	-0.211 ± 2.387e-07
SEMAGRAMME	0.326 ± 4.385e+00	-0.155 ± 4.219e-04	-0.052 ± 2.524e-08
SIMBIOT	0.108 ± 7.249e-08	-0.214 ± 2.863e-08	-0.055 ± 1.780e-08
SMarT	0.043 ± 9.811e-08	-0.087 ± 2.024e-08	-0.038 ± 9.483e-09
SYNALP	-0.041 ± 3.580e-04	-0.201 ± 2.72e-05	-0.355 ± 7.206e-06

Table 8. Results of the t-test applied on each parameter to check if the parameter are significant against pure chance. He
of the test as well as the corresponding p-value are presented. Except for one team marked by *, the parameters are signi

	$\mathrm{TS}(\theta_{11}, 0)$	p-val1	$TS(\theta_{12}, 0)$	p-val2	$TS(\theta_{22}, 0)$	p-val3
BISCUIT	137.162	$\leq 10^{-6}$	-1570.940	$\leq 10^{-6}$	-1386.841	$\leq 10^{-6}$
CAPSID	602.657	$\leq 10^{-6}$	-1090.538	$\leq 10^{-6}$	-1427.932	$\leq 10^{-6}$
CARAMBA	-219.174	$\leq 10^{-6}$	-762.995	$\leq 10^{-6}$	-885.889	$\leq 10^{-6}$
CARTE	102.744	$\leq 10^{-6}$	-881.310	$\leq 10^{-6}$	-787.896	$\leq 10^{-6}$
COAST	-401.836	$\leq 10^{-6}$	-1410.560	$\leq 10^{-6}$	-753.956	$\leq 10^{-6}$
GAMBLE	-486.201	$\leq 10^{-6}$	217.581	$\leq 10^{-6}$	-1158.119	$\leq 10^{-6}$
KIWI	-52.426	$\leq 10^{-6}$	-820.701	$\leq 10^{-6}$	-962.094	$\leq 10^{-6}$
LARSEN	-200.931	$\leq 10^{-6}$	-1926.265	$\leq 10^{-6}$	-906.471	$\leq 10^{-6}$
MADYNES	49.288	$\leq 10^{-6}$	-1767.626	$\leq 10^{-6}$	-1612.832	$\leq 10^{-6}$
MAIA	-1214.453	$\leq 10^{-6}$	-1867.597	$\leq 10^{-6}$	-495.504	$\leq 10^{-6}$
MOSEL	-212.150	$\leq 10^{-6}$	-1037.920	$\leq 10^{-6}$	-1254.859	$\leq 10^{-6}$
MULTISPEECH	-80.189	$\leq 10^{-6}$	-434.440	$\leq 10^{-6}$	-926.255	$\leq 10^{-6}$
NEUROSYS	-244.861	$\leq 10^{-6}$	-392.495	$\leq 10^{-6}$	-1262.550	$\leq 10^{-6}$
ORPAILLEUR	-23.905	$\leq 10^{-6}$	-662.881	$\leq 10^{-6}$	-898.293	$\leq 10^{-6}$
PAROLE	622.812	$\leq 10^{-6}$	-627.453	$\leq 10^{-6}$	325.941	$\leq 10^{-6}$
PESTO	314.309	$\leq 10^{-6}$	-1136.154	$\leq 10^{-6}$	-1338.927	$\leq 10^{-6}$
RESIST*	0.748	4.544e-01	-736.512	$\leq 10^{-6}$	-522.735	$\leq 10^{-6}$
SEMAGRAMME	-61.911	$\leq 10^{-6}$	-356.985	$\leq 10^{-6}$	-321.882	$\leq 10^{-6}$
SIMBIOT	600.459	$\leq 10^{-6}$	-984.834	$\leq 10^{-6}$	-251.961	$\leq 10^{-6}$
SMarT	47.632	$\leq 10^{-6}$	-285.202	$\leq 10^{-6}$	-183.462	$\leq 10^{-6}$
SYNALP	-74.952	$\leq 10^{-6}$	-108.668	$\leq 10^{-6}$	-319.705	$\leq 10^{-6}$