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. Collaboration graphs are relevant sources of information to understand behavioural tendencies of groups of individuals. The study of these collaboration graphs enables figuring out factors that may affect the efficiency and the sustainability of cooperative work. An example of such a collaboration involves researchers who develop relationships with their external counterparts to tackle scientific challenges. We propose a statistical approach that considers edge occurrence in the graph as a labelling process. Our approach combines spatial processes modelling and Exponential Random Graph Models (ERGMs) commonly used to analyse such social processes. Since the normalising constant involved in classical Markov Chain Monte Carlo approaches is not available in an analytic closed form, the inference remains challenging. To overcome this issue, we propose a Bayesian tool that relies on the recent ABC Shadow algorithm. The proposed method is illustrated on real data sets from an open archive of scholarly documents.

1. Introduction

Networks are widely studied mathematical objects (Bollobás, 2013, chapter 5). They describe molecular interactions, relationships between individuals in a social application, collaboration 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 Figure 1. The nodes of the graph are 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 other 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 collaboration 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 of 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 characterise 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 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. Our application considers 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 systems. They induce an extremely high number of configurations. Stochastic modelling allows us to deal with this situation. The approach we propose to consider an appropriate version of exponential graph models to represent collaborations initiated by a community 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 algorithms: the Metropolis-Hastings (MH) dynamics and the Gibbs sampler. Next, in Section 3, we describe the ABC Shadow algorithm (Stoica et al., 2017) used to build posterior based inference. Section 4 demonstrates the relevance of ABC Shadow on simulated data.

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 characterise 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 in sociology (Scott, 1988). We propose to understand intra and inter relation between organisations based on participant collaboration network. From this collaboration graph, we will associate a more relevant one, considering the relation as the object of

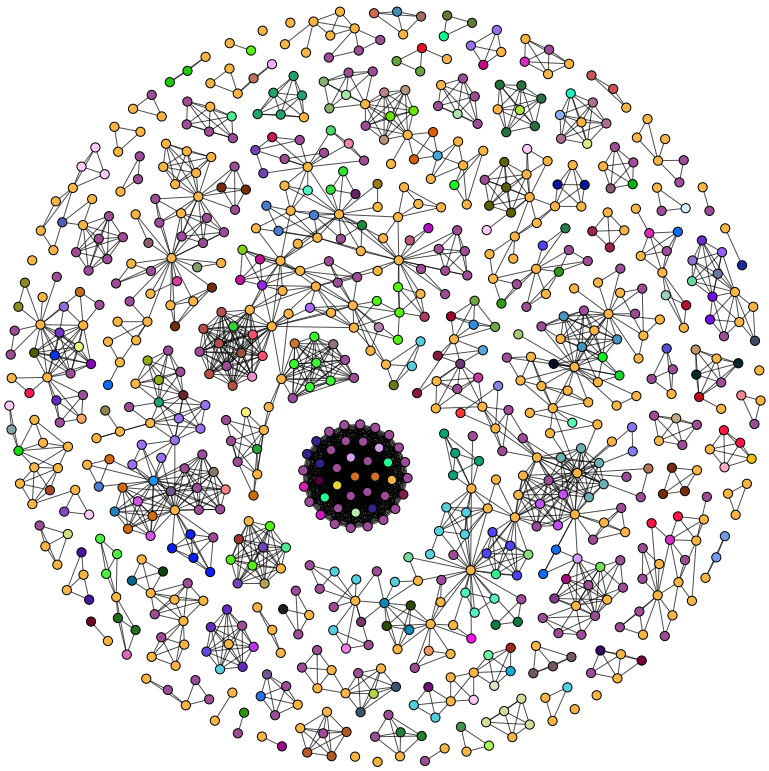


Fig. 1. Collaborations among researchers within the Loria laboratory (2018) – Each node represents a researcher, the edges are collaboration links and nodes' colour represent the affiliation to a laboratory. For example, all Loria members are coloured in yellow, while the members of the other labs are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

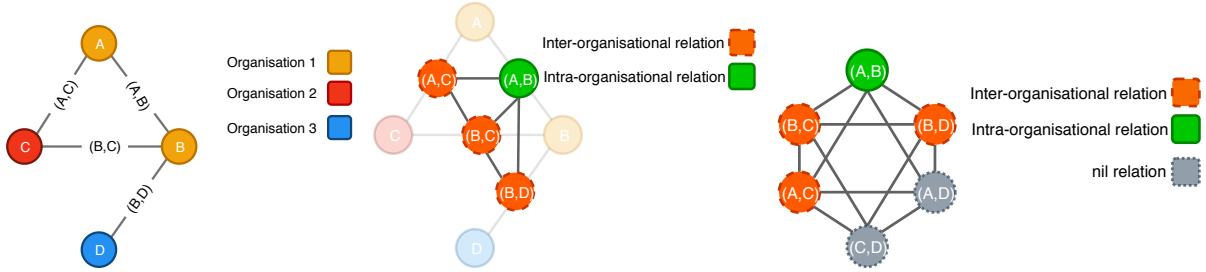
primary interested taking into account that collaboration links can be internal or external. This is what we explain in the next subsection.

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 emphasises relationships over people and allows us to reason on these relationships and the structure they propose.

Throughout this paper, we assess the extent to which 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 (plain outline) node is an intra-organisational relation, the orange (dashed outline) ones are

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



(a) Collaboration graph (b) Line graph representation (c) Augmented line graph representation *taking nil edges into account*

Fig. 2. An example of collaboration 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 labelling 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 the 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 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 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 with $\mathcal{G}$, via the labels in a phase space that we denote $\mathcal{L}$ that have been attached to each vertex. A realisation 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 thorough and rigorous presentation of MRFs we recommend 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 with 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 in hand, the aim is to characterise interactions between researchers. Let us consider the following MRF model:

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

where y is the realisation of the graph representation given by the labels $\{0, 1, 2\}$ associated with 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 of his

own team and a member of a different team. The statistic t_{12} indicates how the researchers exhibit a hub behaviour, since they interact with both kinds 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 with 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 best known sampling algorithms are the MH and the Gibbs sampler. Here, we briefly recall both algorithms and especially the Gibbs sampler (Geman and Geman, 1987), due to its role within the inference procedures used through this paper.

2.3.1. Markov Chain Monte Carlo (MCMC) simulation

The purpose is to sample distributions which are not analytically tractable such as (2). The MH algorithm (Hastings, 1970) provides a solution for this problem. The algorithm works by iterating a two-step 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.

To circumvent the high rejection rate we relied on a Gibbs procedure. Gibbs sampler is a particular case of the MH. The difference between both algorithms resides in the way of picking a new proposition. Instead of selecting a new proposition according to an auxiliary distribution which could be rejected afterwards, Gibbs sampler choose the new proposition y' according to the probability $P(Y|y^c)$ (where y^c is the current configuration). This leads to accept every move and fixing the acceptance ratio : $\alpha_{y \to y'} = 1$.

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 realisations 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's 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 also assumes 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 also be seen 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 closed to 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 by 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 simulated convergent 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 is 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

function ABC_SHADOW(θ_0, y_{obs}, m, Δ) ▷ Where θ_0 - initial parameters, y_{obs} - observation
  $x \sim p(x|\theta_0)$
  for i = 1 to m **do**
    $\theta' \sim \mathcal{U}_\Delta(\theta_{i-1} \to \theta')$
    $\alpha \leftarrow min\left\{1, \exp[(t(y_{obs}) - t(x))(\theta' - \theta_{i-1})]\frac{p(\theta')}{p(\theta_{i-1})}\right\}$
    $accepted \leftarrow \mathcal{U}(0,1)$
    if $\alpha > accepted$ **then**
      $\theta_i \leftarrow \theta'$
    else
      $\theta_i \leftarrow \theta_{i-1}$
    end if
  end for
  return θ_m
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 to use exact simulation methods Huber (2016). Here, for numerical purposes and due also to the rather weak requirements regarding the auxiliary variable,

Table 1. Statistics on the posterior of Binomial distribution

	Q_{10}	Q_{25}	Q_{50}	$mean$	Q_{75}	Q_{95}	MAP	$\hat{\sigma}_\theta$	$\hat{\sigma}_\theta^{MC}$
ABC (θ)	-0.992	-0.69	-0.383	-0.392	-0.075	0.345	-0.408	0.454	4.3×10^{-4}
MH (θ)	-0.961	-0.672	-0.371	-0.377	-0.071	0.353	-0.3718	0.453	4.2×10^{-4}

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 the 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 1.002×10^6 posteriors. The first 2×10^3 samples were cut off and a subsampling kept every 100 samples. This resulted in a chain $(\theta^{(t)})_{t=1,\cdots,T}$ of 10^4 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 $iters = 1.002 \times 10^6$. Like the MH, the output of Algorithm 2 is a chain $(\theta^{(t)})_{t=1,\cdots,T}$ that we subsample, keeping only every 100. 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 3 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 noticing 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 summarised in Table 1. 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.

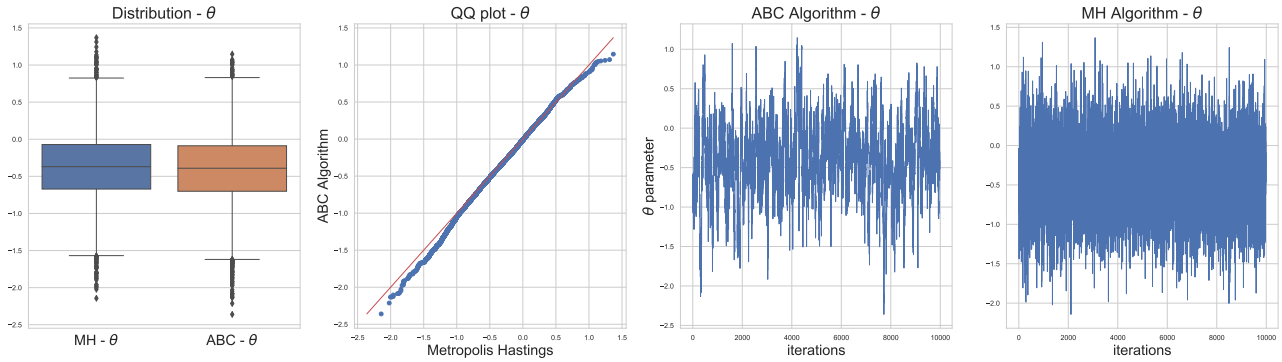


Fig. 3. Posterior sampling of a Bernoulli distribution using Metropolis Hasting and ABC Shadow

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 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.

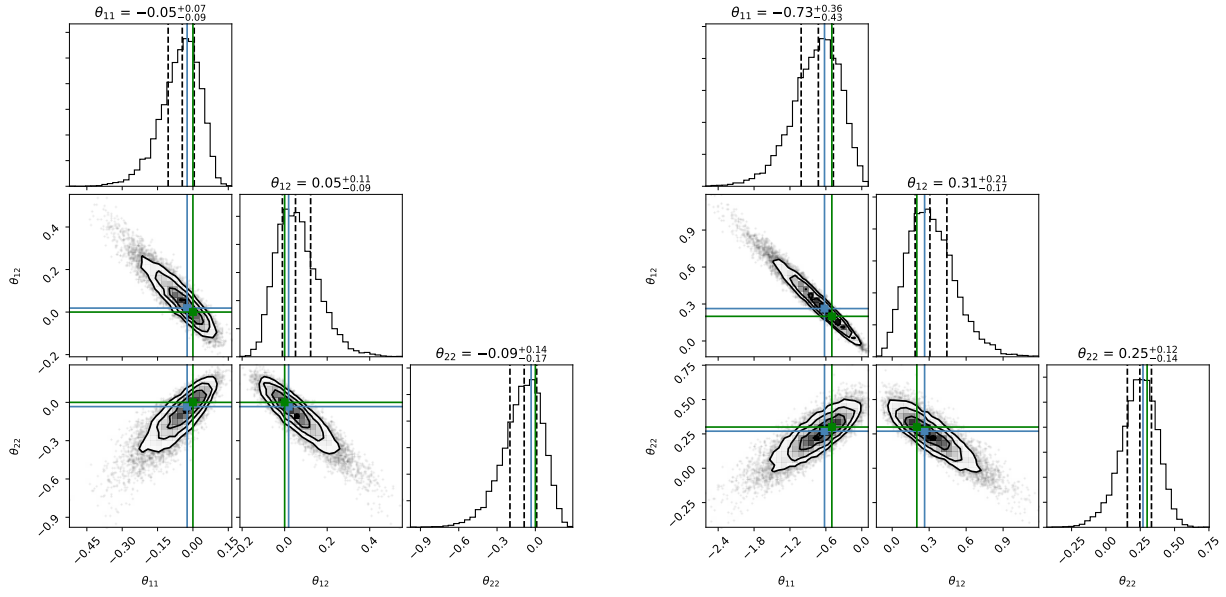
In the first experiment, all interaction parameters were fixed to 0 : $\theta_{11} = \theta_{12} = \theta_{22} = 0$, so that interaction effects are annihilated. Since we have three type of patterns, this leads to a Bernoulli graph model with an occurrence probability for each pattern equal to $\frac{1}{3}$. The observation represents an artificial collaboration involving 12 members of the same organisation and 8 collaborators from the outside i.e. with a $size = (12, 8)$. It was generated from $N = 10^3$ samples yielded by a Gibbs sampler. By averaging sufficient statistics we obtain $\bar{t}(y) = [164.747, 263.495, 83.7645]$ (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 $iters = 1.002 \times 10^6$ samples. We subsampled keeping every 100 value and rejected the 2×10^3 first burn in samples. At each iteration the auxiliary variable x was updated using 200 steps of a Gibbs sampler. Error metrics were computed: the asymptotic standard deviation $\hat{\sigma}_\theta = [0.08, 0.093, 0.144]$ and the Monte Carlo standard deviation $\hat{\sigma}_\theta^{MC} = [3.80 \times 10^{-7}, 5.80 \times 10^{-7}, 1.98 \times 10^{-6}]$.

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: $\hat{\theta} = [-0.0262, 0.036, -0.0436]$. The green lines are the true parameter values : $\theta = [0, 0, 0]$.

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 $size = (12, 8)$ using a Gibbs sampler as we did before. The generated observation yielded the following averaged sufficient statistics: $\bar{t}(y) = [79.1769, 361.796, 296.235]$. Figure 4b represents the resulting posterior distribution. Blue lines representing the MAP are aligned on $\hat{\theta} = [-0.6228, 0.263, 0.2698]$ which is close to the true parameter $\theta = [-0.5, 0.2, 0.3]$ represented with 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 re-

spectively: $[-0.769, 0.325, 0.24]$ and $[-0.728, 0.307, 0.247]$. The error metrics, respectively the asymptotic standard deviation and the Monte Carlo standard deviation are : $\hat{\sigma}_\theta = [0.08, 0.093, 0.144]$ and $\hat{\sigma}_\theta^{MC} = [3.80 \times 10^{-7}, 5.80 \times 10^{-7}, 1.98 \times 10^{-6}]$.

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



(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)*

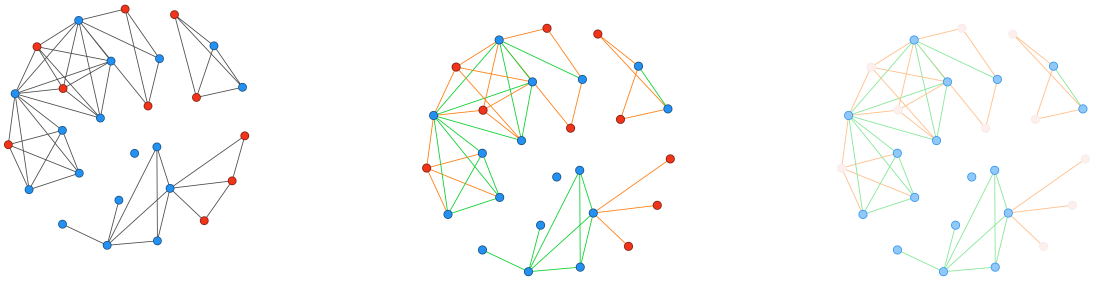
5. Application

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 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 collaborator whereas intra-organisational links connect two collaborators who are affiliated to LORIA. Under the hypothesis of the model, the sufficient statistics were computed. The results are presented in Table 2 (in Appendix A).



(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. An 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 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 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 $iters = 1.002 \times 10^6$ 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 500 iterations of the MH procedure. The first 2000 burn in samples were discarded. In addition, 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 10^3 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, Table 4 in Appendix A present the mean, the median and the estimated MAP of the posterior distribution of θ for each team.

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. At the scale of teams this means 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. It restrict the number of publications and the underlying potential collaborations a researcher is able to undertake.

Regarding the table 4 in Appendix A, both the median and the mean are close to the estimated MAP. Similarly to Van Lieshout and Stoica (2003); Stoica et al. (2017), we computed the asymptotic standard deviation and the Monte Carlo standard deviation. The results are given by Table 5 (in Appendix A). To that end, for each estimated model we performed a simulation providing 10,000 samples (samples were decorrelated by keeping every 10^3 sample out of the 10^7 simulated). Given the Monte Carlo standard deviation, we can determinate the 95% confidence interval reported by Table 6 (in Appendix A).

The closeness of value ranges to 0 raises the question of their significance. Are the three studied patterns more 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 7 (in Appendix A). For most of the teams, the parameters are significantly non-zero since the associated p-values of the t-test are very small. There are only two teams 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. In conclusion, for almost all teams (except the two latter mentioned), the likelihood of link creation is not merely due to 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 collaboration links with the outside 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 in Appendix A). 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 (in

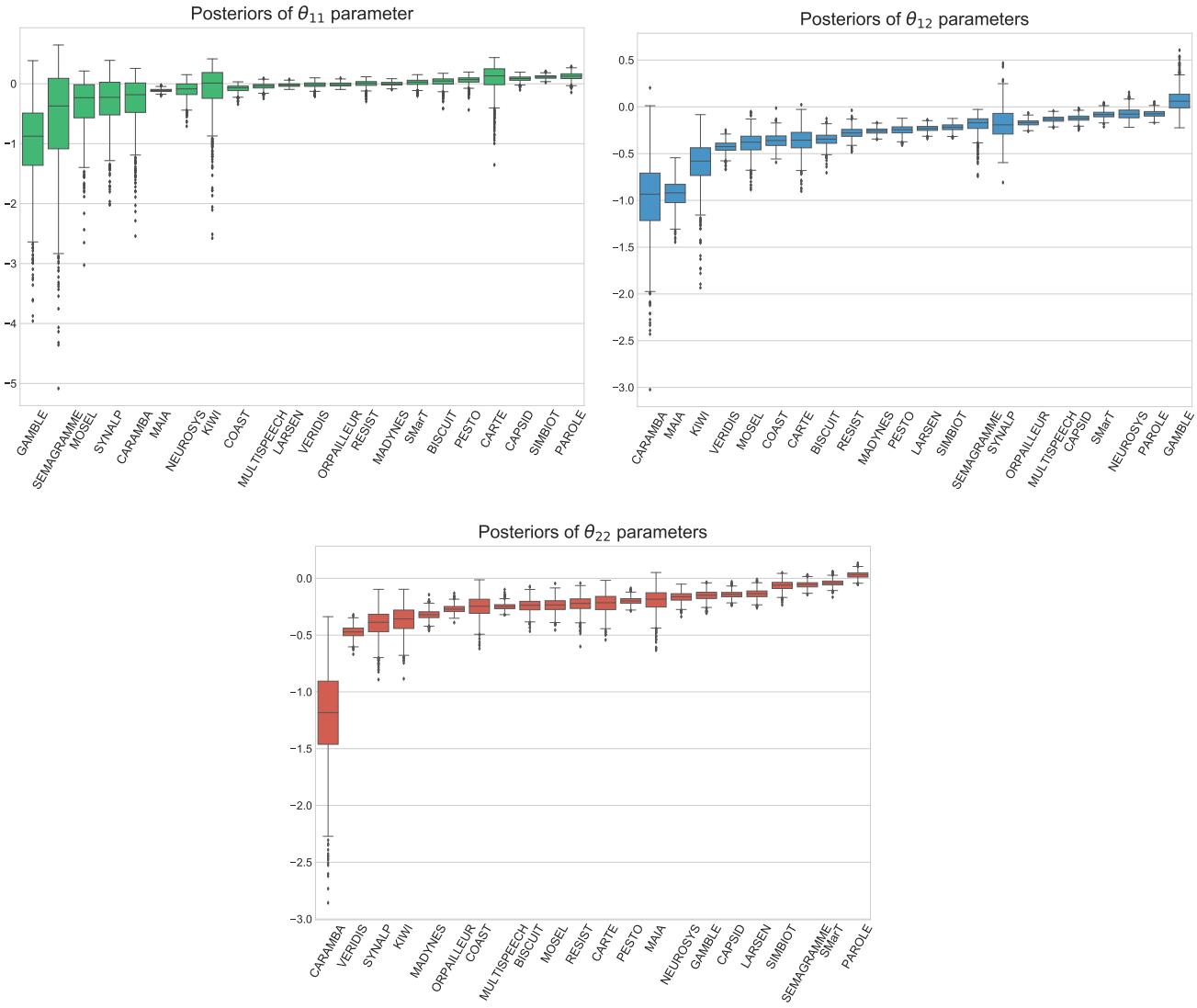


Fig. 6. Box plots of the posterior distributions for the parameters of the Potts model estimated from the collaboration graphs of each team (ordered by ascending mean values)

Appendix A), 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 not 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

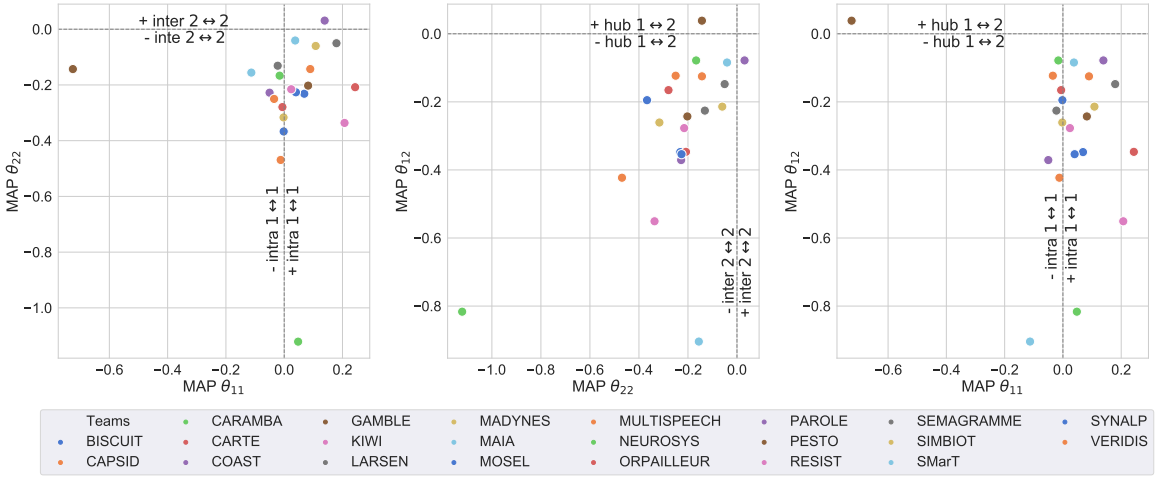


Fig. 7. Scatter plots of estimated MAPs representing the positioning of teams with respect to the different collaborative tendencies controlled by θ_{01}, θ_{02}, θ_{12}

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.

The lab is organised in 5 departments gathering teams working on the same research thematic. The team's names are coloured according to the affiliation to one of these departments. The clustered team's names are not similarly coloured and then, don't necessarily work on the same topic. Consequently, structural patterns are not a feature specific to the research thematic. We also noticed that the closeness between two teams can be related to the fact that one originates from the other. It is not unusual that a researcher keep signing with an old affiliation a long time after the creation of a new team. Also, when a team splits in new teams, members of teams keep collaborating. 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.

6. Conclusion

In this paper, we proposed a method to make inference on structural aspects of collaboration networks. We focused on inter-organisational collaborations yet sparsely addressed by the state of the art. For instance, researchers from different organisation often collaborate to conduct research and write publications. We extracted the collaboration network among researchers by considering the co-authorship of publications from the French open-archive HAL as collaboration links between authors.

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

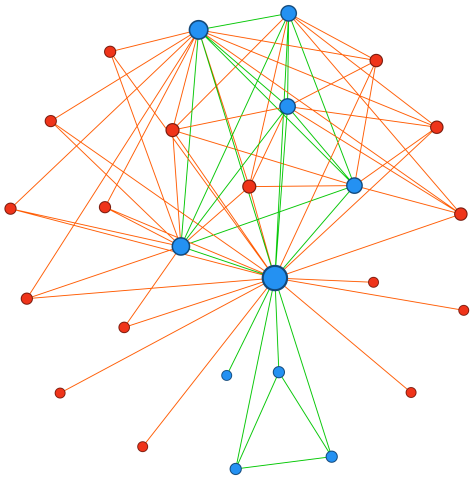


Fig. 8. Co-authorship graph of the team PAROLE. Blue nodes represents the members of LORIA, whereas red nodes are external collaborators. Size of nodes is proportional to their degree. The inter-organisational links are coloured in orange, while the intra-organisational links are in green.

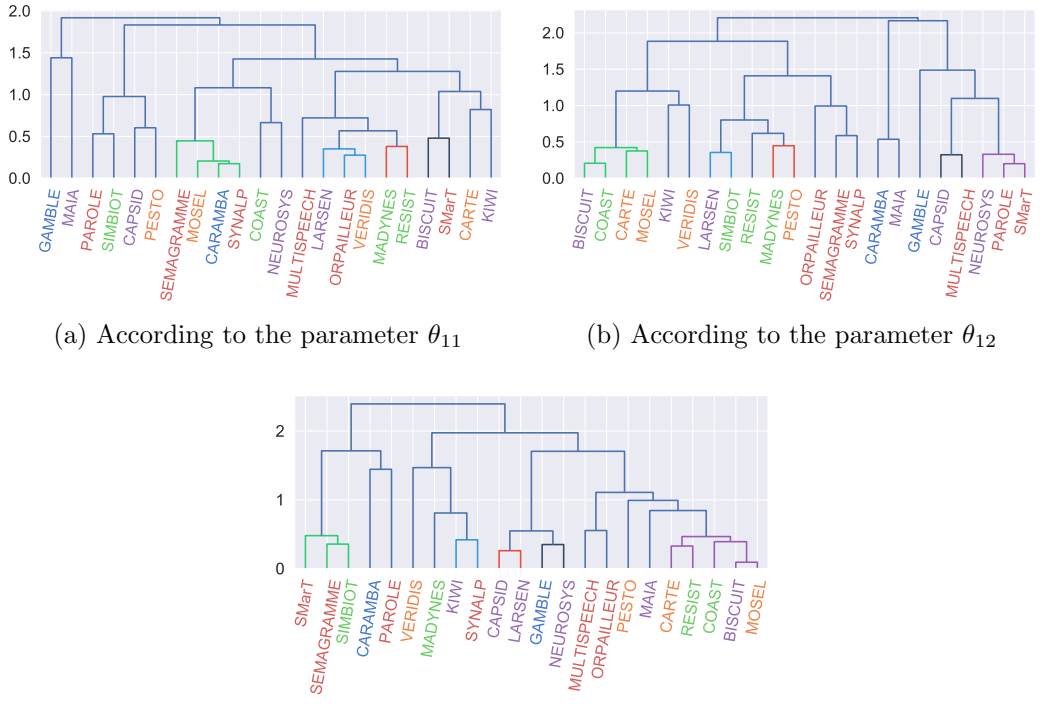
Ising model, the Potts model, to describe the interactions between relations. As for all exponential models, the inference remains difficult due to the intractable normalising constant. To that end, we used a Bayesian tool, the ABC Shadow algorithm which was firstly tuned on tractable model and simulated data. We applied it on collaboration networks of different 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 a few actors play a key role since they connect collaborators of their organisation toward the outside. Given the posteriors, we also demonstrated how to classify the way the different teams collaborate and conclude that structural features at stake are not related to the scientific topic addressed.

The sizes of the teams are relatively disparate and may affect the prevalence of the observed patterns. One of the first perspectives to pursue is to come up with a normalising procedure to put all the observed graph on the same level.

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.

The selection of the model is a very sensitive aspect of our approach which might influence the relevance of the estimates in regard to the observation. This concern should be further investigated in a future work (Caimo and Friel, 2013). For instance, other classes of models such as the Markov Connected Component Fields (Møller and Waagepetersen, 1998) might be good candidates for structural graph pattern analysis.

Finally, the 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



(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.

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.

Caimo, A. and Friel, N. (2013) Bayesian model selection for exponential random graph models. *Social Networks*, **35**, 11–24.

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

Geman, S. and Geman, D. (1987) Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. In *Readings in computer vision*, 564–584. Elsevier.

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*. Chapman and Hall/CRC.

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.

Møller, J. and Waagepetersen, R. P. (1998) Markov connected component fields. *Advances in Applied Probability*, **30**, 1–35.

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. Numerical results and empirical statistics

Table 2. Sufficient statistics of each observed collaboration graph

	t_{11}	t_{12}	t_{22}	LORIA members	External collaborators
BISCUIT	18	36	86	12	17
CAPSID	104	273	268	13	30
CARAMBA	3	4	5	7	9
CARTE	6	16	38	7	12
COAST	55	70	39	17	10
GAMBLE	7	95	204	8	26
KIWI	3	6	27	7	12
LARSEN	212	270	172	25	17
MADYNES	153	195	175	23	33
MAIA	234	18	17	25	8
MOSEL	3	16	78	10	16
MULTISPEECH	153	651	866	30	82
NEUROSYS	43	192	202	13	23
ORPAILLEUR	324	591	453	31	54
PAROLE	95	309	299	10	16
PESTO	22	59	318	14	38
RESIST	47	81	66	14	14
SEMAGRAMME	1	15	801	8	45
SIMBIOT	123	116	102	12	14
SMarT	108	349	325	14	19
SYNALP	5	24	29	7	14
VERIDIS	29	44	142	19	39

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 parameters θ_{11}, θ_{12} and θ_{22}.

	mean θ_{11}	Q50 θ_{11}	MAP θ_{11}	mean θ_{12}	Q50 θ_{12}	MAP θ_{12}	mean θ_{22}	Q50 θ_{22}	MAP θ_{22}
BISCUIT	0.024942	0.044401	0.068886	-0.352912	-0.349398	-0.347587	-0.242338	-0.239895	-0.231520
CAPSID	0.082948	0.084934	0.089377	-0.121275	-0.120743	-0.125009	-0.141760	-0.141644	-0.142977
CARAMBA	-0.281506	-0.172194	0.047788	-0.982179	-0.940077	-0.816491	-1.201228	-1.161985	-1.121262
CARTE	0.075474	0.132348	0.243351	-0.360031	-0.354322	-0.346872	-0.219448	-0.213238	-0.208072
COAST	-0.076206	-0.068637	-0.050184	-0.363103	-0.363962	-0.371046	-0.248358	-0.240338	-0.227805
GAMBLE	-0.975762	-0.876918	-0.725163	0.071242	0.057308	0.038471	-0.151945	-0.149128	-0.143048
KIWI	-0.093009	0.010038	0.207023	-0.611750	-0.584043	-0.550839	-0.362463	-0.353860	-0.336326
LARSEN	-0.024114	-0.024729	-0.022328	-0.228011	-0.227737	-0.225750	-0.133564	-0.132721	-0.131027
MADYNES	0.000685	0.000674	-0.002234	-0.260162	-0.259588	-0.260691	-0.320587	-0.320209	-0.316542
MAIA	-0.111905	-0.112035	-0.112810	-0.929088	-0.918453	-0.904274	-0.197517	-0.181067	-0.155695
MOSEL	-0.330711	-0.226605	0.040198	-0.386187	-0.378969	-0.353764	-0.237875	-0.234533	-0.226377
MULTISPEECH	-0.040704	-0.038330	-0.034666	-0.133189	-0.131164	-0.123168	-0.248131	-0.249479	-0.250201
NEUROSYS	-0.104963	-0.085756	-0.015696	-0.073340	-0.077790	-0.078339	-0.168132	-0.165696	-0.166646
ORPAILLEUR	-0.011312	-0.012008	-0.006116	-0.170487	-0.169386	-0.165536	-0.271443	-0.274956	-0.279405
PAROLE	0.121960	0.128505	0.138845	-0.070416	-0.071778	-0.078252	0.028991	0.030013	0.030928
PESTO	0.057142	0.068986	0.082127	-0.246215	-0.243335	-0.242821	-0.199743	-0.200147	-0.202435
RESIST	-0.000791	0.008758	0.024020	-0.277055	-0.277438	-0.277127	-0.226005	-0.221971	-0.215415
SEMAGRAMME	-0.590010	-0.353481	0.179312	-0.190075	-0.172762	-0.147841	-0.055775	-0.055653	-0.050246
SIMBIOT	0.110817	0.110092	0.108077	-0.217518	-0.216584	-0.214017	-0.063965	-0.062083	-0.060041
SMarT	0.019884	0.027093	0.037714	-0.081083	-0.082015	-0.084403	-0.041422	-0.041232	-0.040489
SYNALP	-0.292751	-0.213665	-0.001850	-0.177759	-0.185366	-0.195060	-0.403532	-0.392149	-0.366982
VERIDIS	-0.022496	-0.016303	-0.012097	-0.425249	-0.422178	-0.422928	-0.472263	-0.471167	-0.469137

Table 5. 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	2.416e-01	8.928e-02	5.240e-02	2.136e-05	1.158e-06	1.256e-07
CAPSID	7.597e-02	3.748e-02	2.725e-02	9.851e-08	1.551e-08	5.371e-09
CARAMBA	2.566e-01	3.25e-01	4.398e-01	3.371e-05	8.405e-05	3.261e-04
CARTE	1.959e-01	1.160e-01	8.626e-02	7.453e-06	1.172e-06	4.700e-07
COAST	5.191e-02	7.527e-02	9.645e-02	7.089e-08	2.589e-07	6.514e-07
GAMBLE	6.955e-01	1.156e-01	3.866e-02	1.786e-03	3.866e-05	1.824e-06
KIWI	5.474e-01	2.129e-01	1.126e-01	6.943e-04	2.772e-05	2.271e-06
LARSEN	3.213e-02	3.416e-02	3.537e-02	4.332e-09	7.235e-09	9.206e-09
MADYNES	6.458e-02	4.198e-02	3.541e-02	5.177e-08	1.317e-08	1.059e-08
MAIA	1.925e-02	1.285e-01	1.171e-01	2.589e-08	2.598e-06	1.653e-06
MOSEL	4.892e-01	1.144e-01	5.718e-02	4.384e-04	9.934e-06	3.933e-07
MULTISPEECH	6.601e-02	2.757e-02	1.77e-02	7.338e-08	9.047e-09	1.969e-09
NEUROSYS	1.095e-01	6.102e-02	4.721e-02	4.127e-07	8.733e-08	4.66e-08
ORPAILLEUR	3.343e-02	2.626e-02	2.555e-02	2.249e-09	1.133e-09	2.319e-09
PAROLE	7.498e-02	3.783e-02	2.716e-02	7.88e-08	1.422e-08	5.773e-09
PESTO	5.353e-01	6.144e-02	2.383e-02	7.103e-04	5.017e-06	4.502e-08
RESIST	7.333e-02	6.691e-02	6.833e-02	9.911e-08	9.090e-08	1.420e-07
SEMAGRAMME	4.327e+00	6.566e-02	9.433e-03	3.473e+00	2.458e-04	1.403e-08
SIMBIOT	5.046e-02	4.174e-02	3.573e-02	2.046e-08	1.263e-08	9.616e-09
SMarT	6.991e-02	3.879e-02	2.830e-02	6.08e-08	1.244e-08	5.721e-09
SYNALP	3.709e-01	1.591e-01	1.315e-01	9.148e-05	9.156e-06	3.786e-06
VERIDIS	3.444e-01	8.289e-02	4.651e-02	1.057e-04	2.592e-06	1.147e-07

Table 6. Ranges of confidence intervals 95% for estimated MAPs computed from the MC standard deviation of Table 5

	CI 95% θ_{11}	CI 95% θ_{12}	CI 95% θ_{22}
BISCUIT	0.068886 ± 4.272e-05	-0.347587 ± 2.316e-06	-0.23152 ± 2.512e-07
CAPSID	0.089377 ± 1.970e-07	-0.125009 ± 3.102e-08	-0.142977 ± 1.074e-08
CARAMBA	0.047788 ± 6.741e-05	-0.816491 ± 1.681e-04	-1.121262 ± 6.523e-04
CARTE	0.243351 ± 1.491e-05	-0.346872 ± 2.344e-06	-0.208072 ± 9.400e-07
COAST	-0.050184 ± 1.418e-07	-0.371046 ± 5.179e-07	-0.227805 ± 1.303e-06
GAMBLE	-0.725163 ± 3.573e-03	0.038471 ± 7.731e-05	-0.143048 ± 3.649e-06
KIWI	0.207023 ± 1.389e-03	-0.550839 ± 5.544e-05	-0.336326 ± 4.543e-06
LARSEN	-0.022328 ± 8.663e-09	-0.22575 ± 1.447e-08	-0.131027 ± 1.841e-08
MADYNES	-0.002234 ± 1.035e-07	-0.260691 ± 2.634e-08	-0.316542 ± 2.119e-08
MAIA	-0.11281 ± 5.178e-08	-0.904274 ± 5.195e-06	-0.155695 ± 3.306e-06
MOSEL	0.040198 ± 8.769e-04	-0.353764 ± 1.987e-05	-0.226377 ± 7.866e-07
MULTISPEECH	-0.034666 ± 1.468e-07	-0.123168 ± 1.809e-08	-0.250201 ± 3.938e-09
NEUROSYS	-0.015696 ± 8.255e-07	-0.078339 ± 1.747e-07	-0.166646 ± 9.32e-08
ORPAILLEUR	-0.006116 ± 4.497e-09	-0.165536 ± 2.267e-09	-0.279405 ± 4.637e-09
PAROLE	0.138845 ± 1.576e-07	-0.078252 ± 2.844e-08	0.030928 ± 1.155e-08
PESTO	0.082127 ± 1.421e-03	-0.242821 ± 1.003e-05	-0.202435 ± 9.005e-08
RESIST	0.02402 ± 1.982e-07	-0.277127 ± 1.818e-07	-0.215415 ± 2.841e-07
SEMAGRAMME	0.179312 ± 6.946e+00	-0.147841 ± 4.916e-04	-0.050246 ± 2.805e-08
SIMBIOT	0.108077 ± 4.091e-08	-0.214017 ± 2.526e-08	-0.060041 ± 1.923e-08
SMarT	0.037714 ± 1.216e-07	-0.084403 ± 2.488e-08	-0.040489 ± 1.144e-08
SYNALP	-0.00185 ± 1.83e-04	-0.19506 ± 1.831e-05	-0.366982 ± 7.572e-06
VERIDIS	-0.012097 ± 2.115e-04	-0.422928 ± 5.184e-06	-0.469137 ± 2.295e-07

Table 7. Results of the t-test applied on each parameter to check if the parameter are significant against pure chance. Here the score of the test as well as the corresponding p-value are presented. Except for two teams marked by *, the parameters are significant.

	TS($\theta_{11}, 0$)	p-val1	$TS(\theta_{12}, 0)$	p-val2	$TS(\theta_{22}, 0)$	p-val3
BISCUIT	30.505	$\leq 10^{-6}$	-497.520	$\leq 10^{-6}$	-435.178	$\leq 10^{-6}$
CAPSID	204.015	$\leq 10^{-6}$	-359.185	$\leq 10^{-6}$	-466.827	$\leq 10^{-6}$
CARAMBA	-68.738	$\leq 10^{-6}$	-237.339	$\leq 10^{-6}$	-293.383	$\leq 10^{-6}$
CARTE	32.086	$\leq 10^{-6}$	-276.357	$\leq 10^{-6}$	-247.134	$\leq 10^{-6}$
COAST	-141.119	$\leq 10^{-6}$	-489.385	$\leq 10^{-6}$	-262.020	$\leq 10^{-6}$
GAMBLE	-143.234	$\leq 10^{-6}$	61.396	$\leq 10^{-6}$	-368.832	$\leq 10^{-6}$
KIWI	-24.105	$\leq 10^{-6}$	-256.393	$\leq 10^{-6}$	-305.132	$\leq 10^{-6}$
LARSEN	-77.447	$\leq 10^{-6}$	-737.265	$\leq 10^{-6}$	-355.826	$\leq 10^{-6}$
MADYNES*	1.829	6.745e-02	-818.858	$\leq 10^{-6}$	-746.175	$\leq 10^{-6}$
MAIA	-431.390	$\leq 10^{-6}$	-662.307	$\leq 10^{-6}$	-187.895	$\leq 10^{-6}$
MOSEL	-78.036	$\leq 10^{-6}$	-326.151	$\leq 10^{-6}$	-398.977	$\leq 10^{-6}$
MULTISPEECH	-91.794	$\leq 10^{-6}$	-418.327	$\leq 10^{-6}$	-907.847	$\leq 10^{-6}$
NEUROSYS	-78.766	$\leq 10^{-6}$	-125.030	$\leq 10^{-6}$	-402.492	$\leq 10^{-6}$
ORPAILLEUR	-33.449	$\leq 10^{-6}$	-549.924	$\leq 10^{-6}$	-755.649	$\leq 10^{-6}$
PAROLE	195.106	$\leq 10^{-6}$	-196.361	$\leq 10^{-6}$	102.888	$\leq 10^{-6}$
PESTO	90.199	$\leq 10^{-6}$	-532.605	$\leq 10^{-6}$	-664.803	$\leq 10^{-6}$
RESIST*	-1.157	2.471e-01	-477.750	$\leq 10^{-6}$	-346.625	$\leq 10^{-6}$
SEMAGRAMME	-66.417	$\leq 10^{-6}$	-209.308	$\leq 10^{-6}$	-206.111	$\leq 10^{-6}$
SIMBIOT	369.346	$\leq 10^{-6}$	-599.731	$\leq 10^{-6}$	-151.456	$\leq 10^{-6}$
SMarT	36.193	$\leq 10^{-6}$	-225.532	$\leq 10^{-6}$	-144.832	$\leq 10^{-6}$
SYNALP	-68.929	$\leq 10^{-6}$	-114.893	$\leq 10^{-6}$	-318.510	$\leq 10^{-6}$
VERIDIS	-44.108	$\leq 10^{-6}$	-709.280	$\leq 10^{-6}$	-914.052	$\leq 10^{-6}$