Aggregation models on hypergraphs

Aggregation models on hypergraphs

Diego Alberici[a], Pierluigi Contucci[a], Emanuele Mingione[a], Marco Molari [a,1,*]

[a] *Dipartimento di Matematica, University of Bologna, Piazza di Porta San Donato 5, 40126 Bologna, Italy*

Abstract

Following a newly introduced approach by Rasetti and Merelli we investigate the possibility to extract topological information about the space where interacting systems are modelled. From the statistical datum of their observable quantities, like the correlation functions, we show how to reconstruct the activities of their constitutive parts which embed the topological information. The procedure is implemented on a class of polymer models on hypergraphs with hard-core interactions. We show that the model fulfils a set of iterative relations for the partition function that generalise those introduced by Heilmann and Lieb for the monomer-dimer case. After translating those relations into structural identities for the correlation functions we use them to test the precision and the robustness of the inverse problem. Finally the possible presence of a further interaction of peer-to-peer type is considered and a criterion to discover it is identified.

Keywords: Networks, hypergraphs, inverse problem, complex systems

1. Introduction and Results

In a recent paper [1] a new perspective for the general problem of data analysis, in the context of Big Data and Complex Systems, has been advanced. By probing the data space encoded as a set of correlation functions, the information content of a phenomenological setting is embedded into a *field theory of data* based on an underlying topological space. This idea is deeply rooted into concepts that have originated from theoretical physics. General Relativity, to mention one of the examples, is the gravitational field theory that describes the motion of particles through space-time where their dynamics is fully determined by the underlying curvature.

*Corresponding author

Email addresses: diego.alberici2@unibo.it (Diego Alberici), pierluigi.contucci@unibo.it (Pierluigi Contucci), emanuele.mingione2@unibo.it (Emanuele Mingione), marco.molari3@studio.unibo.it (Marco Molari)

[1] Present adress: Laboratoire de Physique Statistique, École Normale Supérieure, PSL Research and CNRS UMR8550, Sorbonne Universités UPMC, 24 rue Lhomond, 75005 Paris, France

November 28, 2016

We propose here a very simplified realisation of that program that capitalises on the equivalence of field theories with classical statistical mechanics [2, 3, 4] with the purpose of testing it using the inverse problem approach. The models we consider are hard-core interacting polymer systems on high-dimensional networks (hypergraphs). The choice of this class of models is due to the diversity and richness of the phenomena they describe that span from Physics [5], Biology [6], Computer Science [7, 8, 9], and Social Sciences [10]. We have in mind, in particular, applications in the the socio-technical setting of novel communication systems where groups of people are present in chambers like those of the messaging systems, voip conference calls etc. From a mathematical point of view those are aggregation models of particles that cannot occupy at the same time more than one state (hard-core constraint): in the specific example of the messaging systems an individual is either silent, the monomer state, in a two body conversation, the dimer state, in a three body conversation state called trimer and so on. While the old style phone calls were well described by a standard monomer-dimer model the novel technologies allow for the contemporary presence of multiple individuals in the same virtual room thus requiring higher order objects like hypergraphs for the underlying space and polymers for the fields that represent their state.

In our model the configurations of the system are determined by the occupation number on the elements of the hypergraph (vertices, edges and faces) that takes only two values 0 and 1. We limit the analysis to the rank three case (conversation with maximum three bodies in the mentioned example) but the generalisation to higher ranks is straightforward. The model is assigned by a set of positive weights, the activities, associated to each hyperedge. These weights describe the strength of connections and identify the topology of the hypergraph trough, for instance, the persistent topology methods developed in [11, 12], in [13, 14] and used in [15]. A threshold for the activities could be decided, and the hyperedges below this threshold deleted from the original hypergraph. Instead of studying the topology at an arbitrary threshold, the persistent topology approach consists in exploring the whole filtration of hypergraphs obtained by varying the threshold. Quoting [1], "this filtration process identifies those topological features which persist over a significant parameter range, qualifying them as candidates to be considered as signal, while those that have short-lived features can be assumed to characterize noise". Afterwards this topological signal can be used to compare and classify different datasets.

Our first result is of rigorous mathematical nature: the identification of an iterative relation for the partition function of the model which generalises the Heilmann-Lieb identity [16]. While this relation is introduced in a hypergraph theoretical setting we show that it implies a set of identities directly expressible in terms of the correlation functions of the associated probability measure. They act as a constitutive family of equations for the model that we use in our test and turn out to be an essential tool toward an efficient control of the inverse problem, i.e. the basic question: from a (full or partial) set of the correlation functions can we recover the value of the activities for all the hyperedges?

This work provides a positive answer to the previous question together with

the possible limitations and contains two conceptually different numerical methods which can be used to extract activities from the experimental correlations. The first inversion method is based on the maximisation of the *likelihood* function and works through a recursive gradient-descent algorithm partially inspired by the one used for the learning process in Boltzmann Machines [17]. We tested its performance and found that it converges exponentially at a speed that does not depend on the size of the hypergraph but is influenced by the magnitude of the activities. In particular the convergence speed decreases at higher values of the activities, as expected when reaching the full packing regime. The second method is based on the maximisation of the *pseudo-likelihood* function when additional experimental correlations are known. This has the advantage that it can be applied in a much simpler manner since it provides an explicit expression for the activities.

Finally we study the effects of the presence of a further interaction acting among monomers in the hypergraph. In socio-technical systems this kind of interaction generated by peer-to-peer effects is often very relevant. The extra structure that comes with it is codified by another hypergraph built on the same set of vertices which, in general, is different and independent from the previous one. The two networks indeed can be seen as a bilayer structure like those analysed in [18]. We concentrated on the problem of probing the presence of such an interaction from the set of experimental correlations, and found that the comparison between the two previously introduced inversion methods provides a good test for the detection of the interaction. Moreover, in the high interaction limit, we show how the same comparison can also be used to numerically estimate the parameter magnitude.

2. The theoretical framework

Let $H = V \cup K$ be a *hypergraph* of rank 3, that is a set of vertices V and hyperedges K where $K = E \cup F$ is an union of edges E and faces F (our notation naturally generalises to arbitrary rank). On this topological space we consider configurations of *polymers*, precisely monomers (single particles occupying a vertex), dimers (2-particles occupying an edge), trimers (3-particles occupying a face). Polymers display mutual hard-core interaction: no region of the space can be touched by more than one polymer. At the same time we require all the vertices of the hypergraph to be covered by either a monomer or one of the vertices of a polymer. This last condition that we call *filling*, fully specifies the ensemble and should not be confused with the *full-packing* one where monomers are not allowed.

A suitable way to represent the allowed configurations is to introduce the *occupancy variables* $\alpha = \left(\alpha_h\right)_{h\in H} \in \{0,1\}^H$ with the *hard-core filling* condition

$$\alpha_v + \sum_{\substack{e\in E:\\ e\ni v}} \alpha_e + \sum_{\substack{f\in F:\\ f\ni v}} \alpha_f = 1, \quad v \in V. \tag{1}$$

Notice that because of (1), for any vertex $v \in V$ the quantity α_v, that represents the monomer occupancy of the vertex v, can always be expressed as a function of the dimer and trimer occupancy variables. It is convenient to introduce the admissibility characteristic function $C : \{0,1\}^H \to \{0,1\}$ defined as

$$C(\alpha) = \begin{cases} 1 & \text{if (1) holds} \\ 0 & \text{otherwise} \end{cases} . \tag{2}$$

To fully specify the model we introduce the *polymer activity* of each hyperedge, that is a positive number that measures the propensity of the hyperedge to be occupied by a corresponding polymer. One can show with an elementary computation that the vertex activities can be reabsorbed into the remaining parameters or factorised out of the partition function. We denote by $z_e,\ e \in E$ the *edge activities* (or *dimer activities*) and by $z_f,\ f \in F$ the *face activities* (or *trimer activities*). The topological and analytical data, namely H and z, fully determine a *probability measure* associated to configurations:

$$\mu_z(\alpha) \ = \ \frac{C(\alpha) \prod_{e \in E} z_e^{\alpha_e} \prod_{f \in F} z_f^{\alpha_f}}{Z_H(z)} \ , \quad \alpha \in \{0,1\}^H \tag{3}$$

where Z_H is the normalisation factor usually called *partition function*:

$$Z_H(z) \ = \ \sum_{\alpha \in \{0,1\}^H} C(\alpha) \prod_{e \in E} z_e^{\alpha_e} \prod_{f \in F} z_f^{\alpha_f} \ . \tag{4}$$

We denote by $\langle \cdot \rangle$ the average with respect to the probability measure (3).

Defining $E(v)$ the set of edges with one vertex in v and $F(v)$ the set of faces with one vertex in v, one can prove that the following iterative relation holds:

$$Z_H \ = \ Z_{H-v} + \sum_{e \in E(v)} z_e \, Z_{H-e} + \sum_{f \in F(v)} z_f \, Z_{H-f} \tag{5}$$

which generalises the Heilmann-Lieb relation for monomer-dimer systems [16, 19]. In equation (5), $H - v$ denotes the hypergraph where the vertex v has been removed together with the hyperedges in $E(v) \cup F(v)$; $H - e$ stands for $H - v_1 - v_2$ where $e = \{v_1, v_2\}$; $H - f$ stands for $H - v_1 - v_2 - v_3$ where $f = \{v_1, v_2, v_3\}$. Equation (5) can be easily proven by splitting the sum (4) according to the three possible states of a given vertex v, indeed the three terms on the right hand side of (5) correspond to sum over the configurations α where v is occupied by a monomer ($\alpha_v = 1$), a dimer ($\alpha_e = 1$ for some $e \in E(v)$) or a trimer ($\alpha_f = 1$ for some $f \in F(v)$) respectively.
The previous family of relations (5) for the partition function of the model implies the following *topological constraint relations* for the correlation functions. For every edge $e = \{i, j\}$ and for every observable g that does not depend on α_e, α_i and α_j it holds:

$$\langle \alpha_e \, g \rangle \ = \ z_e \, \langle \alpha_i \alpha_j \, g \rangle \ . \tag{6}$$

Similarly, for every face $f = \{i, j, l\}$ and for every observable g that does not depend on α_f , α_i, α_j and α_l it holds:

$$\langle \alpha_f \, g \rangle \; = \; z_f \, \langle \alpha_i \alpha_j \alpha_l \, g \rangle \; . \tag{7}$$

In particular for $g \equiv 1$ one obtains an explicit expression of the activities in terms of correlations

$$z_e \; = \; \frac{\langle \alpha_e \rangle}{\langle \alpha_i \alpha_j \rangle} \quad , \quad z_f \; = \; \frac{\langle \alpha_f \rangle}{\langle \alpha_i \alpha_j \alpha_l \rangle} \; . \tag{8}$$

Equations (6) and (7) can be easily proven by observing that the admissible configurations α such that $\alpha_e = 1$ are in one-to-one correspondence with the admissible configurations α such that $\alpha_i = \alpha_j = 1$.

3. The inverse problem

In the last few years several new ideas and techniques have been developed [20, 21, 22] for the *inverse problem* of the Ising model. We will discuss the inverse problem for the class of hard-core polymer models introduced in the previous section. The general task is to extract the parameters of a given theoretical model from experimental measures on the observables. The problem clearly displays different features according to the types of data that become available. In this work we will focus on two experimental database settings. In the first one the dataset is composed by the empirical densities of dimers and trimers, while in the second one some empirical correlations for the monomers are also included:

A) the *empirical polymer densities*, that is $\langle \alpha_e \rangle_{\rm exp}$ for very edge $e \in E$ and $\langle \alpha_f \rangle_{\rm exp}$ for every face $f \in F$;

B) the previous *empirical polymer densities* plus the *empirical monomer correlations*, that is $\langle \alpha_i \alpha_j \rangle_{\rm exp}$ for every edge $e = \{i, j\} \in E$ and $\langle \alpha_i \alpha_j \alpha_l \rangle_{\rm exp}$ for every face $f = \{i, j, l\} \in F$.

The symbol $\langle \; \rangle_{\rm exp}$ denotes the empirical average, that is if M polymer configurations $\alpha^{(1)}, \ldots, \alpha^{(M)}$ are observed independently then $\langle g \rangle_{\rm exp} \equiv \frac{1}{M} \sum_{s=1}^{M} g(\alpha^{(s)})$.

3.1. Maximum likelihood and maximum pseudo-likelihood approximations

We start by shortly recalling the application of the Maximum Likelihood and the Maximum Pseudo-Likelihood Methods to our model. These methods are used in the solution of the inverse problem, since they provide a rule to choose the values of the model parameters that are best able to reproduce some empirically observed features in the data. The general framework is the following: fix the hypergraph H and assume the model is described by an unknown value of the activities z to be determined. Consider a set of M observations of polymer configurations $\bar{\alpha} = \{\alpha^{(s)}\}_{s=1,\ldots,M}$, where $\alpha^{(s)} = (\alpha_k^{(s)})_{k \in K}$ and $\alpha_k^{(s)}$ encodes

the presence/absence of a polymer on the hyperedge k in the $s^{\rm th}$ experimental observation. Suppose that $\bar{\alpha}$ is a set of independent observations sampled from the same probability distribution μ_z, for a certain value of the activities $z = z^*$.

We use two standard methods that give an optimal value z^* to fit the dataset $\bar{\alpha}$: the *maximum likelihood estimation* (MLE) and the *maximum pseudo-likelihood estimation* (MPLE). Let us briefly recall these methods.

The optimal estimate z^* in the MLE sense maximizes the *likelihood function* defined as

$$\mathcal{L}(z;\bar{\alpha}) \,=\, \prod_{s=1}^{M} \mu_z(\alpha^{(s)}) \,. \tag{9}$$

Standard computations show that $\log \mathcal{L}(z;\bar{\alpha})$ is a concave function in the variables $\log z$ and it attains its maximum at the point z^* satisfying the following system of $|K|$ equations:

$$\langle \alpha_k \rangle_{z^*} = \langle \alpha_k \rangle_{\rm exp} \,, \quad k \in K = E \cup F \,, \tag{10}$$

where as before $\langle \alpha_k \rangle_{\rm exp} \equiv \frac{1}{M} \sum_{s=1}^{M} \alpha_k^{(s)}$ is the experimental average value of the presence of a polymer in the hyperedge k. This approach naturally fits the experimental situation of case A), where the available data is the set of *empirical polymer densities*. Let us observe that the likelihood function $\mathcal{L}(z;\bar{\alpha})$ is strictly related to the Kullback-Leibler divergence of the measure μ_z from the empirical measure μ^*, defined as

$$D_{\rm KL}(\mu_z|\mu^*) \,=\, \sum_{\alpha} \mu^*(\alpha) \, \log \frac{\mu^*(\alpha)}{\mu_z(\alpha)} \tag{11}$$

where $\mu^*(\alpha) \equiv \frac{1}{M} \sum_{s=1}^{M} \delta(\alpha = \alpha^{(s)})$. Precisely the following relation holds:

$$\frac{1}{M} \log \mathcal{L}(z;\bar{\alpha}) \,=\, -D_{\rm KL}(\mu_z|\mu^*) \,+\, C \tag{12}$$

with $C = \sum_{\alpha} \mu^*(\alpha) \log \mu^*(\alpha)$.

Now let us consider the pseudo-likelihood instead of the likelihood. The optimal estimate z^* in the MPLE sense maximizes the *pseudo-likelihood function* defined as

$$\mathcal{L}^P(z;\bar{\alpha}) \,=\, \prod_{s=1}^{M} \prod_{k \in K} \mu_z\big(\alpha_k^{(s)} \big| \alpha_{\neq k}^{(s)}\big) \tag{13}$$

where, for a given sample s and hyperedge k, $\alpha_{\neq k}^{(s)}$ encodes the experimental observation of a polymer on all the hyperedges different from k. It is possible to show that $\mathcal{L}^P$ attains its maximum at the point z^{**} explicitly defined by the following $|K|$ conditions:

$$\langle \alpha_k \rangle_{\rm exp} \,=\, z_k^{**} \, \big\langle \prod_{v \in k} \alpha_v \big\rangle_{\rm exp} \,, \quad k \in K = E \cup F \tag{14}$$

where $\alpha_v^{(s)}$ denotes the experimental observations of a monomer on the vertex v in the sth trial and $\langle \prod_{v\in k} \alpha_v \rangle_{\exp} \equiv \frac{1}{M}\sum_{s=1}^{M} \prod_{v\in k} \alpha_v^{(s)}$ is the empirical monomer correlation of the vertices in k. This time the set of equation naturally fits case B), where the set of empirical monomer correlations is known.

3.2. The Kullback-Leibler method

In case A) the *Maximum Likelihood Estimation* (MLE) can be used. Let us denote by μ_z and by $\langle\ \rangle_z$ respectively the probability measure defined by (3) and the associated expectation. As we previously proved, the MLE of the polymer activities $z^* = (z_k^*)_{k\in K}$ satisfies the following set of $|K|$ conditions

$$\begin{aligned} \langle \alpha_e \rangle_{z^*} &= \langle \alpha_e \rangle_{\exp}\,, \quad e \in E \\ \langle \alpha_f \rangle_{z^*} &= \langle \alpha_f \rangle_{\exp}\,, \quad f \in F\,. \end{aligned} \tag{15}$$

The set of equations (15) determines implicitly the activities. We approach its solution by means of a gradient descent algorithm since the Maximum Likelihood function is a concave function. Precisely at step $n+1$ ($n \geq 0$) we update the vector of polymer activities $z^{(n)} \equiv \left(z_k^{(n)}\right)_{k\in K}$ as follows

$$z^{(n+1)} = z^{(n)} - \eta^{(n+1)} \frac{\nabla(z^{(n)})}{\sqrt{\sum_{k\in K}\left(\partial_k(z^{(n)})\right)^2}}\,. \tag{16}$$

The vector $\nabla(z) \equiv \left(\partial_k(z)\right)_{k\in K}$ is the gradient of the Kullback-Leibler divergence $D_{KL}(\mu_z|\mu^*)$, defined by:

$$\partial_k(z) = -\frac{\langle \alpha_k \rangle_{\exp} - \langle \alpha_k \rangle_{\mathrm{z}}}{z_k} \tag{17}$$

and it gives to the update step $\Delta z^{(n+1)} \equiv z^{(n+1)} - z^{(n)}$ the direction of the gradient of the likelihood function, or equivalently the direction of minus the Kullback-Leibler divergence gradient (see eq. (12)). The positive number $\eta^{(n+1)}$ tunes the magnitude of the update steps $\Delta z^{(n+1)}$. By fixing $\eta^{(n)} \equiv \eta$, the speed of convergence of relation (16) is linear, while it can be improved by introducing an adaptive learning rate defined iteratively as:

$$\eta^{(n+1)} = \eta^{(n)} \exp\left\{ \gamma \frac{\sum_{k\in K} \Delta z_k^{(n)}\, \Delta z_k^{(n-1)}}{\sqrt{\sum_{k\in K}\left(\Delta z_k^{(n)}\right)^2}\sqrt{\sum_{k\in K}\left(\Delta z_k^{(n-1)}\right)^2}} \right\} \tag{18}$$

γ is a positive parameter to be chosen. The relation (18) is based on the scalar product between two consequent updates of the activities. If it is positive, which means that the last update steps $\Delta z_k^{(n)}$, $\Delta z_k^{(n-1)}$ were performed along similar directions, then the next update $\Delta z_k^{(n+1)}$ will have a greater magnitude. If it is negative, which means that the last two updates were performed along opposite

directions, then we are in proximity of the solution and a greater precision is needed, so the magnitude of the next update step is diminished.

The recursion stops when the value of the activities $z^{(n_f)}$ is sufficiently close to the exact MLE solution of the inverse problem z^*. In our case we used two different stopping criteria. The first one can be used only when testing the performance of the algorithm on a priori known models, since it requires the knowledge of the exact values of the activities. In this case a value of precision $\epsilon_f > 0$ is chosen, and the recursion stops when the maximum relative error over the set of activities is less than ϵ_f:

$$\epsilon^{(n_f)} = \max_{k \in K} \left| \frac{z_k^* - z_k^{(n_f)}}{z_k^*} \right| < \epsilon_f \ . \tag{19}$$

The second criterion can be applied when solving the inverse problem on experimental data, since it does not assume the knowledge of the exact value of the activities. Again a final precision value $\hat{\epsilon}_f > 0$ is chosen, and the recursion stops as soon as the set of equations (15) is satisfied with precision of at least $\hat{\epsilon}_f$:

$$\hat{\epsilon}^{(n_f)} = \max_{k \in K} |\log\langle\alpha_k\rangle_{z^{(n_f)}} - \log\langle\alpha_k\rangle_{\exp}| < \hat{\epsilon}_f \ . \tag{20}$$

In order to assess the reliability and stability of this method we performed numerical tests on the speed of convergence of the algorithm (16) to the solution of the equation (15) on random hypergraphs.

In particular we made use of a class of random hypergraph which represents the extension of the notion of Erdős-Rény random graph. This choice allows us to test the performance of our algorithm over different topologies. Moreover, real-world data is often constituted by many similar instances of the model, whose topologies can be considered as extracted from some random distribution (see [10] for instance).

We observed that the convergence of the algorithm is exponentially fast in the number of iterations n (Figure 1). Moreover the distribution of the speed of convergence does not seem to depend on the number of vertices N in the random hypergraph (Figure 2). Anyway we stress the fact that the larger N is, the longer it takes to compute each step of the algorithm, since the evaluation of $\langle\alpha_k\rangle_{z^{(n)}}$ is more demanding. On the contrary the speed of convergence depends on the intensity of the activities (Figure 3). In particular in the limit of large polymer activity the exponential rate of convergence vanishes. This limit is equivalent to the *full-packing* regime, in fact when polymer activities are high the presence of monomers is repressed in favour of higher order particles.

Precisely, to obtain these results, we have generated data as follows:

1. A random hypergraph $H = V \cup K$ over N vertices is generated by placing each hyperedge independently. Each 2-edge is present with probability $p_1 = 2c_1/(N-1)$ and each 3-edge with probability $p_2 = 6c_2/(N-1)(N-2)$.
2. An activity z_k is assigned to each hyperedge $k \in K$. For simplicity when generating the dataset we chose $z_k = z$ constant for all $k \in K$. Details of this choice are specified in each case.

3. All the possible monomer-dimer-trimer configurations $\alpha = (\alpha_k)_{k\in K}$ on the hypergraph are computed. We assign to each configuration its probability and we evaluate the expectations $\langle \alpha_k \rangle_z$.

The gradient descent algorithm was then applied, using as input parameters $\langle \alpha_k \rangle_{\rm exp} = \langle \alpha_k \rangle_z$. Clearly, this choice entails that z solves eq. (15) and the recursion converges to the value $z^* = z$. We set $z_k^{(0)} = 1$ for all $k \in K$ and $\gamma = 0.2$. We used eq. (19) as stopping criterion setting $\epsilon_f = 10^{-10}$.

To conclude this subsection we notice that, according to step 1), our random hypergraphs ensemble is defined by the probability distribution:

$$P(H) = {p_1}^{|E|_H} \, (1-p_1)^{\binom{N}{2}-|E|_H} \, {p_2}^{|F|_H} \, (1-p_2)^{\binom{N}{3}-|F|_H} \quad . \tag{21}$$

It is easy to show that this probability distribution can be derived from a maximum entropy principle. In fact this is the probability distribution over the set of hypergraphs with N vertices which maximises the entropy function $S(H) = -\sum_H P(H) \ln P(H)$ with a pair of "soft" constraint, namely a fixed average number of edges $\bar{e} = \sum_H P(H)|E|_H$ and average number of faces $\bar{f} = \sum_H P(H)|F|_H$. In particular the relation among the probabilities appearing in ([**?**])and those average quantities are given by:

$$p_1 = \frac{\bar{e}}{\binom{N}{2}}, \quad p_2 = \frac{\bar{f}}{\binom{N}{3}} \ , \tag{22}$$

from which it follows that the constants $c_1 = \bar{e}/N$, $c_2 = \bar{f}/N$ represent respectively the average number of edges and faces per node in the hypergraph. This random ensemble is the natural generalisation of the concept introduced in [23] of exponential random simplicial complex to the case of hypergraphs (see also [24, 25]).

3.3. The effects of an imitative perturbation

It is important to notice that in case B) the number of observables is two times the number of degrees of freedom of the model defined by (3), since the dataset contains the *empirical polymer densities* $\langle \alpha_k \rangle_{\rm exp}$ and the *empirical monomer correlations* $\langle \prod_{v\in k} \alpha_v \rangle_{\rm exp}$ while the model is determined only by the activities z_k, $k \in K$.

As we stated in section 3.1, a possible way to deal with this overdetermined case is to consider the *Maximum Pseudo-Likelihood Estimation* (MPLE). This method can be seen as an approximation of the MLE where the joint distribution is replaced with a suitable conditional probability: we look at the probability to observe an occupied hyperedge conditionally on the states of all the others. As we showed in eq. (14) the MPLE of the activities z^{**} satisfies the following set of $|K|$ conditions

$$\begin{aligned} \langle \alpha_e \rangle_{\rm exp} &= z_e^{**} \, \langle \alpha_i \alpha_j \rangle_{\rm exp} \, , \quad e = \{i,j\} \in E \\ \langle \alpha_f \rangle_{\rm exp} &= z_f^{**} \, \langle \alpha_i \alpha_j \alpha_l \rangle_{\rm exp} \, , \quad f = \{i,j,l\} \in F \ . \end{aligned} \tag{23}$$

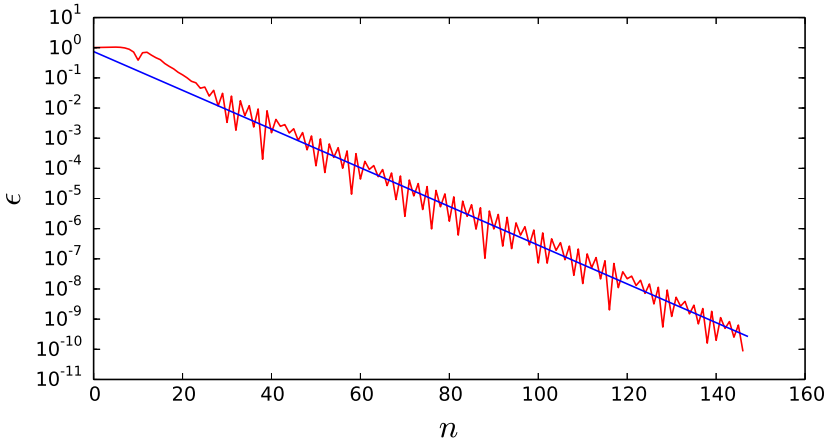

Figure 1: **Relative error of the gradient descent algorithm**: (Colour online) $\epsilon^{(n)} = |z_k^{(n)} - z_k|/z_k$ versus number of iterations n (red curve, linear-log scale). *The convergence is exponentially fast in the number of iterations*: to test this hypothesis we performed a linear fit (blue line) according to the relation $\log \epsilon^{(n)} = -A\,n + B$. We chose a random hypergraph with $N = 15$, $c_1 = c_2 = 1$ and $z_k = 0.5$ for all $k \in K$. The fit is performed on the data after removing the initial 20% of iterations.

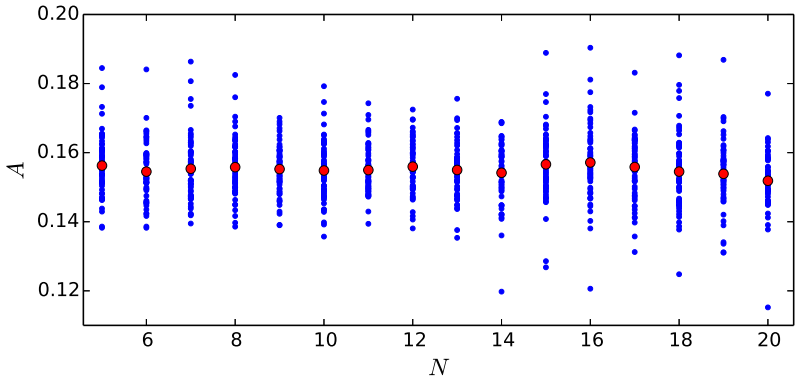

Figure 2: **Exponential rate of convergence A of the gradient descent algorithm versus number of vertices** N, according to the fit $\log \epsilon^{(n)} = -A\,n + B$. (Colour online) The number of vertices ranges from 5 to 20. *The distribution of the velocity of convergence does not seem to depend on the number of vertices. Anyway we stress the fact that the larger N is, the longer it takes to compute each step of the algorithm.* For each value of N we performed 60 trials (samples) on different random hypergraphs, taking always $c_1 = c_2 = 1$ and $z_k = 0.5$ for all $k \in K$. The red dots represent the mean values of A for each set of trials with the same value of N. To test the accuracy of the exponential fit we computed the correlation coefficient R: its average value and standard deviation over these 960 tests are $R = -0.945 \pm 0.023$.

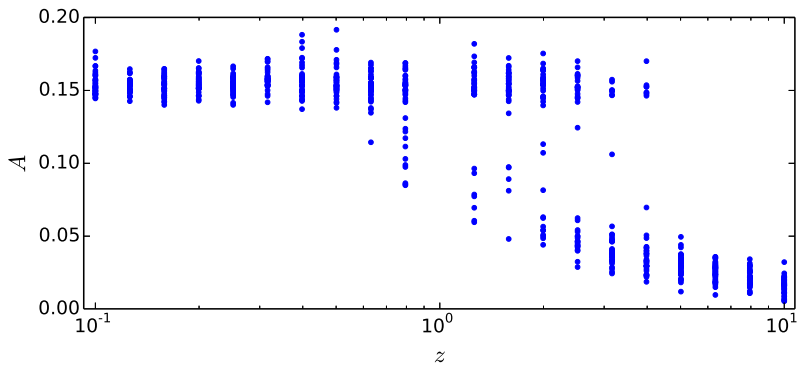

Figure 3: **Exponential rate of convergence A of the gradient descent algorithm versus polymer activity** z (log-linear scale), according to the fit $\log \epsilon^{(n)} = -A\,n + B$. The activity is the same for each hyperedge ($z_k = z\ \forall\, k \in K$) and takes values $z = 10^h$, $h = -1, -0.9, \ldots, 1$, excluding $h = 0$ which is by default the starting point of our algorithm. *The distribution of the rate of convergence depends on the intensity of the activity*: it is constant for $z \leq 10^{-0.2}$, then for $10^{-0.1} \leq z \leq 10^{0.6}$ it splits in two regions, and for $z \geq 10^{0.6}$ only the slower region survives and the rate of convergence decreases to zero. For each value of z we performed 40 trials on different random hypergraphs, taking always $N = 20$, $c_1 = c_2 = 1$. The underlying hypothesis of exponential convergence is supported by the correlation coefficient $R = -0.968 \pm 0.028$ over these 800 tests.

We observe two important features: the analogy between (23) and the exact relations (8) and the fact that these relations provide an explicit form for the activities.

Another way to exploit the additional information given by the empirical monomer correlations is to modify the model defined in (3) by introducing a new family of parameters $J = (J_k)_{k \in K}$ that tune the monomer correlations:

$$\mu_{z,J}(\alpha) \;=\; \frac{C(\alpha)\,\prod_{k\in K} z^{\alpha_k}\,\exp\Big(\sum_{k\in K} J_k \prod_{v\in k}\alpha_v\Big)}{Z_H(z,J)}\,, \quad \alpha \in \{0,1\}^H\,. \tag{24}$$

We denote by $\langle\cdot\rangle_{z,J}$ the average with respect to this probability measure. While this fact could appear as a mere technical device, it has instead a deep phenomenological meaning: the monomers can indeed directly interact beyond the hard-core repulsion, a situation largely expected in socio-technical systems due to the peer-to-peer effect among individuals. In other words in the experiments the presence of a coupling J between monomers cannot be excluded *a priori*. For this reason in this second part of our work we have generated the *empirical polymer densities* and *empirical monomer correlations* according to a perturbed distribution $\mu_{z,J}$.

The following extension of the Heilmann-Lieb identity for the partition function of the measure (24) holds:

$$Z_H \;=\; Z^*_{H-v} \;+\; \sum_{\substack{k\in K\\ k\ni v}} z_k\, Z_{H-k}\;, \quad v\in V \tag{25}$$

where in the partition function Z^*_{H-v} a monomer activity $e^{J_{u\sim v}} := \prod_{k\in K,\, k\ni u,v} e^{J_k}$ is introduced on every vertex u which was connected to v. We call *hypertree* a hypergraph H such that, after having removed the edges included in some face, its line graph is a tree. On hypertrees the relation (25) provides the following useful estimate:

$$\frac{\langle\alpha_k\rangle_{z,J}}{\langle\prod_{v\in k}\alpha_v\rangle_{z,J}} \;=\; \frac{z_k}{\prod_{\substack{h\in K,\\ |h\cap k|>0}} e^{J_h}}\,\theta_k\;, \quad k\in K \tag{26}$$

where the term θ_k goes to 1 as $z_p\, e^{-J_p}$ vanishes for every polymer $p\in K$ at distance 1 from k, and even better:

$$1 \;\le\; \theta_k \;\le\; \prod_{\substack{v\in V,\\ v\sim k}} \left(1 + \sum_{\substack{p\in K,\\ p\ni v,\, |p\cap k|=0}} z_p \prod_{\substack{q\in K,\\ q\ni v,\, |q\cap k|=0}} e^{-J_q}\right). \tag{27}$$

As said before, we have generated data $\langle\alpha_k\rangle_{\rm exp}$, $\langle\prod_{v\in k}\alpha_v\rangle_{\rm exp}$ according to the distribution (24) in the presence of an interaction $J\neq 0$: the quantities $\langle\alpha_k\rangle_{z,J}$ and $\langle\prod_{v\in k}\alpha_v\rangle_{z,J}$ have been computed exactly on random hypergraphs, following a procedure analogous to Section 3.2. Starting from these data we have computed the MLE and MPLE as if the interaction was not present. We guessed that while the two resulting estimates z^* and z^{**} of the activities agree in case $J=0$, they may differ when $J\neq 0$, and thus they may be used to probe the presence of an interaction. To make this guess more precise, we performed the following test, which could be applied also to real data.

- The gradient descent algorithm (16) is executed using as input $\langle\alpha_k\rangle_{\rm exp} = \langle\alpha_k\rangle_{z,J}$. If the algorithm converges, its limit is a vector of activities z^* such that:

  $$\langle\alpha_k\rangle_{z^*} \;=\; \langle\alpha_k\rangle_{z,J}\;, \quad k\in K\;. \tag{28}$$

  We set $z_k^{(0)} = 1$ and $\gamma = 0.2$. We used eq. (20) as stopping criterion setting $\hat\epsilon_f = 10^{-5}$, together with a bound for the number of iterations that stops the recursion at $n = 5000$ even if the precision $\hat\epsilon_f$ has not been reached yet.

- The closed inversion formula (23) is applied, as if the coupling potential was not present:

  $$z_k^{**} \;=\; \frac{\langle\alpha_k\rangle_{z,J}}{\langle\prod_{v\in k}\alpha_v\rangle_{z,J}}\;, \quad k\in K\;. \tag{29}$$

- We study the parameter

  $$\delta \;=\; \frac{1}{|K|}\sum_{k\in K}\left(\log z_k^{**} - \log z_k^*\right). \tag{30}$$

For zero coupling potential δ is close to zero, since both z_k^{**} and z_k^* equal the true value of the activity z_k (up to the precision of the gradient descent algorithm).

We observed that δ, together with the final precision $\hat{\epsilon}$, can indeed be used as a test-parameter to understand whether the real system obeys a pure hard-core interaction or there are other types of non-negligible interactions. In fact it allows to distinguish between the following three regimes (Fig. 4):

- For $J < 0$ the gradient descent algorithm is not guaranteed to converge in the prescribed number of iterations since the precision $\hat{\epsilon}$ ranges from 10^{-5} to 10^0. The value of δ is negative and its modulus grows linearly with J.
- For $0 < J < J_0$ the convergence of the gradient descent method is attained. The parameter δ is close to zero, positive, and shows a non-monotonic behaviour in J.
- For $J > J_0$ the convergence of the gradient descent method becomes abruptly poor and for J sufficiently large $\hat{\epsilon}$ is larger that 10^1. δ is positive and exhibits a large variance over different random hypergraphs.

When J is positive and sufficiently large, we propose a method to estimate its value. Compare the relations (26) for the measure $\mu_{z,J}$ with the exact relations (8) for the measure μ_z. It becomes clear that if the experimental parameter $\rho_k \equiv \log\left(\langle\alpha_k\rangle_{\rm exp}/\langle\prod_{v\in k}\alpha_v\rangle_{\rm exp}\right)$ shows a correlation with the number of hyperedges intersecting k, $\nu_k \equiv \mathrm{Card}\{h \in K,\, |h \cap k| > 0\}$, then the system presents other interactions beyond the hard-core one. In particular in the case of constant J and z, the equation (26) gives

$$\rho_k(z, J) \approx \log z - J\,\nu_k\,, \quad k \in K \tag{31}$$

when $J\,\mathrm{Card}\{q \in K,\, q \ni v,\, |q \cap k| = 0\}$ is sufficiently large with respect to $\log z_p$, for all hyperedges p intersecting k and all vertices v neighbouring k. Therefore J and z can be found by performing a linear fit between ρ_k and ν_k (Fig. 5).

4. Conclusions and Outlooks

With the purpose to investigate the possibility to discover topological information from the data space we introduced in this work a model in which polymers are deposited on the hyperedges of an hypergraph with a probability determined according to the hyperedges activities. The idea underlying the model is that simple graphs are no longer able to account for the structure of many modern socio-technical systems, such as those of virtual messaging systems or voip calls. In these systems the communications do not occur only between pairs of users, but may involve larger groups [26]. We believe that this context may give rise to new interesting behaviours, where topology plays a crucial role.

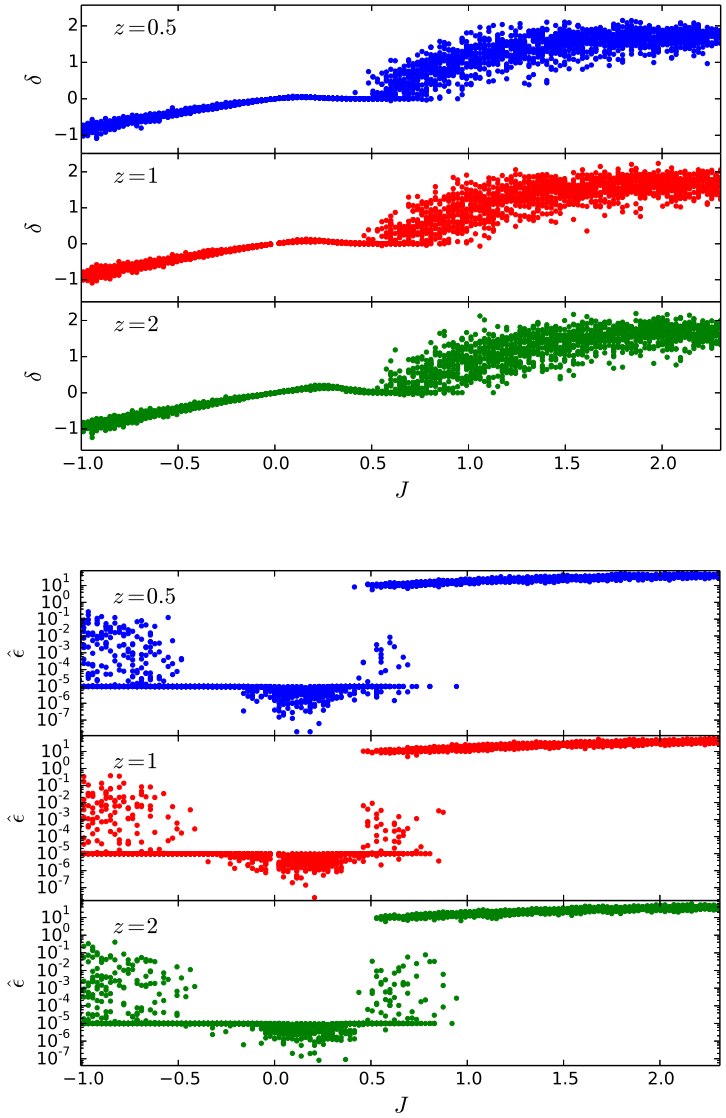

Figure 4: **Tests for the presence of imitative interaction.** (Colour online) *On top:* Parameter $\delta = \frac{1}{|K|}\sum_{k\in K}(\log z_k^{**} - \log z_k^{*})$ evaluated through the use of both the analytic inversion formula and the gradient descent method, as if the imitative interaction was not present, versus imitative potential J. *A value $\delta < 0$ reveals that $J < 0$. On the other hand, the order of magnitude of δ and its variance grow abruptly when J crosses a positive critical value. On bottom:* Precision $\hat{\epsilon} = \max_{k\in K} |\log\langle\alpha_k\rangle_{z^*} - \log\langle\alpha_k\rangle_{z,J}|$ of the gradient descent algorithm built as if the imitative interaction was not present, versus imitative potential J (linear-log scale). *The convergence is always reached for J close to* 0*, while it is never reached for J larger than a critical value.* The polymer activity and the imitative potential are the same for each hyper-edge: $z_k = z,\ J_k = J\ \forall\, k \in K$. For each value of J we performed 20 trials on different random hypergraphs, taking always $N = 20$, $c_1 = c_2 = 1$ and $z = 0.5$ (blue), $z = 1$ (red), $z = 2$ (green).

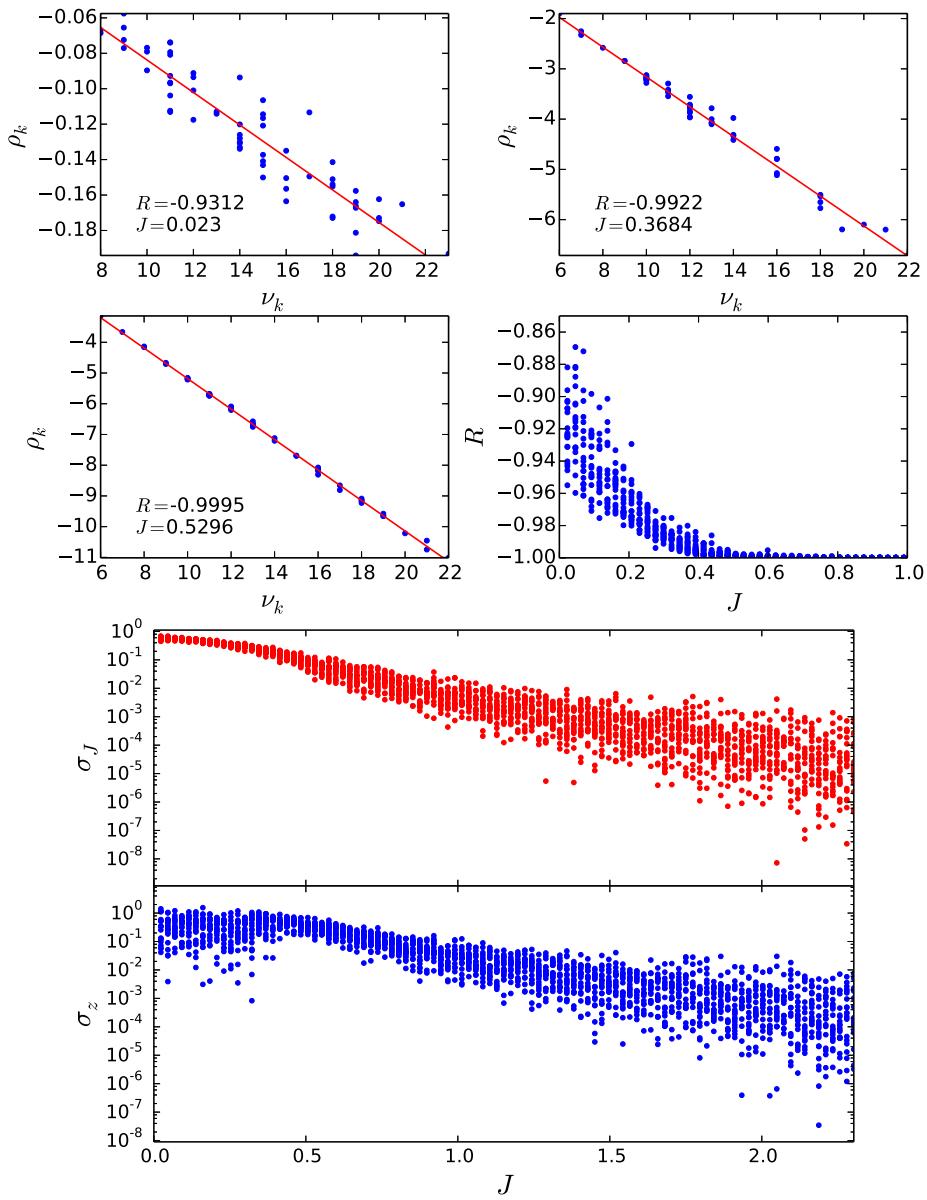

Figure 5: **Estimate of the imitative potential J.** (Colour online) *On top:* parameter $\rho_k = \log\langle\alpha_k\rangle / \langle\prod_{v\in k}\alpha_v\rangle$ versus $\nu_k = \mathrm{Card}\{h \in K \,|\, |h \cap k| > 0\}$ for every hyperedge k in a random hypergraph (blue dots). The polymer activity and the coupling are the same for each hyperedge: $z_k = z$, $J_k = J \;\forall\, k \in K$. The test is performed on a random hypergraph taking $N = 25$, $c_1 = c_2 = 1$, $z = 1$ and different values of J: $J = 0.023$, $J = 0.3684$, $J = 0.5296$. *The relation between ρ_k and ν_k is linear for J sufficiently large*: a linear fit (red line) is performed according to the relation $\rho_k = -\alpha\,\nu_k + \beta$. The reliability of this fit is tested by plotting the correlation coefficient R versus J. *On bottom:* relative errors $\sigma_J = \left|\frac{\alpha - J}{J}\right|$ (red) and $\sigma_z = \left|\frac{\beta - \log z}{\log z}\right|$ (blue), versus J. *According to the relation* (31)*, the slope of the fit α is used as an estimate of the coupling J, when J is sufficiently large.*

With these possible applications in mind we tackled the inverse problem. After finding an extension of the Heilmann-Lieb relations that fits the higher-dimensional case, we introduced the *Maximum Likelihood Estimation* (MLE) and the *Maximum Pseudo-Likelihood Estimation* (MPLE) solutions of the inverse problem. While the latter constitutes a more rough estimate but has an explicit form in terms of experimental quantities, the former provides a more precise but implicit solution, which can nonetheless be numerically evaluated by the gradient descent algorithm we proposed. We found that by introducing a variable update step size the algorithm converges with exponential precision in the number of steps. However we stress that the time it takes to compute each step of the algorithm grows with the size of the hypergraph, since all the admissible configurations have to be computed exactly. A possible solution to this problem could be to evaluate average quantities through Markov chain Monte Carlo sampling. We tested the algorithm on simple instances for different values of the parameters, and found that while the exponential convergence does not seem to be influenced by the number of vertices in the hypergraphs, it does depend on the values of the activities. A further analysis of this dependence could be performed, for example with respect to the variance of the activity distribution.

We then considered the presence of an interaction between the monomers in the configurations. The meaning of this interactions can be understood by thinking to the social-technical systems that our model describes where monomer interaction is the direct peer-to-peer effect that people share in real life, outside the chat room. We found that a comparison between the MLE and the MPLE solution of the inverse problem can be used to detect the presence of such an interaction. The same comparison can moreover lead to the estimation of the interaction magnitude in the "strong interaction" regime.

The next step and most natural continuation of this work would be the application of such a model on real-world data. By testing the model on data we could verify whether it is able to accurately describe the behaviour of users in virtual messaging services and what type of predictive ability it comes with. For instance, this could be done by measuring the Kullback-Leibler distance between the experimental probability distribution and the probability distribution resulting from the Maximum Likelihood Estimation. In case the model is accurate it would allow us to measure of user activities in chat rooms, and even determine whether the system is subject to peer-to-peer monomer interactions.

Aknowledgments The authors are deeply indebted to Mario Rasetti for inspiring this work and for many illuminating discussions. We also thank Massimo Ferri, Giovanni Petri, Federico Ricci-Tersenghi, Alina Sîrbu and Francesco Vaccarino for interesting discussions. This work was partially supported by FIRB (grant number RBFR10N90W), PRIN (grant number 2010HXAW77) and INdAM-GNFM (Progetto Giovani 2015).

References

[1] M. Rasetti, E. Merelli, Topological field theory of data: mining data beyond complex networks, in: Contucci, Giardina (Eds.), Advances in disordered systems, random processes and some applications, Cambridge University Press, 2016.

[2] R. P. Feynman, Space-time approach to non-relativistic quantum mechanics, Reviews of Modern Physics 20 (2) (1948) 367.

[3] J. Schwinger, On the euclidean structure of relativistic field theory, Proceedings of the National Academy of Sciences 44 (9) (1958) 956–965.

[4] K. Symanzik, Euclidean quantum field theory. i. equations for a scalar model, Journal of Mathematical Physics 7 (3) (1966) 510–525.

[5] T. Chang, Statistical theory of the adsorption of double molecules, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences (1939) 512–531.

[6] K. T. O'Neil, W. F. DeGrado, A thermodynamic scale for the helix-forming tendencies of the commonly occurring amino acids, Science 250 (4981) (1990) 646–651.

[7] C. Bordenave, M. Lelarge, J. Salez, Matchings on infinite graphs, Probability Theory and Related Fields 157 (1-2) (2013) 183–208.

[8] R. M. Karp, M. Sipser, Maximum matching in sparse random graphs, in: Foundations of Computer Science, 1981. SFCS'81. 22nd Annual Symposium on, IEEE, 1981, pp. 364–375.

[9] L. Zdeborová, M. Mézard, The number of matchings in random graphs, Journal of Statistical Mechanics: Theory and Experiment 2006 (05) (2006) P05003.

[10] A. Barra, P. Contucci, R. Sandell, C. Vernia, Integration indicators in immigration phenomena. a statistical mechanics perspective., Tech. rep. (2013).

[11] P. Frosini, Measuring shapes by size functions, in: Intelligent Robots and Computer Vision X: Algorithms and Techniques, International Society for Optics and Photonics, 1992, pp. 122–133.

[12] A. Verri, C. Uras, P. Frosini, M. Ferri, On the use of size functions for shape analysis, Biological cybernetics 70 (2) (1993) 99–107.

[13] H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification, in: Proc. 41st IEEE Symp. Found. Comput. Sci., 2000, pp. 454–463.

[14] H. Edelsbrunner, J. Harer, Persistent homology-a survey, Contemporary mathematics 453 (2008) 257–282.

[15] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. Hellyer, F. Vaccarino, Homological scaffolds of brain functional networks, Journal of The Royal Society Interface 11 (101) (2014) 20140873.

[16] O. J. Heilmann, E. H. Lieb, Theory of monomer-dimer systems, in: Statistical Mechanics, Springer, 1972, pp. 45–87.

[17] D. H. Ackley, G. E. Hinton, T. J. Sejnowski, A learning algorithm for boltzmann machines, Cognitive science 9 (1) (1985) 147–169.

[18] G. Bianconi, Statistical mechanics of multiplex networks: Entropy and overlap, Phys. Rev. E 87 (2013) 062806.

[19] O. J. Heilmann, E. H. Lieb, Monomers and dimers, Phys. Rev. Lett. (24) (1970) 1412–1414.

[20] E. Aurell, M. Ekeberg, Inverse ising inference using all the data, Physical review letters 108 (9) (2012) 090201.

[21] V. Sessak, R. Monasson, Small-correlation expansions for the inverse ising problem, Journal of Physics A: Mathematical and Theoretical 42 (5) (2009) 055001.

[22] Y. Roudi, J. Tyrcha, J. Hertz, Ising model for neural data: model quality and approximate methods for extracting functional connectivity, Physical Review E 79 (5) (2009) 051915.

[23] K. Zuev, O. Eisenberg, D. Krioukov, Exponential random simplicial complexes, Journal of Physics A: Mathematical and Theoretical 48 (46) (2015) 465002.

[24] A. Costa, M. Farber, Large random simplicial complexes, i, Journal of Topology and Analysis (2015) 1–31.

[25] M. Kahle, Topology of random simplicial complexes: a survey, AMS Contemp. Math 620 (2014) 201–222.

[26] O. T. Courtney, G. Bianconi, Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes, Phys. Rev. E 93 (2016) 062311.