Algebraic characteristics and satisfiability threshold of random Boolean equations

Binghui Guo, Wei Wei,[*] Yifan Sun, and Zhiming Zheng

LMIB and School of Mathematics and Systems Science, Beihang University, 100191 Beijing, China

(Received 29 August 2009; revised manuscript received 18 January 2010; published 23 March 2010)

The satisfiability of a class of random Boolean equations named massive algebraic system septated to linear and nonlinear subproblems is studied in this paper. On one hand, the correlation between the magnetization of generators and the clustering of solutions of the linear subproblem is investigated by analyzing the Gaussian elimination process. On the other hand, the characteristics of maximal elements of solutions of the nonlinear subproblem are studied by introducing the partial order among solutions. Based on the algebraic characteristics of these two subproblems, the upper and lower bounds of satisfiability threshold of massive algebraic system are obtained by unit-clause propagation and leaf-removal process, and coincide as the ratio of nonlinear equations $q > 0.739$ in which analytical values of the satisfiability threshold can be derived. Furthermore, a complete algorithm with heuristic decimation is proposed to observe the approximation of the satisfiability threshold, which performs more efficiently than the classical ones.

DOI: 10.1103/PhysRevE.81.031122 PACS number(s): 05.70.Fh, 89.20.Ff, 75.10.Nr, 02.70.−c

I. INTRODUCTION

As basic theoretical models of the computational complexity literature in computer science, random constraint satisfaction problems (CSPs) are studied in a wide range of theoretical and applicable areas [1–3]. Over the past twenty years, much attention has been paid to study the phase transition phenomena of some hard NP-complete CSPs, which cannot be determined whether they are satisfiable or not in polynomial time in the worst case by any known algorithm, e.g., satisfiability problem [4,5], number partitioning [6], and graph coloring [7]. It is considered to be an open problem in computer science to find the rigorous description of the satisfiability threshold, which is important to understand the phase transition intrinsically for a class of CSPs.

In recent years, how to identify the exact values of the thresholds of satisfiability and depict the complexity evolution of CSPs as the constraint density increases has been extensively investigated in mathematics, computer science, and statistical physics [8–10]. When equation density is close to the satisfiability threshold, local searching algorithms may face a sparse solution space with complicated organization. This brings great obstruction for searching procedures, which leads to high computational complexity. There are also some mathematical estimations of boundaries on satisfiability threshold for some random CSPs. Based on the technique of finite-size scaling window, the phase transition behavior is proved to be continuous with an order parameter critical exponent for 2-SAT problem [11,12]. Using the first-moment and second-moment methods, the upper and lower bounds for the sharp SAT/UNSAT threshold of random satisfiability problem have been studied [13,14].

Some characteristics of solution space, e.g., clustering and freezing of solutions, are considered as the essential hardness for algorithms. On intuitive grounds, the clustering phenomenon is considered to be in an intermediate SAT phase and responsible for blocking many local searching algorithms. Techniques from statistical physics of glassy systems have been introduced to investigate these characteristics of organization of solutions [15–17]. For random k-SAT problem, the analytical value of satisfiability threshold and the clustering phenomenon in solution space are investigated by the mean-field cavity method [18]. Specifically, survey propagation algorithm has been proposed, which can find the solutions of random formulas in the satisfiable regime very successfully in the case of enormous number of variables [19]. For another problem, XORSAT, the satisfiability threshold and the clustering phenomenon have been studied rigorously by cavity method in the viewpoint of the geometrical organization of solution space [20–22].

In this paper, the correlation between the algebraic properties and the organization of solutions of XORSAT problem, the self-averageness of the number of solutions and the magnetization of generators with clustering, are studied by analyzing the process of Gaussian elimination. Furthermore, the statistical characteristics of maximal elements of solutions of a class of nonlinear Boolean equations are studied to achieve a general landscape of the evolution of the solution space in the viewpoint of set theory. As a combination of linear and nonlinear Boolean equations, a model massive algebraic system (MAS) is proved to be NP-complete. By analyzing the unit-clause propagation and leaf-removal algorithm, the upper bounds and lower bounds of the satisfiability threshold with different ratios of nonlinear equations are obtained. The lower and upper bounds coincide as the ratio of nonlinear equations q is larger than 0.739. Based on the study on the algebraic properties of subproblems of MAS, we propose a complete algorithm whose performance is more effective than that of classical ones, and approximate the location of the satisfiability threshold by this algorithm. The statistical algebraic study of MAS provides rigorous ways to reveal the complicated organizations of the solutions, to locate the satisfiability threshold exactly and to design effective algorithms relying on mathematical properties of the problems.

[*] Corresponding author; weiw@buaa.edu.cn

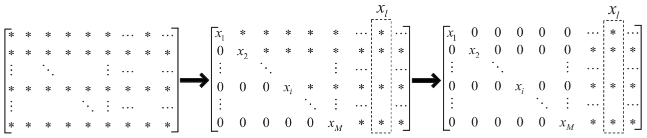

FIG. 1. The evolution of the coefficient matrix in the procedure of the Gaussian and reversed elimination for XORSAT problem. The symbol $*$ represents 1 or 0, which suggests whether the variable is in the corresponding equation or not.

II. GENERATORS AND MAGNETIZATION OF XORSAT PROBLEM

As a variant of k-satisfiability problem, the XORSAT problem is important for decoding of the low density parity checks of practical interest [23]. The detailed characterization of the phase diagram of this problem, including the location of threshold and clustering phenomenon, has been studied in [20] by both cavity method and leaf-removal analysis. Specially, the self-averageness of the percolation core after leaf-removal is proved by the moment method. Further, the x *-satisfiability* threshold of the XORSAT problem is investigated to measure the distances among solutions in different clusters [21]. In this section, we will provide some understanding to the self-averageness of the organization of the solutions and the formation of clusters in the viewpoint of algebra.

An XORSAT formula is defined by a string of Boolean variables and equations with the operation of Boolean addition, which can be solved in polynomial time by Gaussian elimination [24–26]. In this paper, we specialize to the instances of random k-XORSAT problem. A k-XORSAT instance is defined as: considering a set of N variables $x_1,x_2,\ldots,x_N$, we choose $M=\alpha N$ equations randomly and independently that involve k variables with the operation of Boolean addition

$$\sum_{i=1}^{k} x_{a_i}=J_a(\mathrm{mod}\ 2), \quad \text{for all } a=1,2,\ldots,M, \tag{1}$$

where $a_i\in\{1,2,\ldots,N\}$ and $J_a\in\{0,1\}$.

A. Self-averageness of the number of generators and solutions

For satisfiable instances of k-XORSAT problem, all the equations are linearly independent with probability one, otherwise it is unsatisfiable with some positive probability which is larger than $\frac{1}{2}$. To solve the problem, an important alternative method is Gaussian elimination. In the satisfiable phase of XORSAT problem, Gaussian elimination proceeds until the coefficient matrix of these linear equations is transformed to be an upper triangular matrix. This global algorithm will take at most $O(N^3)$ steps in the worst case and may take $O(N^2)$ steps when the coefficient matrix is sparse [24]. Furthermore, to obtain the mechanism of the organization of solutions, a reversed elimination is added after the Gaussian elimination procedure which is shown in Fig. 1.

Without loss of generality, assume that the diagonal elements in the resulted matrix after the eliminations are $x_1,x_2,\ldots,x_M$, respectively. Then, each of the diagonal elements $x_1,x_2,\ldots,x_M$ only appears in one single row of the resulted matrix. It is easy to verify that any assignment of $x_{M+1},x_{M+2},\ldots,x_N$ uniquely fixes the assignment of the other variables $x_1,x_2\ldots,x_M$ to achieve a single solution. Since all the variables $x_{M+1},x_{M+2},\ldots,x_N$ are free to take values in $\{0,1\}$, there are $N-M$ degrees of freedom for the variables in the solutions.

Considering an instance of k-XORSAT $\{\sum_{i=1}^{k}x_{a_i}=J_a\}_{a=1,2,\ldots,M}$ with solution space S, by the knowledge of algebra, the solution space S' of a homogeneous instance $\{\sum_{i=1}^{k}x_{a_i}=0\}_{a=1,2,\ldots,M}$ is isomorphic to S. Then, it is easy to verify that S' is a group of solutions by the operation of Boolean addition. By algebraic approach, there are $N-M$ generators of all the solutions of the homogeneous k-XORSAT problem, which can be in the form of

$$\{(\ldots,x_{M+1}=0,\ \ldots,x_{i-1}=0,x_i=1,x_{i+1}=0,\ \ldots,x_N=0), \\ \times M+1\leq i\leq N\}. \tag{2}$$

For that arbitrary assignment on $x_{M+1},\ldots,x_N$ can fix only one solution, there are 2^{N-M} solutions for any random instance with high probability. This result has a clear algebraic description of the formation and configuration of the solution space of k-XORSAT problem and verifies the self-averageness of the number of solutions in the satisfiable phase.

B. Magnetization of the generators with the clustering

In recent years, the space of solutions breaking into many disconnected clusters is studied by statistical mechanics [15,18]. Roughly speaking, solutions in different clusters have Hamming distance $O(N)$ and cannot reach each other by finite flipping. For the existence of this geometrical organization of the solutions, the invalidity of searching algorithms is ascribed to the extensive distances and sparseness among the solutions [27]. Long-range correlation among the variables with distance at least $O(\log N)$ on the factor graph is considered as the origin of clustering phenomenon [28], i.e., cycles on the factor graph make the variables correlated. For random 3-XORSAT, a percolation core with long-range correlation exists when $0.818469<\alpha<0.917935$, in which flipping the assignment of any variable forces infinite other variables changing assignments to keep the satisfiability and the solution space splits into many clusters [20].

To gain further understanding of the clustering phenomenon by algebra, we precisely focus on the statistical characteristics of the magnetization of generators which plays core status in the group and affects the organization of the solutions. Define the magnetization function on the assignments as

$$m(s^a)=m(x_1^a,\ \ldots,x_N^a)=\sum_{i=1}^{N}x_i^a, \tag{3}$$

which counts the number of 1*s* of a configuration s^a, $a\in\{1,2,\ldots,2^N\}$.

Since the formation of a set of generators that can generate all the solutions is not unique, it is meaningful to find the simplest formation with the lowest magnetization. If all such generators with the lowest magnetization have finite magnetization, all the solutions can reach each other by flipping finite values of variables and are in the same cluster. Considering a set of generators $\{g^1,g^2,\ldots,g^{M-N}\}$ with finite magnetization, any solution s can be written as the Boolean addition of several generators, e.g., $s=g^1+g^2+\ldots+g^k$. To view the distance between generator g^1 and solution s, we construct a generating chain as $\{g^1,g^1+g^2,\ldots,g^1+g^2+\ldots+g^{k-1},s\}$. Herein, the distance between any adjacent elements in such generating chain is finite. By this chain effect, we obtain that the solution s must be in the same cluster as the generator g^1. Thus, all the solutions generated by $\{g^1,g^2,\ldots,g^{M-N}\}$ must be in one single cluster.

We can obtain the corresponding correlation between the magnetization of the generators in simplest formation and the percolation core after leaf-removal process by the property of locked occupation problem [27,29] and the leaf-removal analysis [20]. For XORSAT problem, when the percolation core does not exist, one only need to flip finite variables to go from one solution to another, which can be derived from the literature of reconstruction on trees [29]. In this case, for the trivial solution $\{x_i=0,i=1,\ldots,N\}$ and some variable x_i, assigning value 1 to x_i and considering its influence propagation as in [20,30,31], another solution with only finite variables (including x_i) taking 1s can be obtained which with lowest magnetization is chosen to be one of the generators in the simplest formation. When the percolation core exists, one must flip at least a closed loop of variables, and there must exist at least one generator with magnetization $O[\log(N)]$ even $O(N)$ in any formation of the generators. In order to calculate the ratio t of variables in the percolation core, we quote the expression in [20]

$$1-t=(1-e^{-3\alpha(1-t)})^2, \tag{4}$$

in which α is the equation density. When $\alpha<0.818469$, Eq. (4) only has two trivial solutions $t=0$ and $t=1$, the generators with the simplest formation have finite magnetization, and then all the solutions are in one single cluster. When $\alpha>0.818469$, Eq. (4) has some nontrivial solution, there exists some generators in the simplest formation with infinite magnetization, which indicates that the solution space splits into clusters.

At the ending of this section, we discuss the magnetization of generators for different degree distributions of the factor graph of XORSAT problem [20]. For the case of Poisson distribution, the existence of generators with infinite magnetization fits well with the results gained from both Eq. (4) and leaf-removal. But for the case of power-law distribution, similar deduction as Eq. (4) is not in effect and the existence of generator with infinite magnetization only corresponds to the existence of percolation core after leaf-removal procedure. Numerical results of leaf-removal algorithm and Eq. (4) of XORSAT with different degree distributions are shown in Fig. 2.

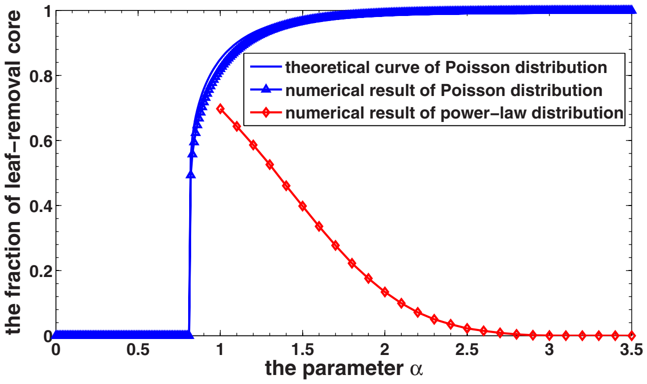

FIG. 2. (Color online) Numerical results of leaf-removal algorithm and theoretical curves of Eq. (4) of 3-XORSAT problem. The blue triangles represent the ratio of variables in the percolation core after leaf-removal procedure in the Poisson distribution case by numerical simulation with $N=10^5$, where the parameter α represents the equation density. The red triangles represent the ratio of variables in the percolation core after leaf-removal procedure in the power-law case by numerical simulation with $N=10^4$, where the parameter α represents the power exponent.

III. PARTIAL ORDER RELATION AND MAGNETIZATION IN NONLINEAR BOOLEAN EQUATIONS

The XORSAT problem only consists of linear equations with the operation of Boolean addition, which makes it a tractable problem. To break the symmetry among the variables, nonlinear factors (multiplication) are recognized as an alternative choice in mathematics. On Galois field $\mathbb{Z}_2$, multiplication is Boolean conjunction indeed. As a variant of satisfiable problem involving both Boolean addition and multiplication, a model named MAS nonlinear [30] proposed to investigate detailed organizations of the solution space of a class of CSPs rigorously, is studied by the set theory and magnetization in this section.

In a MAS-nonlinear instance, there are N Boolean variables and $M=\alpha N$ equations chosen randomly from the ensemble of all the possible equations with the form

$$x_i+x_j\cdot x_k=0,\quad 1\le i,j,k\le N. \tag{5}$$

It is easy to obtained that two trivial solutions $\{x_i=0,i=1,\ldots,N\}$ and $\{x_i=1,i=1,\ldots,N\}$ always exist in the solution space for any MAS-nonlinear instance. Supposing $s^1=(x_1^1,\ldots,x_N^1)$ and $s^2=(x_1^2,\ldots,x_N^2)$ are two solutions of a MAS-nonlinear instance $\{x_{(a,i)}+x_{(a,j)}\cdot x_{(a,k)}=0\}_{a=1}^M$, for each equation, we have

$$x^1_{(a,i)}=x^1_{(a,j)}\cdot x^1_{(a,k)},$$

$$x^2_{(a,i)}=x^2_{(a,j)}\cdot x^2_{(a,k)}\Rightarrow x^1_{(a,i)}\cdot x^2_{(a,i)}=(x^1_{(a,j)}\cdot x^2_{(a,j)})\cdot(x^1_{(a,k)}\cdot x^2_{(a,k)}). \tag{6}$$

Therefore, $s^3=s^1\cdot s^2=(x_1^1\cdot x_1^2,\ldots,x_N^1\cdot x_N^2)$ can satisfy all the M equations in this instance and is also a solution. It means that the solution space $\mathcal{S}_{nl}$ of a MAS-nonlinear instance forms a semigroup with Boolean multiplication. The magnetization of the configuration of MAS-nonlinear is defined the same as Eq. (3).

To get a clear description of the magnetization of solutions of MAS-nonlinear, a partial order '$\geq$' among the configurations is introduced as

$$s^1 \geq s^2 \Leftrightarrow \bigwedge_{1\leq i\leq N} (x_i^1 \geq x_i^2) \Rightarrow [m(s^1) \geq m(s^2)]. \tag{7}$$

This partially ordered relation can be derived by the Boolean multiplication relation among configurations as following:

$$s^3 = s^1 \cdot s^2 \Rightarrow (s^1 \geq s^3) \wedge (s^2 \geq s^3).$$

Therefore, all the solutions in MAS-nonlinear problem form a partially ordered set.

In the viewpoint of algebra, semigroup is completely determined by its generators. The formation of its generators plays a key role in understanding the organization of the solution space of MAS-nonlinear. In the view point of set theory, maximal elements possess a crucial status in a partially ordered set. Except the solution $\{x_i=1, i=1,\dots,N\}$, all the generators of the semigroup have one-one correspondence with the maximal elements under the above partially ordered relation. That is to say, generators have larger magnetization than ordinary solutions and cannot be generated by the Boolean multiplications of others.

As the generators of the solutions of MAS-nonlinear are maximal elements under the partially ordered relation '$\geq$', the fewer 0s are assigned in a solution, the more possible it acts as a generator. To identify whether a solution is a maximal element or not, the influence propagation process when one variable takes value 0 should be investigated.

To ensure the satisfiability of an instance when we fix the assignments to variables stepwise, the influence propagation of a variable assigned to 0 in equation $x_i+x_j\cdot x_k=0$ should be

$$x_i = 0 \Rightarrow (x_j = 0) \vee (x_k = 0),$$

$$x_j = 0 \Rightarrow x_i = 0,$$

$$x_k = 0 \Rightarrow x_i = 0. \tag{8}$$

Then, by analyzing the influence propagation from some variable fixed to 0, the formation of the maximal elements can be obtained which undergoes three different phases through $\alpha=\frac{1}{4}$ and $\alpha=\frac{1}{3}$ [30].

As a paralleled study, the influence propagation of a variable taking 1 in equation $x_i+x_j\cdot x_k=0$ is in the following:

$$x_i = 1 \Rightarrow (x_j = 1) \wedge (x_k = 1),$$

$$x_j = 1 \Rightarrow x_i = x_k,$$

$$x_k = 1 \Rightarrow x_i = x_j. \tag{9}$$

It has been obtained that $O(N)$ variables have to be fixed to 1s by the influence propagation for some variable assigned to 1 when the equation density $\alpha > 0.5$ [30].

By the above analysis of the influence propagation of 0 and 1, when $\alpha < \frac{1}{4}$, the magnetization of almost all the maximal elements concentrates to some typical value which is as large as N ($\sim N$); when $\frac{1}{4} \leq \alpha < \frac{1}{3}$, the magnetization of the maximal elements concentrates to several typical values which are as large as N, and the number of maximal elements with lower magnetization is much larger than those with higher magnetization; when $\frac{1}{3} \leq \alpha < \frac{1}{2}$, the magnetization of the maximal elements concentrates to several typical values, some of which are of a finite fraction of N $[O(N)]$ and some of which are as large as N, and the number of maximal elements with lower magnetization $[O(N)]$ is greatly larger than those with higher magnetization ($\sim N$); when $\alpha \geq \frac{1}{2}$, as the emergency of the percolation core of both the influence propagation of 0 and 1, the maximal elements can be classified into two different classes, between which the Hamming distance is scaled by the size of the percolation core induced by the influence propagation of [30]. Therefore, the set of the maximal elements (generators) of the solution space undergoes four different phases shown in Fig. 3, which makes the formation of the maximum elements more and more complicated and the solution space possesses increasingly structural complexity.

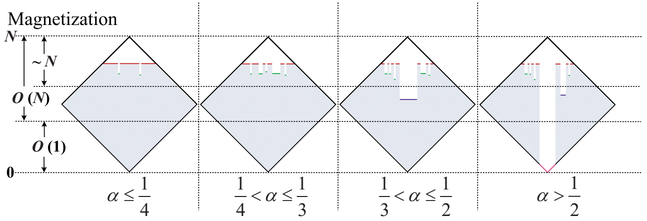
Magnetization
N
~N
O (N)
O (1)
0
$\alpha \leq \frac{1}{4}$
$\frac{1}{4} < \alpha \leq \frac{1}{3}$
$\frac{1}{3} < \alpha \leq \frac{1}{2}$
$\alpha > \frac{1}{2}$

FIG. 3. (Color online) A schematic view for the magnetization variation in maximal elements of solutions with equation density α in different regions. The chromatic bands and the regions filled with shadow represent the maximal elements and the set of assignments in which the solutions generated by the maximal elements involve, respectively. The red, blue, and green bands represent the numbers of maximal elements are $\sim(N)$, $\sim$pol(N), and $\sim$exp(N), respectively, which can be obtained similarly as in our previous work [30]. The gap corresponding the purple bands in the last graph suggests the clustering of maximal elements, which results in the splitting of the solutions.

IV. SATISFIABILITY THRESHOLD OF MASSIVE ALGEBRAIC SYSTEM

Combing the two types of Boolean equations from XOR-SAT and MAS-nonlinear, a model named MAS is proposed in [32]. A MAS instance is defined as: considering a set of N variables $x_1, x_2, \dots, x_N$, we choose $M=\alpha N$ equations randomly and independently with the form of

$$x_i + x_j + x_k = J_{ijk} \quad \text{or} \quad x_i + x_j \cdot x_k = 0, \tag{10}$$

where $J_{ijk} \in \{0,1\}$. This model is proposed for building the correspondence between the classical phase transitions and the mathematical thresholds in the viewpoint of algebra. Here we focus on the algorithmic characteristics of the ensembles of the random instances of MAS and the properties of the two independent parts which are septated artificially but coupled together indeed as the constraints to the solutions.

A. Proof for NP-completeness of MAS

As a constraint satisfiability problem, MAS is NP-complete. A proof is given in the following.

Suppose $f \in \mathbb{Z}_2[x_1, \ldots, x_n]$ is a Boolean quadratic polynomial with the form

$$f(x) = f(x_1, \ldots, x_n) = \sum_{\langle i,j \rangle} x_i \cdot x_j + \sum_{l=1}^{t} x_{k(l)} + c, \qquad (11)$$

where $1 \leq i \leq j \leq n$, $1 \leq k(l) \leq n$, and c is a constant. Replace the quadratic items in $f(x)$ by nonlinear formulas $\tau_w = x_i \cdot x_j$, $w = 1, \ldots, s$, in which s is the number of quadratic items. Then, $f(x)$ is equivalent to

$$f(x_1, \ldots, x_n, \tau_1, \ldots, \tau_s) = \sum_{w=1}^{s} \tau_w + \sum_{l=1}^{t} x_{k(l)} + c, \qquad (12)$$

$$\tau_w = x_i \cdot x_j, \quad w = 1, \ldots, s. \qquad (13)$$

After introducing some new intermediate Boolean variables g, h, ys, zs, a more concise form can be obtained,

$$g + h + \tau_1 = c, \quad \tau_w + x_i \cdot x_j = 0, \quad w = 1, \ldots, s, \qquad (14)$$

$$g + \tau_2 + y_2 = 0, \ldots, y_{s-2} + \tau_{s-1} + \tau_s = 0, \qquad (15)$$

$$h + x_{k(1)} + z_1 = 0, \quad z_1 + x_{k(2)} + z_2 = 0,$$

$$\ldots, z_{t-2} + x_{k(t-1)} + x_{k(t)} = 0. \qquad (16)$$

Denoting the above system as $P(x, x')$, where

$$x' = (\tau_1, \ldots, \tau_s, g, h, y_2, \ldots, y_{s-2}, z_1, \ldots, z_{t-2}),$$

we have

$$f(x) = 0 \Leftrightarrow \exists\, x' \in \{0,1\}^{2s+t-3} s.t. \ \ P(x,x') = 0. \qquad (17)$$

As the number s of quadratic items should be less than $n(n-1)/2$, the total number of variables in P must not be larger than $2n^2$.

Therefore, for a set of Boolean quadratic polynomials $\mathcal{F} = \{f_1, \ldots, f_m\} \subset \mathbb{Z}_2[x_1, \ldots, x_n]$, there is a system $\mathcal{P} = \{P_1, \ldots, P_m\}$ associated with it and

$$\mathcal{F} \text{ is satisfiable} \Leftrightarrow \mathcal{P} \text{ is satisfiable},$$

in which $\mathcal{F}$ has n variables and $\mathcal{P}$ has at most $2n^2$ variables and mn^2 equations. The system $\mathcal{P}$ is with the form of MAS, and a Boolean quadratic polynomial system can be reduced to MAS in polynomial time (no more than mn^2 steps indeed). For that solving Boolean quadratic polynomial system is a NP-complete problem [33], it is evident that MAS is NP-complete.

B. Upper and lower bounds of satisfiability threshold by unit-clause propagation

Since MAS problem consists of two types of different equations, it is interesting to consider the satisfiability threshold of this problem. Defining a parameter q as the ratio of nonlinear equations in MAS, it is easy to see that MAS changes from a linear system to a nonlinear system, when q varies from 0 to 1.

In this section, we propose the upper bounds and lower bounds of the satisfiability thresholds of MAS with different values of parameter q by analyzing a variant of unit-clause algorithm [34]. This algorithm is for the satisfiability judgment of Boolean equation constraints and still named UC for short in this paper. The basic UC resolution round consists of a free step and several forced steps [35]. In a free step, the algorithm chooses a variable randomly or according to some heuristic strategy. In a forced step, the unit clause propagates under the condition of preserving the satisfiability of the original problem. Therefore, the forced steps in each round can be viewed as a branching process.

Considering an instance of MAS with the equation density α and parameter q, we analyze the transition matrix $\mathcal{M}$ of the branching process for forced steps of UC algorithm. Writing t as the number of rounds completed so far and $X(t)$ as the number of variables which have been already set to some values so far. To obtain $X(t)$ dynamically, we focus on the evolution of the expected number of variables assigned to 1/0 in each round, which are denoted as m_1/m_0, respectively, by analyzing the state of equations after the free step or forced steps of each round. Following a free step, there are six cases of the resulted equations:

Case A. $x_i + x_j \cdot x_k = 0$, where $i,j,k \in \{1,2,\ldots,N\}$. The equations in form of case A are the original nonlinear equations of the problem, the number of which is denoted as N_A;

Case B. $x_i + x_j + x_k = 1$, where $i,j,k \in \{1,2,\ldots,N\}$. The equations in form of case B are the original linear equations of the problem, the number of which is denoted as N_B;

Case C. $x_i + x_j + x_k = 0$, where $i,j,k \in \{1,2,\ldots,N\}$. The equations in form of case C are the original linear equations of the problem, the number of which is denoted as N_C;

Case D. $x_i + x_k = 0$, where $i,j,k \in \{1,2,\ldots,N\}$. The equations in form of case D are the resulted equations by assigning the linear-part variable of equations in case A to 1 or assigning one variable of the equations in case B/C to 1/0, the number of which is denoted as N_D;

Case E. $x_j \cdot x_k = 0$, where $i,j,k \in \{1,2,\ldots,N\}$. The equations in form of case E are the resulted equations by assigning the linear-part variable of equations in case A to 0, the number of which is denoted as N_E;

Case F. $x_i + x_k = 1$, where $i,j,k \in \{1,2,\ldots,N\}$. The equations in form of case F are the resulted equations by assigning one variable of the equation in case B/C to 0/1, the number of which is denoted as N_F.

If X variables have been set to some values after t completed rounds, the probability of a variable appearing in a given equation with two variables is $2/(N-X)$, and the probability of a variable appearing in a given equation with 3 variables is $3/(N-X)$. Based on the insights gained from the study on the constraints propagation process of the subproblem MAS nonlinear, we can obtain the expected number of unit clauses (equations) created by the assignment to some variable in free step. When some variable is set to 1, the rest variables appearing in the equations of both case E and F are forced to be assigned to 0, the number of which are $2N_E/(N-X)$ and $2N_F/(N-X)$, respectively. Meanwhile, the rest variables appearing in the equations of both case A and D are forced to be assigned to 1, the numbers of which are $2N_A/(N-X)$ and $2N_D/(N-X)$, respectively. When some variable is set to 0, the rest variables appearing in the equations of both case A and D are forced to be assigned to 0, the numbers of which are $2N_A/(N-X)$ and $2N_D/(N-X)$, respectively. In the same condition, the rest variables appearing in the equations of case F are forced to be assigned to 1, the number of which is $2N_F/(N-X)$. Thus, we have the expression of the transition matrix with respect to X as following:

$$\mathcal{M}(X) = \frac{2}{N-X}\begin{pmatrix} N_A(X)+N_D(X) & N_F(X) \\ N_E(X)+N_F(X) & N_A(X)+N_D(X) \end{pmatrix}. \tag{18}$$

Concerning on the fraction of the variables assigned to certain value, this transition matrix can be rewritten as

$$\mathcal{M}(x) = \frac{2}{1-x}\begin{pmatrix} n_A(x)+n_D(x) & n_F(x) \\ n_E(x)+n_F(x) & n_A(x)+n_D(x) \end{pmatrix}, \tag{19}$$

where $x=X/N$, $n_A(x)=N_A(X)/N$, $n_D(x)=N_D(X)/N$, $n_E(x)=N_E(X)/N$, and $n_F(x)=N_F(X)/N$. The largest eigenvalue of $\mathcal{M}(x)$ is

$$\lambda_{\max}(x) = \frac{2}{1-x}\{n_A(x)+n_D(x)+\sqrt{n_F(x)[n_E(x)+n_F(x)]}\}. \tag{20}$$

If the largest eigenvalue of $\mathcal{M}(x)$ is less than 1, the expected number of variables fixed to certain values can be calculated by the following formula:

$$\begin{pmatrix} m_1 \\ m_0 \end{pmatrix} = [I+\mathcal{M}(x)+\mathcal{M}(x)^2+\ldots]\cdot P_0 = [I-\mathcal{M}(x)]^{-1}\cdot P_0, \tag{21}$$

where $P_0=(p,1-p)^T$ represents the initial probability vector of expected population of unit clauses and I is the identity matrix. When $\lambda_{\max}<1$, the numbers m_0 and m_1 are remaining $O(1)$ in one round, and the UC algorithm succeeds with positive probability. On the other hand, when $\lambda_{\max}>1$, the number of variables forced to assigned certain values proliferates to $O(N)$ with high probability and contradiction exists with a positive probability.

The next step is to analyze the lower bound from the condition of $\lambda_{\max}>1$ for all values of x by random heuristic strategy in the free step of each round. Since there are two possible values for a randomly chosen variable with uniform probability from the unassigned variables, we define a parameter r to represent the probability that a chosen variable is set to 1 [to 0 with probability $(1-r)$]. In round t and $t+1$, we obtain the expressions of the expected changes in the number of fixed variables between the beginnings of two rounds,

$$E[\Delta X(t)] = (m_0+m_1), \tag{22}$$

$$E[\Delta N_A(t)] = -(m_0+m_1)\frac{3N_A(X)}{N-X}+o(1), \tag{23}$$

$$E[\Delta N_B(t)] = -(m_0+m_1)\frac{3N_B(X)}{N-X}+o(1), \tag{24}$$

$$E[\Delta N_C(t)] = -(m_0+m_1)\frac{3N_C(X)}{N-X}+o(1), \tag{25}$$

$$E[\Delta N_D(t)] = -(m_0+m_1)\frac{2N_D(X)}{N-X}+m_1\frac{2N_A(X)+3N_B(X)}{N-X} + m_0\frac{3N_C(X)}{N-X}+o(1), \tag{26}$$

$$E[\Delta N_E(t)] = -2(m_0+m_1)\frac{N_A(X)+N_E(X)}{N-X}+m_0\frac{N_A(X)}{N-X}+o(1), \tag{27}$$

$$E[\Delta N_F(t)] = -(m_0+m_1)\frac{2N_F(X)}{N-X}+m_0\frac{3N_B(X)}{N-X} + m_1\frac{3N_C(X)}{N-X}+o(1). \tag{28}$$

When the number of variables N tends to infinity, as in [34], we can rewrite them in the form of differential equations where the $o(1)$ items are ignored,

$$\frac{dn_A(x)}{dx} = \frac{-3n_A(x)}{1-x}, \tag{29}$$

$$\frac{dn_E(x)}{dx} = \frac{m_0}{m_0+m_1}\frac{n_A(x)}{1-x} - \frac{2n_E(x)}{1-x}, \tag{30}$$

$$\frac{dn_D(x)}{dx} = \frac{m_1}{m_0+m_1}\frac{2n_A(x)+3n_B(x)}{1-x} + \frac{m_0}{m_0+m_1}\frac{3n_C(x)}{1-x} - \frac{2n_D(x)}{1-x}, \tag{31}$$

$$\frac{dn_B(x)}{dx} = \frac{-3n_B(x)}{1-x}, \tag{32}$$

$$\frac{dn_C(x)}{dx} = \frac{-3n_C(x)}{1-x}, \tag{33}$$

$$\frac{dn_F(x)}{dx} = \frac{m_0}{m_0+m_1}\frac{3n_B(x)}{1-x} + \frac{m_1}{m_0+m_1}\frac{3n_C(x)}{1-x} - \frac{2n_D(x)}{1-x}. \tag{34}$$

The initial conditions of these equations are $n_A(0)=\alpha q$, $n_B(0)=n_C(0)=(1/\alpha)(1-q)$, and $n_D(0)=n_E(0)=n_F(0)=0$. By solving these differential equations, we get the expressions of the expected numbers of the variables in case A, B, C, and F,

$$n_A(x) = \alpha q(1-x)^3, \tag{35}$$

$$n_B(x) = n_C(x) = \frac{\alpha}{2}(1-q)(1-x)^3, \tag{36}$$

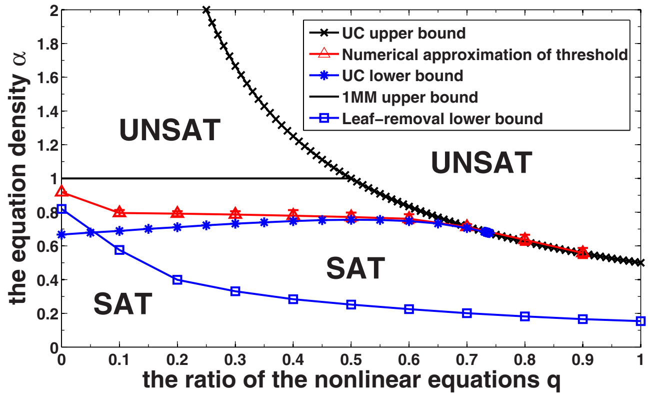

FIG. 4. (Color online) The phase diagram of MAS.

$$n_F(x) = \frac{2}{3}\alpha(1-q)x(1-x)^2. \tag{37}$$

For a given value of parameter r, we can derive the numerical results from Eqs. (20), (21), and (29)–(34). Using the method in [35], we find that the random heuristic parameter r equals to 0, which implies the best choice to lead to fewest implication by numerical experiment results. The lower bound of the satisfiability threshold of MAS with respect to parameter q is chosen as the maximum element in the set $\{\alpha|\lambda_{\max}(x,\alpha,q)<1, x\in[0,1]\}$.

When a single round cascade from the initial variables could fix a finite fraction of the unassigned variables in the large N limit, the process of unit-clause propagation in forced steps will cause contradiction of assignments with high probability. By this condition, we can obtain the upper bound when $O(N)$ variables are percolated with a positive probability in some rounds, i.e., the number of residual unassigned variables X equals to zero on the edge of some critical rounds. From this point of view, there will be such contradiction if $\lambda_{\max}(0)=2q\alpha>1$, and the upper bound of SAT/UNSAT transition is

$$\alpha_{\text{upper}} = \frac{1}{2q}. \tag{38}$$

By the results of the above computation, we find that the lower and upper bounds coincide as the ratio of nonlinear equations q is larger than 0.739. It means that the satisfiability threshold can be exactly located by the upper and lower bounds when the ratio of nonlinear equations is in the interval (0.739,1). Figure 4 shows the upper bounds derived by unit-clause propagation and the first-moment method (1MM) [32] and the lower bounds by unit-clause confliction and leaf-removal process.

C. Leaf-removal analysis for massive algebraic system

Leaf-removal analysis is a classical method which can be used to find the percolation core of large-scale problems, such as XORSAT [20] and Vertex-Cover problems [36,37]. In the following, the leaf-removal process for massive algebraic system is analyzed, to reduce the scale of instance and estimate lower bounds of the satisfiability thresholds of MAS with different values of parameter q.

All the MAS instances can be represented as bipartite factor graphs [19] with N variable nodes and $M=\alpha N$ equation nodes. Each equation node only connects with 3 variable nodes. By the random graph theory, as N sufficiently large, each variable node connects to k equation nodes with probability

$$f_\lambda(k) = \frac{\lambda^k}{k!}e^{-\lambda}, \quad \lambda = 3\alpha, \quad k \in 0 \cup \mathbb{N}.$$

For the linear equation $x_i+x_j+x_k=J_{ijk}$ in MAS, the equation node connects to 3 variables nodes with solid lines; and for the nonlinear equation $x_i+x_j\cdot x_k=0$, the equation node connects to the linear-part variable x_i with solid line and the two nonlinear part variables x_j and x_k with dashed lines. By random graph theory, each variable node connects with k linear equation nodes with probability

$$f_{\lambda_l}(k) = \frac{\lambda_l^k}{k!}e^{-\lambda_l}, \quad \lambda_l = 3(1-q)\alpha, \quad k \in 0 \cup \mathbb{N},$$

and connects with k nonlinear equation nodes with probability

$$f_{\lambda_{nl}}(k) = \frac{\lambda_{nl}^k}{k!}e^{-\lambda_{nl}}, \quad \lambda_{nl} = 3q\alpha, \quad k \in 0 \cup \mathbb{N}.$$

By the graphical representation of MAS instances, a systematic analysis for the leaf-removal process will be provided by both theoretical and numerical ways. For a MAS instance, there exist three leaf-removal conditions which keep its satisfiability unchanged.

Case I. If some variable appears in exact one linear equation, i.e., this variable corresponds to a leaf variable node on the factor graph, then there always exists some assignment of the variable to satisfy this equation;

Case II. If some variable appears as the linear-part variable in exact one nonlinear equation, i.e., this variable corresponds to a leaf variable node on the factor graph, then there always exists some assignment of the variable to satisfy this equation;

Case III. If the two nonlinear-part variables of a nonlinear equation never appear in other equations, i.e., these two variables correspond to two leaf variable nodes connected to the same equation node on the factor graph, then there always exists a pair of assignments of these two variables to satisfy this equation.

On the original factor graph corresponding to the original MAS instance, for that the leaves in the above Case I, II, and III do not affect the satisfiability of the whole instance, re-

TABLE I. The lower bounds of satisfiability threshold by leaf-removal.

q	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
α_L	0.819	0.576	0.399	0.331	0.284	0.252	0.225	0.201	0.182	0.166	0.154

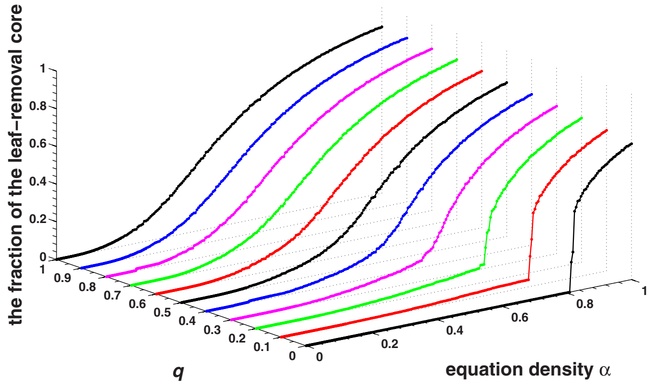

FIG. 5. (Color online) The numerical results of the size of the percolation core after leaf-removal process with different values of q for MAS.

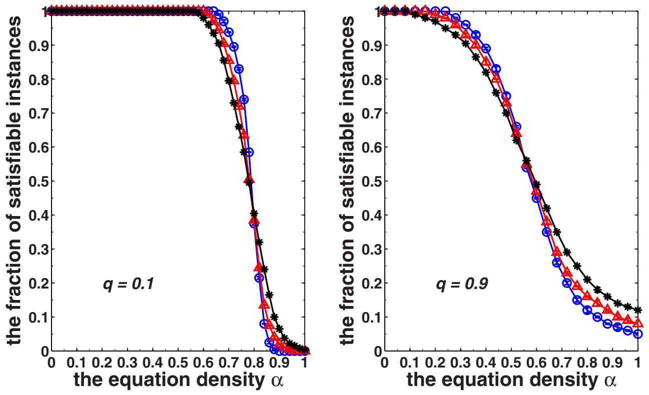

FIG. 6. (Color online) The phase transition of satisfiability of MAS with q=0.1 and 0.9, respectively, by numerical experiments on 500 instances with N=100, 200, and 300. The left graph shows that the points of intersection locate in the interval $0.778<\alpha<0.812$. The right graph shows that the points of intersection locate in the interval $0.529<\alpha<0.587$.

moving these leaves with their connecting equations will produce a new equivalent instance of the original one with smaller scale under the consideration of satisfiability. After the leaf-removal, there will exist new produced leaves on the residual factor graph satisfying the conditions in Case I, II, and III, which allows the continued leaf-removal on the residual graph. Thus, the leaf-removal can be performed until no such leaves satisfying the conditions in Case I, II, and III exist. The final residual graph is percolation core1 in the literature of leaf-removal [20,36,37]. By careful inspection, it is easy to find that the satisfiability of the percolation core is equivalent to that of the original instance.

In the following, we provide a theoretical analysis for the leaf-removal, in which the Case III is neglected as the amount of leaves in condition of Case III occupies a relatively small fraction. Let $f_k(t)$ be the probability of a variable node with degree k after tN equation nodes are removed. It is easy to see that t ranges from 0 to α and $f_k(0)=e^{-3\alpha}(3\alpha)^k/k!=f_\lambda(k)$. Then the evolution equations of $f_k(t)$ are read as

$$\frac{\partial f_0(t)}{\partial t}=\left[2\cdot\frac{f_1(t)}{m(t)}+1\right]\left\{\left[1-\frac{2}{3}q(t)\right]\right\}, \qquad (39)$$

$$\frac{\partial f_1(t)}{\partial t}=\left(2\cdot\frac{2f_2(t)-f_1(t)}{m(t)}-1\right)\left(1-\frac{2}{3}q(t)\right), \qquad (40)$$

$$\frac{\partial f_k(t)}{\partial t}=\left(2\cdot\frac{(k+1)f_{k+1}(t)-kf_k(t)}{m(t)}\right)\left(1-\frac{2}{3}q(t)\right),\quad k\geq 2, \qquad (41)$$

in which $q(t)=\frac{q}{(1-q)e^{-2/3t}+q}$, $m(t)=3\{\alpha-\int_0^t[1-\frac{2}{3}q(t)]dt\}$. In the above equations, $q(t)$ is the ratio of nonlinear equations after tN steps of leaf-removal and $q(0)=q$. After each step of leaf-removal, the number of variable nodes with degree 0 increases with 1 if the chosen leaf is in a linear equation or a linear-part variable in a nonlinear equation, which makes the contribution of $[1-\frac{2}{3}q(t)]$.

For a MAS instance, if the final residual graph (percolation core) has no edges, i.e., all the equations can be removed by the strategy above, then it must have an assignment satisfying all the equations. By the theoretical analysis and experimental results, when the final residual graph of a random instance with given parameters q and α after leaf-removal is almost of isolated variable nodes with probability 1, this instance is satisfiable with high probability. Thus, the lower-bound estimation α_L for satisfiability threshold of random MAS instances can be obtained by the leaf-removal analysis above. The numerical results of leaf-removal for random MAS instance are shown in Table I and Fig. 5.

D. Algorithm and numerical experiments for the satisfiability

Based on the above heuristic strategies of leaf-removal and Gaussian decimation, we propose a complete algorithm to solve the instances of MAS. This algorithm consists of three parts:

(a) the first part is the leaf-removal process to decimate the leaves of both linear and nonlinear type equations, which is to simplify the equations;

(b) the second part is a variant of the Gaussian decimation algorithm, which determines the nondiagonal and diagonal variables which play different roles in the judgment process;

TABLE II. The numerical results of satisfiability threshold.

q	0	0.1	0.2	0.3	0.4
α_c	0.918 ± 0.005	0.795 ± 0.017	0.791 ± 0.015	0.786 ± 0.019	0.779 ± 0.033
q	0.5	0.6	0.7	0.8	0.9
α_c	0.771 ± 0.026	0.761 ± 0.026	0.713 ± 0.0153	0.636 ± 0.030	0.558 ± 0.029

TABLE III. The efficiency comparison between LRR and DPLL algorithms.

α	0.2	0.4	0.6	0.8	1
$R(N=25)$	14.9359	36.0753	298.1746	89.7436	8.6022
$R(N=50)$	96.5909	133.7766	582.3853	565.8602	132.8846
$R(N=100)$	161.9892	424.8925	753.1269	41.1174	20.8237

(c) the third part is the branching process with backtracking based on the properties of the influence propagation among the variables, which could be considered as a complete algorithm on the ordered set of variables to check the satisfiability of MAS.

A more detailed version of this algorithm is shown in the following.

Leaf-removal Reordered Algorithm

```
INPUT: EquationSet $S$, Linear-equationSet $S_I$ and
Nonlinear-equationSet $S_{II}$
OUTPUT: SAT or UNSAT

while ($S$ contains equations satisfying the leaf-removal conditions in section IV.C)
  {delete these equations from $S$}
if $S=\emptyset$
  { return SAT}
else
  {run Gaussian decimation on $S$
  get the resulted coefficient matrix $\mathcal{A}$ by column permutation
  reorder variables from off-diagonal variables to diagonal variables in $\mathcal{A}$
  while (the current variable is off-diagonal)
    { if (the current variable is not assigned)
        {assign the current variable to 1
        assign related variables by influence propagation as in Eq. (9)}
      if (contradiction occurs)
        {assign the current variable to 0
        assign related variables by influence propagation as in Eq. (8)}
      if (contradiction occurs)
        {backtrack}
      else
        {go to the next variable}
      }
if(the current variable is diagonal)
    { return SAT}
else
    { return UNSAT}
}
```

We run leaf-removal reordered (LRR) algorithm on instances of MAS for different values of q to estimate the locations of satisfiability thresholds. Here, the points of intersection of different curves are considered as the approximate locations of satisfiability thresholds. Figure 6 shows the phase transition of satisfiability of MAS instances with $N=100$, 200, and 300 when the parameter q takes the values of 0.1, and 0.9, respectively. In the numerical experiments on instances with $N=500$, we take the midpoint of the interval which consists of the possible points of intersection in the case of certain value of q as the approximations of satisfiability threshold α_c. These approximations of satisfiability thresholds and variations are shown in Table II and as the red triangles in Fig. 4.

In order to compare the LRR algorithm to classical Davis-Putnam-Logemann-Loveland (DPLL) algorithm [38], we study results of the numerical experiments on these two algorithms. We run 1000 instances with 25, 50, and 100 variables in the case of different equation density $\alpha=0.2$, 0.4, 0.6, 0.8, and 1, respectively, when $q=0.5$. There exists a large gap of the time of computation between these two algorithms. Table III shows the ratios of the running time of LRR algorithm to DPLL algorithm for the instances of MAS.

V. CONCLUSION

In this work, we study the correlation between the magnetization of generators and the clustering of solutions of XORSAT problem by Gaussian elimination process. Then, we investigate the partially ordered correlation and the characteristics of maximal elements of solutions of MAS-nonlinear problem. Furthermore, we estimate the lower and upper bounds of the satisfiability threshold of MAS by analyzing unit-clause and leaf-removal mechanism. Taking the advantage of the explicit algebraic and geometrical characteristics of solutions of two subproblems septated artificially, a complete algorithmic frame to solve MAS problem is proposed and used to approximate the satisfiability threshold.

By studying the algebraic properties of Boolean equations, we can explore the implicated connection to the onset of algorithmic hardness and organization of solutions. To investigate the essential hardness for NP problems, algebraic methods and geometrical measurement will be enlightening alternatives. It is interesting to study how the generators of the subproblems construct the whole solution space and the detailed organizations of solutions by the intersection of a group and a semigroup.

ACKNOWLEDGMENTS

This work is supported by the National Key Basic Research Project of China under Grant No. 2005CB321902 and the Innovation Foundation of BUAA for Ph.D. Graduates.

[1] P. Van Hentenryck, *Constraint Satisfaction in Logic Programming* (MIT Press, Cambridge, MA, 1989).

[2] G. Istrate, Discrete Appl. Math. **153**, 141 (2005).

[3] F. Krzakala, A. Montanari, F. Ricci-Tersenghi, G. Semerjian, and L. Zdeborova, Proc. Natl. Acad. Sci. U.S.A. **104**, 10318 (2007).

[4] R. Monasson and R. Zecchina, Phys. Rev. Lett. **76**, 3881 (1996).

[5] M. Mézard and R. Zecchina, Phys. Rev. E **66**, 056126 (2002).

[6] S. Mertens, Phys. Rev. Lett. **81**, 4281 (1998).

[7] R. Mulet, A. Pagnani, M. Weigt, and R. Zecchina, Phys. Rev. Lett. **89**, 268701 (2002).

[8] R. Monasson, R. Zecchina, S. Kirkpatrick, B. Selman, and L. Troyansky, Nature (London) **400**, 133 (1999).

[9] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).

[10] S. Kirkpatrick and B. Selman, Science **264**, 1297 (1994).

[11] D. Achlioptas, L. Kirousis, E. Kranakis, and D. Krizanc, Theor. Comput. Sci. **265**, 109 (2001).

[12] B. Bollobás, C. Borgs, J. T. Chayes, J. H. Kim, and D. B. Wilson, Random Struct. Algorithms **18**, 201 (2001).

[13] D. Achlioptas and Y. Peres, J. Am. Math. Soc. **17**, 947 (2004).

[14] D. Achlioptas, P. Beame, and M. Molloy, J. Comput. Syst. Sci. **68**, 238 (2004).

[15] J. Ardelius and L. Zdeborova, Phys. Rev. E **78**, 040101(R) (2008).

[16] D. Achlioptas and F. Ricci-Tersenghi, SIAM J. Comput. **39**, 260 (2009).

[17] A. Hartmann, A. Mann and W. Radenbach, J. Phys.: Conf. Ser. **95**, 012011 (2008).

[18] M. Mézard, T. Mora, and R. Zecchina, Phys. Rev. Lett. **94**, 197205 (2005).

[19] A. Braunstein, M. Mézard, and R. Zecchina, Random Struct. Algorithms **27**, 2 (2004).

[20] M. Mézard, F. Ricci-Tersenghi, and R. Zecchina, J. Stat. Phys. **111**, 314 (2003).

[21] M. Mézard and T. Mora, J. Stat. Mech.: Theory Exp. **2006**, P10007.

[22] O. Dubois and J. Mandler, Proceedings of 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS 2002), 2002 (unpublished).

[23] D. MacKay, *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003).

[24] A. Braunstein, M. Leone, F. Ricci-Tersenghi, and R. Zecchina, J. Phys. A **35**, 7559 (2002).

[25] S. Mertens, M. Mézard, and R. Zecchina, Random Struct. Algorithms **28**, 3 (2005).

[26] E. Friedgut, J. Am. Math. Soc. **12**, 1017 (1999).

[27] L. Zdeborová and M. Mézard, Phys. Rev. Lett. **101**, 078702 (2008).

[28] H. Zhou, Phys. Rev. Lett. **94**, 217203 (2005).

[29] L. Zdeborová and M. Mézard, J. Stat. Mech.: Theory Exp. (2008) P12004.

[30] W. Wei, B. Guo and Z. Zheng, J. Stat. Mech.: Theory Exp. (2009) P02010.

[31] S. N. Dorogovtsev, J. F. F. Mendes, and A. N. Samukhin, Phys. Rev. E **64**, 025101(R) (2001).

[32] W. Wei, B. Guo, and Z. Zheng, Proceedings of MACIS, 2007 (unpublished), pp. 1–14.

[33] L. Blum, F. Cucker, M. Shub, and S. Smale, *Complexity and Real Computation* (Springer-Verlag, New York, 1997).

[34] V. Kalapala and C. Moore, e-print arXiv:cs/0508037v2.

[35] C. Ming-Te and J. Franco, Inf. Sci. (N.Y.) **51**, 289 (1990).

[36] M. Weigt, Eur. Phys. J. B **28**, 369 (2002).

[37] M. Bauer and O. Golinelli, Eur. Phys. J. B **24**, 339 (2001).

[38] C. Sinz, J. Autom. Reason. **39**, 219 (2007).