Probabilities of 2-Xor Functions

Élie de Panafieu[1]*, Danièle Gardy[2]**, Bernhard Gittenberger[3]***, and Markus Kuba[3]†

[1] Univ. Paris Diderot, Sorbonne Paris Cité, LIAFA, UMR 7089, 75013, Paris, France.
[2] PRISM, Univ. of Versailles, France.
[3] Institute of Discrete Mathematics and Geometry, TU Wien, Austria.

Abstract. The problem 2-Xor-Sat asks for the probability that a random expression, built as a conjunction of clauses $x \oplus y$, is satisfiable. We consider here a refinement of this question, namely the probability that a random expression computes a specific Boolean function. The answer involves a description of 2-Xor expressions as multigraphs, and uses classical methods of analytic combinatorics.

Keywords: multigraphs, probability of Boolean functions, 2-Xor expressions

1 Introduction

In constraint satisfaction problems we ask for the probability that a random expression, built on a finite set of Boolean variables according to some rules (k-Sat, k-Xor-Sat, NAE, . . .), is (un)satisfiable. The behaviour of this probability, when the number n of Boolean variables and the length m of the expression (usually defined as the number of clauses) tend to infinity, most specially the existence and location of a threshold from satisfiability to unsatisfiability as the ratio m/n grows, has given rise to numerous studies. The literature in this direction is vast, for Xor-functions see e.g. [1–5].

Defining a probability distribution on Boolean functions through a distribution on Boolean expressions is *a priori* a different question. Quantitative logic aims at answering such a question, and many results have been obtained when the Boolean expression, or equivalently the random tree that models it, is a variation of well-known combinatorial or probabilistic tree models (Galton-Watson and Pólya trees, binary search trees, etc).

So we have two frameworks: on the one hand we try to determine the probability that an expression is satisfiable; on the other hand we try to identify probability distributions on Boolean functions. It is only natural that we should wish to merge these two approaches: what if we set satisfiability problems into the framework of quantitative logic (this only requires to choose a suitable model of expressions), and ask for

* Supported by the ANR projects BOOLE (2009-13) and MAGNUM (2010-14).

** Part of the work of this author was done during a long-term visit at the Institute of Discrete Mathematics and Geometry, TU Wien. Supported by the P.H.C. Amadeus project *Probabilities and tree representations for Boolean functions* and by the ANR project BOOLE (2009-13).

*** Supported by the FWF (Austrian Science Foundation), Special Research Program F50, grant F5003-N15, and by the ÖAD, grant Amadée F03/2013.

† Supported by the ÖAD, grant Amadée F03/2013.

the probability of FALSE – this is the classical satisfiability problem – *and* of the other Boolean functions? This amounts to refining the satisfiable case, which gathers together all the functions differing from FALSE, into subcases according to the exact (class of) Boolean function(s) that is computed.

Within this unified framework one could, e.g., ask for the probability that a random expression computes a function that is satisfied by a specific number of assigments. Although this may turn out to be out of our reach for most classical satisfiability problems, there are some problems for which we may still have hope to obtain a (partial) description of the probability distribution on Boolean functions. The case of 2-Xor expressions is such a problem, and this paper is devoted to presenting our results in this domain.

The 2-Xor-Sat satisfiability problem has been studied by Creignou and Daudé [1] who established the existence of a threshold for $m = \frac{n}{2}$, then proved in [4] that this threshold is coarse. Further work by Daudé and Ravelomanana [6] and by Pittel and Yeum [7] led to a precise understanding of the transition in a window of size $n^{2/3}$.

The paper is organized as follows. We present in the next section the 2-Xor problem and the set of Boolean functions that can be attained by such expressions, then give a modelization in terms of multigraphs, before considering in Section 3 how enumeration results on classes of multigraphs allow us to compute probabilities of Boolean functions. We then give explicit results for several classes of functions in Section 4, and conclude with a discussion on the relevance and of possible extensions of our work.

2 Boolean Expressions and Functions, and Multigraphs

2.1 2-Xor Expressions and Boolean Functions

Starting from an infinite set $\{x_1, x_2, \ldots\}$ of Boolean variables, we define a 2-Xor expression as a finite conjunction of clauses $l \oplus l'$, where l and l' are literals, i.e. either some x_i or $\bar{x_i}$. We shall denote by m the number of clauses of an expression. Now each 2-Xor expression defines a Boolean function on a finite number of variables, but not all Boolean functions on a finite number of variables can be obtained from a 2-Xor expression. We define $\mathcal{X}$ as the set of functions from $\{0,1\}^{\mathbb{N}}$ to $\{0,1\}$, such that there exists at least one 2-Xor expression representing them. We also define, for each $n \geq 1$, the set $\mathcal{X}_n$ of functions in $\mathcal{X}$ such that there exists an expression representing the function, that does not use any of the variables $x_{n+1}, x_{n+2}, \ldots$ This implies that $\mathcal{X}_{n_1} \subset \mathcal{X}_{n_2}$ for $n_1 \leq n_2$, and that $\mathcal{X} = \cup_{n \geq 1} \mathcal{X}_n$.[4]

Consider now the expressions that can represent a function of $\mathcal{X}_n$. The literals in a clause are ordered (the clauses $x \oplus y$ and $y \oplus x$ are distinct), hence there are $4n^2$ distinct clauses. We assume that the m clauses are drawn with a uniform probability and with replacement (i.e., a clause can appear several times), and are unordered, i.e. we are dealing with a *set* of clauses. This framework allows us to define a probability distribution on the set $\mathcal{X}_n$: $\Pr_{[m,n]}(f) = \frac{N_{[m,n]}(f)}{N_{[m,n]}}$, with $N_{[m,n]}$ the total number of expressions with m clauses on the variables $x_1, \ldots, x_n$, and $N_{[m,n]}(f)$ the number of these expressions that compute f.

[4] For brevity's sake, "(the set of) Boolean functions" in the sequel is to be understood as either the set $\mathcal{X}_n$ or the set $\mathcal{X}$, according to the context.

2.2 The Sets $\mathcal{X}_n$

Rewriting a clause $l_1 \oplus l_2$ as $l_1 \sim \bar{l_2}$ (i.e., l_1 and l_2 must take opposite values for the clause to evaluate to TRUE), we see that the functions we obtain can be written as a conjunction of equivalence relations on literals: $(l_1 \sim \cdots \sim l_p) \wedge (l_{p+1} \sim \cdots \sim l_q) \wedge \cdots \wedge (l_{r+1} \sim \cdots \sim l_s)$. E.g., for $n = 7$ the expression $(x_1 \oplus x_3) \wedge (\bar{x_6} \oplus x_5) \wedge (x_7 \oplus \bar{x_7}) \wedge (x_2 \oplus \bar{x_3})$ computes a Boolean function f that we can write as $(x_1 \sim \bar{x_2} \sim \bar{x_3}) \wedge (x_5 \sim x_6)$, and this function partitions the Boolean variables into the subsets $\{x_1, x_2, x_3\}$, $\{x_4\}$, $\{x_7\}$ and $\{x_5, x_6\}$.

If a relation $l \sim \bar{l}$ appears in at least one of the equivalence relations, the expression simply computes FALSE. In other words: *For any $n \geq 1$, the set $\mathcal{X}_n$ comprises exactly the function* FALSE *and those functions that partition the set of the n Boolean variables into subsets, as follows: the variables (or their negations) in a given part are equivalent; a variable which appears in no clause of an expression computing the function, or only as $l \sim l$, is put in a singleton.*

We now define the following equivalence relation on $\mathcal{X}_n$. *Two Boolean functions f and g on n variables are equivalent, if g can be obtained from f by permuting the variables and flipping some of them.* We denote by $\mathcal{C}(f)$ the equivalence class of a function f. All the Boolean functions in $\mathcal{C}(f)$ share the same probability $\Pr_{[m,n]}(f)$.

Let $f \in \mathcal{X}$; we say that a Boolean variable x is an *essential* variable w.r.t. f iff $f|_{x=1} \neq f|_{x=0}$. Let $f \notin \{\text{TRUE}, \text{FALSE}\}$ and $e(f) \leq n$ be the number of its essential variables[5]; then $f \notin \mathcal{X}_{e(f)-1}$ but $f \in \mathcal{X}_{e(f)}$. In our example, $e(f) = 5$.

It is not hard to see that, with the exception again of FALSE that is in a class by itself, the classes we have thus defined on $\mathcal{X}_n$ are in bijection with partitions of the integer n; in our example the function f partitions the integer 7 as $1 + 1 + 2 + 3$.

Let $\mathbf{i} = (i_\ell)_{\ell \geq 1}$ be an integer partition of n, written in its part-count representation. Hence $i_\ell \geq 0$ for all ℓ and $s(\mathbf{i}) := \sum_\ell \ell\, i_\ell = n$; the total number of parts (or *blocks*) is $\xi(\mathbf{i}) := \sum_\ell i_\ell$ and i_ℓ is the number of parts of size ℓ. Partitions of the type $i_\ell = 0$ except $i_n = 1$ appear regularly in the sequel; we shall denote such a partition by $\mathbf{imax}(n)$. We write $\mathbf{i}(f)$ for the integer partition associated to a Boolean function f, and we extend the notation for the equivalence class into $\mathcal{C}_{\mathbf{i}} = \mathcal{C}(f)$ when $\mathbf{i} = \mathbf{i}(f)$.

Our running example corresponds to the integer partition $(n-5, 1, 1, 0, 0, 0)$ on $n \geq 5$ variables, which has $n - 3$ parts; the set partition it induces on the set of Boolean variables may be taken, for example, equal to $\{x_1, x_2\}, \{x_3, x_4, x_5\}$. The function TRUE corresponds to the integer partition $(n, 0, \ldots, 0)$ and is computed by the expressions that have only clauses of the type $l \oplus \bar{l}$.

Proposition 1. *Set $p(n)$ as the number of partitions of n; the number of classes of computable Boolean functions is then $p(n) + 1$. The class associated to a partition $\mathbf{i} = (i_\ell)$ has cardinality $\frac{2^{n-\xi(\mathbf{i})}\, n!}{\prod_{\ell \geq 1} i_\ell! (\ell!)^{i_\ell}}$. The number of assigments satisfying a function $f \in \mathcal{C}_{\mathbf{i}}$ is $2^{\xi(\mathbf{i})}$.*

[5] Although the constant functions can only be written as 2-Xor expressions invoving one or more variables, they have no essential variable: $e(\text{TRUE}) = e(\text{FALSE}) = 0$.

2.3 2-Xor Expressions as Colored Multigraphs

Consider multigraphs, i.e. graphs where we allow loops and multiple edges. Set $M_{m,n}$ as the number of multigraphs on n vertices[6] and m edges or loops, each multigraph being weighted as follows: every loop contributes a multiplicative factor $1/2$ to the weight, each k-fold edge a factor $1/k!$. The generating function for weighted multigraphs is (see Janson, Knuth, Luczak and Pittel [8])

$$M(z,v) = \sum_{m,n} M_{m,n} z^m \frac{v^n}{n!} = \sum_{n \geq 0} e^{\frac{n^2}{2} z} . \frac{v^n}{n!} .$$

A multigraph being a set of connected components, the g.f. for *connected* multigraphs is

$$C(z,v) = \log M(z,v) = \sum_{r \geq -1} z^r C_r(zv), \tag{1}$$

where we have set $r = m - n$, the *excess* of the multigraph, and where $C_r(z)$ enumerates the *connected* multigraphs of fixed excess r.

We are now ready to define a bijection between Boolean expressions of m clauses on n variables, and *colored* multigraphs on n vertices and with m edges, i.e. multigraphs with different types (colors) of edges between any two vertices, as follows.

- Each Boolean variable x_ℓ corresponds to a vertex, and each 2-Xor clause to an edge between two distinct vertices, or to a loop on one vertex; each loop or edge can be repeated.
- A loop on vertex x can appear in four colors : $x \oplus x$, $x \oplus \bar{x}$, $\bar{x} \oplus x$ or $\bar{x} \oplus \bar{x}$.
- An edge between two distinct vertices x_i and x_j can appear in eight colors: $l_i \oplus l_j$ or $l_j \oplus l_i$, where l_i and l_j are respectively equal to x_i or its negation, and x_j or its negation.

Proposition 2. *There is a bijection between 2-Xor expressions, and multigraphs where loops are 4-colored and other edges are 8-colored. Hence the generating function for 2-Xor expressions is $M(8z, v)$.*

Let $f \in \mathcal{X}_n$; then $\xi(\mathbf{i}(f))$ is the number of connected components of the associated multigraph.

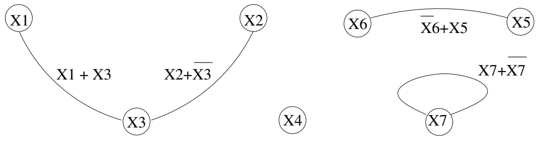

Fig. 1. The colored multigraph for our running example.

[6] As is usual when enumerating such structures, we consider labels on the vertices, say $1, \ldots, n$.

2.4 The Different Ranges

We shall consider in the sequel the range where m and n are related, and set $m \sim \alpha n$ (α is usually assumed to be a constant). It is well known ([6]) that the probability that a random expression is satisfiable decreases from 1 to 0 when α increases, with a (coarse) threshold at $\frac{1}{2}$. However, *a Boolean function corresponding to a partition of the n Boolean variables into p blocks cannot appear before at least $n - p$ clauses have been drawn, i.e. before $m \geq n - p$.* E.g., the function $x_1 \sim \cdots \sim x_n$ cannot appear for $m < n - 1$, which means that it has a non-zero probability only for $\alpha \geq 1$, much later than the threshold – and at this point the probability of FALSE is $1 - o(1)$. This leads us to define regions according to the value of α when $n, m \to +\infty$:

- $\alpha < 1/2$. Here the probability of satisfiability is non-zero, but the attainable functions cannot have more than $n(1 - \alpha)$ blocks.
- $1/2 < \alpha < 1$. Some Boolean functions still have probability zero, but now the probability of satisfiability is $o(1)$ and the probability of FALSE is $1 - o(1)$. Thus any other attainable Boolean function has a vanishing probability $o(1)$.
- $1 \leq \alpha$. At this point all the attainable Boolean functions have non-zero probability, but again the probability of FALSE is tending to 1.

3 Probabilities on Boolean Functions

We consider here how we can obtain the probability of satisfiability (or equivalently of FALSE), or of any function in $\mathcal{X}_n$. The reader should recall that the probabilities given in the sequel are actually distributions on $\mathcal{X}_n$, i.e. they depend on n and m. Letting n and $m = m(n)$ grow to infinity amounts to specializing the probability distribution $\Pr_{[m,n]}(f)$ (defined in Section 2.1 for $f \in \mathcal{X}_n$) into $Pr_{[m(n),n]}(f)$. We shall be interested in its limit when $n \to +\infty$ and f is a function of $\mathcal{X}$. We begin with the case $f = \text{FALSE}$ (which is the usual satisfiability problem) and derive anew the probability of satisfiability in the critical window, before turning to general Boolean functions.

3.1 Probability of Satisfiability

Theorem 1. *The probability that a random expression is satisfiable is*

$$\Pr_{[m,n]}(\mathrm{S}at) = \frac{[z^m v^n]\sqrt{M(4z, 2v)}}{[z^m v^n]M(8z, v)}.$$

Its asymptotic value for $n \to +\infty$ and $m = \frac{n}{2}(1 + \mu n^{-1/3})$ is

$$n^{-1/12}\sqrt{2\pi}\sum_{r \geq 0} \frac{e_r^{(1/2)}}{2^r} A(3r + 1/4, \mu),$$

where

$$e_r^{(\sigma)} = [z^{2r}] \left(\sum_{k \geq 0} \frac{(6k!) z^{2k}}{2^{5k} 3^{2k} (2k)! (3k)!} \right)^{\sigma},$$

$$A(y, \mu) = \frac{e^{-\mu^3/6}}{3^{(y+1)/3}} \sum_{k \geq 0} \frac{(3^{2/3}\mu/2)^k}{k! \mathit{\Gamma}\left(\frac{y+1-2k}{3}\right)}.$$

Proof. To obtain the g.f. for satisfiable expressions, we shall count the number of pairs {satisfiable expression, satisfying assignment}, then get rid of the number of satisfying assignments. We can assign TRUE or FALSE to each variable, and one of eight colors to an edge, hence $M(8z, 2v)$ counts all pairs {expression, assignment}.

Once we have chosen an assignment of variables, for an expression to be satisfiable we have to restrict the edges we allow. Say that x and y are assigned the same value; then the edges colored by $x \oplus y$, $y \oplus x$, $\bar{x} \oplus \bar{y}$ or $\bar{y} \oplus \bar{x}$ cannot appear in a satisfiable expression. For a similar reason, the only loops allowed are $x \oplus \bar{x}$ or $\bar{x} \oplus x$. We thus count multigraphs with 2 colors of loops and 4 colors of edges, which gives a g.f. equal to $M(4z, 2v)$.

Now consider the generating function $S(z, v)$ for satisfiable expressions: we claim that it is equal to $\sqrt{M(4z, 2v)}$. To see this, choose an expression computing a Boolean function f, and consider how many assignments satisfy it: we have seen (cf. Proposition 1) that their number is equal to $2^{\xi(f)}$, with $\xi(f)$ the number of connected components (once we have chosen the value of a single variable in a block, all other variables in that block have received their values if the expression is to be satisfiable). This means that, writing $S(z, v) = \exp \log S(z, v)$ with $\log S(z, v)$ the function for connected components, the g.f. enumerating the pairs {expression, satisfiable assignment} is equal to $\exp(2 \log S(z, v)) = S(z, v)^2$. As we have just shown that it is also equal to $M(4z, 2v)$, the value of $\Pr_{[m,n]}(Sat)$ follows.

To obtain the asymptotics in the critical window $m = n/2 + \mathcal{O}(n^{2/3})$, we use Lemma 1 below, which is an easy variation of [8, Lemma 3]. The function $A(y, \mu)$ is a variation of the classical Airy function; see for example [8, Lemma 3], [9, Theorem 11] or [10, Theorem IX.16].

Lemma 1. *Let us consider a positive real value σ, a bounded parameter μ and $m = \frac{n}{2}(1 + \mu n^{-1/3})$. Then, with the notations of Theorem 1,*

$$n! [z^m v^n] M(z, v)^{\sigma} \sim \frac{n^{2m}}{2^m m!} \sigma^{n-m} n^{(\sigma-1)/6} \sqrt{2\pi} \sum_r \sigma^r e_r^{(\sigma)} A(3r + \sigma/2, \mu).$$

Theorem 2. *The probability for a random satisfiable expression with n variables and m clauses to be satisfied by a random input, in the range $m = \frac{n}{2}(1 + \mu n^{-1/3})$, is*

$$\Pr_{[m,n]}(Sat) = \frac{[z^m v^n] M(4z, 2v)}{2^n [z^m v^n] \sqrt{M(4z, 2v)}} \sim \frac{n^{1/12}}{2^m} \frac{\sum_r e_r^{(1)} A(3r + 1/2, \mu)}{\sum_r 2^{-r} e_r^{(2)} A(3r + 1/4, \mu)}.$$

3.2 Probability of a Given 2-Xor Function

We now refine the probability of satisfiability, by computing the probability of a specific Boolean function $\neq$ FALSE. We first give in Proposition 3 the generating functions for all Boolean functions (except again FALSE), then use it to provide in Theorem 3 a general expression for the probability of a Boolean function, or rather of all the functions of an equivalence class $C_{\mathbf{i}}$. This theorem is at a level of generality that does not give readily precise probabilities, and we delay until Section 4 such examples of asymptotic probabilities.

Proposition 3. *For* $\mathbf{i}$ *an integer partition, define* $\phi_{\mathbf{i}}(z)$ *as the generating function for Boolean expressions that compute a specific Boolean function* f *in the class* $\mathcal{C}_{\mathbf{i}}$*:* $\phi_{\mathbf{i}}(z) = \sum_m N_{[m,n]}(f)\, z^m$*. When* $\mathbf{i} = \mathbf{imax}(n)$*, we set* $\phi_n(z) := \phi_{\mathbf{imax}(n)}(z)$*. Then*

$$\phi_n(z) = \left[\frac{v^n}{n!}\right] C(4z, v); \qquad \phi_{\mathbf{i}}(z) = \prod_{\ell \geq 1} \left(\phi_\ell(z)\right)^{i_\ell}.$$

Proof. A canonical representant of the class $\mathbf{imax}(n)$ is the function $x_1 \sim \cdots \sim x_n$. Any expression that computes it corresponds to a connected multigraph, where we only allow the 2 types of loops that compute TRUE, and the 4 types of edges between x_i and x_j $(i \neq j)$ that compute $x_i \sim x_j$; this gives readily the expression of $\phi_n(z)$.

As for functions whose associated multigraphs have several components, such multigraphs are a product of connected components; hence the global generating function is itself the product of the generating functions for each component.

Theorem 3. *1. The probability that a random expression of* m *clauses on* n *variables computes the function* $x_1 \sim \cdots \sim x_n$ *is*

$$\Pr_{[m,n]}(x_1 \sim \cdots \sim x_n) = \frac{\left[z^m \frac{v^n}{n!}\right] C(4z, v)}{\left[z^m \frac{v^n}{n!}\right] M(8z, v)} = \frac{m!}{n^{2m}} \left[\frac{v^n}{n!}\right] C_{m-n}(v).$$

2. *Let* f *be a function of* $\mathcal{X}$*, with* $q = \xi(\mathbf{i}(f))$*, and* $B_1, \ldots, B_q$ *be the blocks of* $\mathbf{i}(f)$*, with* r_j $(1 \leq j \leq q)$ *the excess of the block* B_j*. The probability that a random expression of* m *clauses on* n *variables computes* f *is*

$$\Pr_{[m,n]}(f) = \frac{m!}{n^{2m}} \sum_{\substack{r_1, \ldots, r_q \geq -1 \\ r_1 + \cdots + r_q = m-n}} \prod_{j=1}^{q} \left[\frac{v^{|B_j|}}{|B_j|!}\right] C_{r_j}(v).$$

Proof. By the correspondance between 2-Xor expressions and weighted multigraphs the probability that an expression of m clauses on n variables computes a function f can be expressed as follows:

$$\Pr_{[m,n]}(f) = \frac{[z^m]\phi_{\mathbf{i}}(z)}{[z^m \frac{v^n}{n!}] M(8z, v)}.$$

Expressing $\phi_{\mathbf{i}}$ in terms of coefficients of powers of $C(4z, v)$, then substituting the expression (1) for C, gives the result after careful management of the coefficients.

4 Explicit Probability Computations

We now show on examples how Theorem 3 allows us to compute the asymptotic probability of a specific function.

We consider first a Boolean function f with a fixed number $e(f)$ of essential variables, and consider how its probability varies when $n \to +\infty$ (i.e. when we add non-essential variables), then turn to functions that vary with n, either with a fixed number of blocks (this includes functions that are "close to" FALSE in the sense that they have few blocks, hence few satisfying assigments), or with a number of blocks that grows with n (e.g., $\frac{n}{j}$ blocks of size j for some $j \geq 2$).

4.1 Probability of a Fixed Function

We compute here the probability of any specific function, when m is large enough so that it can be obtained, and see how it varies when $n, m \to +\infty$ with fixed ratio α.

Proposition 4. *Let $f \in \mathcal{X}_n$, with $\mathrm{e}(f)$ the number of its essential variables, and $\mathbf{i}(f) = (i_1, i_2, \dots)$ its associated integer partition. Assume $m = \alpha n \geq n - \xi(\mathbf{i}(f))$; then*

$$P_{[\alpha n, n]}(f) \sim \frac{e^{\alpha\, e(f)}}{(2n)^{\alpha n}} \prod_{\ell \geq 2} \left(\ell! \phi_\ell \left(\frac{\alpha}{2}\right)\right)^{i_\ell} \qquad (n \to +\infty).$$

4.2 Asymptotics for a Single-Block Function

We consider here the class of $x_1 \sim \cdots \sim x_n$, and the range $m \geq n - 1$. This corresponds to (a subset of) the third range of Section 2.4. From Theorem 3 , we have

$$\mathrm{Pr}_{[m,n]}(x_1 \sim \cdots \sim x_n) = \frac{m!}{n^{2m}} . \left[\frac{v^n}{n!}\right] C_{m-n}(v).$$

We now specialize this expression according to the possible values for the excess $r = m - n$. For the first three cases, we use the fact that for each fixed excess r, there is an explicit constant K_r such that

$$\left[\frac{v^n}{n!}\right] C_r(z) \sim K_r .\, n^{n + \frac{3r-1}{2}}.$$

We use the result of [11] and the alternative proof of [12] to derive the remaining cases.

1. For $r = -1$, we have $\mathrm{Pr}_{[m,n]}(x_1 \sim \cdots \sim x_n) = \frac{(n-1)!}{n^n} \sim \sqrt{\frac{2\pi}{n}}\; e^{-n}$.
2. For $r = 0$, we get $\mathrm{Pr}_{[m,n]}(x_1 \sim \cdots \sim x_n) \sim \frac{\pi}{2}\; e^{-n}$.
3. For $r \geq 1$ but still fixed, $\mathrm{Pr}_{[m,n]}(x_1 \sim \cdots \sim x_n) \sim K_r\, e^{-n} n^{r/2}$ where the constant K_r can be made explicit.
4. For $r \to \infty$ and $r = o(\sqrt{n})$, $\mathrm{Pr}_{[m,n]}(x_1 \sim \cdots \sim x_n) \sim \sqrt{\frac{3}{2}} \frac{e^{r/2}}{(2\sqrt{3})^r} e^{-n} \left(\frac{n}{r}\right)^{r/2}$.
5. For $r = cn$ for a constant $c > 0$, $\mathrm{Pr}_{[m,n]}(x_1 \sim \cdots \sim x_n) \sim K \left(\frac{(1+c)^c \cosh\zeta}{(2\zeta)^c e^{1+c}}\right)^n$ where $\zeta \coth\zeta = 1 + c$ and $K = \sqrt{1+c} \frac{e^{2\zeta} - 1 - 2\zeta}{\sqrt{\zeta(e^{4\zeta} - 1 - 4\zeta e^{2\zeta})}}$.

6. When $r \to +\infty$ and $2m/n - \log(n)$ is bounded - which covers the two previous cases - then $\mathrm{Pr}_{[m,n]}(x_1 \sim \cdots \sim x_n) \sim \frac{K}{(2\zeta)^r} \left(\frac{\sinh\zeta}{\zeta}\right)^n \frac{(1+r/n)^{n+r+1/2}}{e^{n+r}}$ where ζ is the positive solution of $\zeta \coth\zeta = 1 + r/n$ and $K = \frac{e^{2\zeta}-1-2\zeta}{\sqrt{\zeta(e^{4\zeta}-1-4\zeta e^{2\zeta})}}$. This formula is an adaptation for multigraphs of Theorem 3 of [12] and appeared originally (for graphs) in [13].
7. Finally, when $2m/n - \log(n) \to +\infty$ as $n \to +\infty$, $\mathrm{Pr}_{[m,n]}(x_1 \sim \cdots \sim x_n) \sim \frac{1}{2^m}$ because almost all multigraphs are connected.

4.3 Asymptotics for a Two-Blocks Function

We now consider a function in the class of $x_1 \sim \cdots \sim x_p, x_{p+1} \sim \cdots \sim x_n$ (the block sizes are p and $n-p$), which has cardinality $2^{n-2}\frac{n!}{p!(n-p)!}$. We are again in the third range: $m \geq n-2$, i.e. $r \geq -2$. Theorem 3 gives the generating function as

$$\phi_{\mathbf{j}}(z) = \left[\frac{v^p}{p!}\right] C(4z, v) \cdot \left[\frac{w^{n-p}}{(n-p)!}\right] C(4z, w),$$

from which we readily obtain that

$$\mathrm{Pr}_{[m,n]}(f) = \frac{m!}{n^{2m}} \sum_{d=-1}^{r+1} \left[\frac{v^p}{p!}\right] C_d(v) \cdot \left[\frac{w^{n-p}}{(n-p)!}\right] C_{r-d}(w).$$

We now consider several cases. For simplicity we assume that, when r is large, it is equal to cn for a fixed positive value c.

1. Fixed excess r, and a single large part. In the range we are working in, p and d belong to a fixed, finite set. For some explicitly computable constant k_f,

$$\mathrm{Pr}_{[m,n]}(f) \sim k_f \,.\, n^{\frac{r+3}{2}-p}\, e^{-n}.$$

2. Fixed excess r, and two proportional large parts. By symmetry, we can assume that $p \leq n-p$. We have that

$$\mathrm{Pr}_{[m,n]}(f) \sim \frac{2\pi}{e^n n^{2n+2r}} (n-p)^{2n+\frac{3r}{2}} \left(\frac{p}{n-p}\right)^{2p} \sum_{d=-1}^{r+1} K_d\, K_{r-d} \left(\frac{p}{n-p}\right)^{\frac{3d}{2}}. \qquad (2)$$

Now assume for simplicity that $p = \gamma n$, then

$$\mathrm{Pr}_{[m,n]}(f) \sim k_f\, n^{-\frac{r+1}{2}}\, \beta^{2n}\, e^{-n} \qquad \textit{with} \qquad \beta = (1-\gamma)^{1-\gamma}\, \gamma^{\gamma}.$$

3. Fixed excess r, and two non-proportional large parts. In this case, the expression (2) is still valid, but now $p/(n-p) = o(1)$ i.e. $p = \varepsilon\, n$ with $\varepsilon = o(1)$. Then

$$\mathrm{Pr}_{[m,n]}(f) \sim k_f\, e^{-n}\, n^{\frac{r-1}{2}}\, \varepsilon^{n\varepsilon-1}\, (1-\varepsilon)^{(1-\varepsilon)n}.$$

A more precise evaluation of probabilities requires to know the order of growth of p w.r.t. n. E.g.,

   (a) $p = \sqrt{n}$: then $\varepsilon = n^{-1/2}$ and the probability of the function is of order $n^{-\frac{r}{2}+\frac{3}{4}}\, e^{-n-2\sqrt{n}}\, n^{-\sqrt{n}}$.
   (b) $p = \log n$: now $\varepsilon = \frac{\log n}{n}$, the probability is of order $\left(\frac{\log n}{n}\right)^{\log n - 1} n^{\frac{r+1}{2}}\, e^{-n}$.

4. Large excess r and a single large part. If $r = cn$ and p is fixed, we obtain for some explicitly computable constant k_f

$$\Pr_{[m,n]}(f) \sim \frac{k_f}{n^{p-1}} \left(\frac{(1+c)^c \cosh(\zeta)}{e^{1+c}(2\zeta)^c} \right)^n$$

   where $\zeta \coth \zeta = 1 + \frac{cn+1}{n-p}$.

5. Large excess r and two proportional large parts. If $r = cn$ and $p = \gamma n$,

$$\Pr_{[m,n]}(f) \sim \frac{k_f}{n} \left(\frac{\gamma^\gamma (1-\gamma)^{1-\gamma}(1+c)^{1+c}}{2^c e^{1+c}} g(a_0) \right)^n$$

   where k_f is a computable constant, and $g(a_0)$ is the unique maximum in $[0;1]$ of the function

$$g(a) = \left(\frac{\cosh(\zeta_1(a))}{1 + \frac{ac}{\gamma}} \right)^\gamma \left(\frac{\cosh(\zeta_2(a))}{1 + \frac{(1-a)c}{1-\gamma}} \right)^{1-\gamma} \left(\frac{\gamma}{\zeta_1(a)} \right)^{ac} \left(\frac{1-\gamma}{\zeta_2(a)} \right)^{(1-a)c}$$

   where the functions ζ_1 and ζ_2 are implicitly defined by $\zeta_1(a) \coth \zeta_1(a) = 1 + \frac{ac}{\gamma}$ and $\zeta_2(a) \coth \zeta_2(a) = 1 + \frac{(1-a)c}{1-\gamma}$.

4.4 Number of Blocks Proportional to n

A general approach via Theorem 3 seems difficult, so we assume a certain regularity: Let f denote a boolean function whose associated integer partition representation has the form $\mathbf{i}(f) = (0, \ldots, 0, n/g, 0, \ldots)$, with $g \geq 2$. Note that the corresponding multigraph has to have at least $m = (g-1) \cdot \frac{n}{g}$ edges. Thus, in contrary to the previously discussed cases, the excess is no more bounded from below, as $n \to \infty$. Such functions may now appear even close to the threshold $1/2$. In Proposition 5, we derive an exact result for those functions, and an asymptotic result in Proposition 6.

Using directly the definition $C(z, v) = \log M(z, v)$ we can show:

Proposition 5. *The number of expressions $N_{[m,n]}(f)$ with n variables and m clauses computing a function f with associated integer partition representation of the form $\mathbf{i}(f) = (0, \ldots, 0, n/g, 0, \ldots)$, i.e. n/g blocks of size g, is given by*

$$N_{[m,n]}(f) = 4^m (g!)^{\frac{n}{g}} [z^m] \Big(\sum_{j=1}^{g} \frac{(-1)^{j-1}}{j} e_{j,g-j}(z) \Big)^{\frac{n}{g}} \tag{3}$$

with

$$e_{j,n}(z) = \sum_{\substack{\sum_{\ell=1}^{j} k_\ell = n \\ k_\ell \geq 0}} \binom{n}{k_1, \ldots, k_j} \frac{\exp\left(\sum_{\ell=1}^{j} \frac{(k_\ell+1)^2 z}{2} \right)}{\prod_{r=1}^{j} (k_\ell + 1)!}.$$

For example, in the case $g = 2$ we get

$$\Pr_{[m,n]}(f) = \frac{1}{n^{2m}} \sum_{\ell=0}^{\frac{n}{2}} \binom{\frac{n}{2}}{\ell} (n+\ell)^m (-1)^{\frac{n}{2}-\ell},$$

and for $g = 3$

$$\Pr_{[m,n]}(f) = \frac{1}{n^{2m}} \sum_{\ell=0}^{\frac{n}{3}} \sum_{j=0}^{\ell} \binom{\frac{n}{3}}{\ell} \binom{\ell}{j} (\frac{n}{2} + \ell + 2j)^m (-3)^{\ell-j} 2^{\frac{n}{3}-\ell}.$$

It turns out that there is no qualitative difference between constant and large excess. The relevant quantity here is the distance from the minimal possible excess. Thus we start with small $g = 2, 3, \ldots$ and assume that $m = \frac{g-1}{g} \cdot n + \kappa_n$, with $\kappa_n \geq 0$. According to [4] the interesting range is $\frac{n}{2} + \Theta(n^{2/3})$. Hence we also assume $\kappa_n = O(n^{2/3})$.

The above expression for $N_{[m,n]}(f)$, Equation (3), is a fixed function $G(z)$ raised to a large power. Moreover, it is not hard to show that $G(z) = \sum_{\ell \geq g-1} a_\ell z^\ell$. Thus in case of constant κ_n of Equation (3) becomes a finite sum which can be computed explicitly (at least in principle).

For $\kappa_n \to \infty$ the saddle point method applies and we can compute $N_{[m,n]}(f)$ asymptotically, though the expressions quickly become messy as g grows. For $g = 2$ we obtain

Proposition 6. *The number of expressions $N_{[m,n]}(f)$ with n variables and m clauses computing a function f with associated integer partition representation of the form $\mathbf{i}(f) = (0, n/2, 0, 0, \ldots)$, i.e. $n/2$ blocks of size 2, is given by*

$$N_{[m,n]}(f) = \frac{2}{\sqrt{6\pi}} 4^{m+\frac{n}{4}} r_n^{-m+\frac{n}{2}} \exp\left(\frac{3nr_n}{4} + \frac{1}{48} n r_n^2 + O(n r_n^4) \right).$$

where r_n is the unique positive solution of $\frac{z(2e^z-1)}{e^z-1} = 1 + \frac{2\kappa_n}{n}$, and satisfies

$$r_n = \frac{4}{3} \cdot \frac{\kappa_n}{n} + \mathcal{O}\Big(\frac{\kappa_n^2}{n^2}\Big).$$

5 Discussion

We have analysed the probability of Boolean functions generated by random 2-Xor expressions. This is strongly related to the 2-Xor-SAT problem. For people working in SAT-solver design the structure of solutions of satisfiable expressions, which corresponds to the component structure of the associated multigraphs, is also important.

We derived expressions in terms of coefficients of generating functions for the probability of satisfiability in the critical region ($m \sim \frac{n}{2} + \Theta(n^{2/3})$) as well as a general expression for the probability of any function (Theorem 3). Unfortunately, this expression is too complicated to be used for an asymptotic analysis of general functions. So, we discussed several particular classes of functions: Single block functions are completely analyzed. The asymptotic probability very much depends on the range of the excess. For two block functions, the only missing case is that of two large components which are not proportional in size. All those functions are rather close to FALSE. Finally, functions on the other edge (close to TRUE) were studied and, under some regularity conditions on the block sizes, we were able to get the asymptotic probability.

What is missing is an asymptotic analysis of functions on the boundaries TRUE and FALSE having a more irregular component structure as well as the study of functions in the intermediate range.

Acknowledgments. We thank Hervé Daudé and Vlady Ravelomanana for fruitful discussions.

References

1. Creignou, N., Daudé, H.: Satisfiability threshold for random XOR-CNF formulas. Discrete Applied Mathematics **96-97** (1999) 41–53
2. Creignou, N., Daudé, H.: Smooth and sharp thresholds for random k-XOR-CNF satisfiability. Theor. Inform. Appl. **37**(2) (2003) 127–147
3. Creignou, N., Daudé, H., Dubois, O.: Approximating the satisfiability threshold for random k-XOR-formulas. Combin. Probab. Comput. **12**(2) (2003) 113–126
4. Creignou, N., Daudé, H.: Coarse and sharp transitions for random generalized satisfiability problems. In: Mathematics and Computer Science III. Trends Math. Birkhäuser, Basel (2004) 507–516
5. Creignou, N., Daudé, H., Egly, U.: Phase transition for random quantified XOR-formulas. J. Artif. Intell. Res. **29** (2007) 1–18
6. Daudé, H., Ravelomanana, V.: Random 2XorSat phase transition. Algorithmica **59**(1) (2011) 48–65
7. Pittel, B., Yeum, J.A.: How frequently is a system of 2-linear equations solvable? Electronic Journal of Combinatorics **17** (2010)
8. Janson, S., Knuth, D., Luczak, T., Pittel, B.: The birth of the giant component. Random Structures and Algorithms **4**(3) (1993) 233–358
9. Banderier, C., Flajolet, P., Schaeffer, G., Soria, M.: Random maps, coalescing saddles, singularity analysis, and Airy phenomena. Random Struct. Algorithms **19**(3-4) (2001) 194–246
10. Flajolet, P., Sedgewick, R.: Analytic combinatorics. Cambridge University Press, Cambridge (2009)
11. Bender, E.A., Canfield, E.R., McKay, B.D.: the asymptotic number of labeled connected graphs with a given number of vertices and edges. Random Structures and Algorithm **1** (1990) pp. 129–169
12. Pittel, B., Wormald, N.C.: Counting connected graphs inside-out. J. Comb. Theory, Ser. B **93**(2) (2005) 127–172
13. Bender, E.A., Canfield, E.R., McKay, B.D.: The asymptotic number of labeled connected graphs with a given number of vertices and edges. Random Struct. Algorithms **1**(2) (1990) 127–170