2-Xor revisited: satisfiability and probabilities of functions

Élie de Panafieu* Danièle Gardy† Bernhard Gittenberger‡ Markus Kuba§

November 25, 2015

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 revisit this classical problem by giving an alternative, explicit expression of this probability. We then consider a refinement of it, namely the probability that a random expression computes a specific Boolean function. The answers to both problems involve a description of 2-Xor expressions as multigraphs and use classical methods of analytic combinatorics by expressing probabilities through coefficients of generating functions.

Keywords: multigraph enumeration, probability of Boolean functions, satisfiability, 2-Xor expressions, asymptotics.

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, has given rise to numerous studies, most of them concentrating on the existence and location of a threshold from satisfiability to unsatisfiability as the ratio m/n grows. The literature in this direction is vast; for Xor-functions see e.g. [9, 10, 12, 11, 13].

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 such as Galton-Watson and Pólya trees, binary search trees, etc ([30, 6, 5, 34, 37, 24, 27, 29, 7, 25, 23, 26]).

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 the set of Boolean functions. It is only natural that we should wish to merge these two approaches: We set satisfiability problems into the framework of quantitative logic (this only requires choosing a suitable model of expressions) and ask for the probability of FALSE – this is the classical satisfiability problem – *and* for the probabilities of the other Boolean functions as well. This amounts to refining the satisfiable case and taking all the functions different from FALSE also into account. The set of Boolean expressions is then partitioned into subsets according to the (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

*COMPLEX NETWORKS, University of Paris 6. Supported by the ANR projects BOOLE (2009-13) and MAGNUM (2010-14), the P.H.C. Amadeus project (2013-14), the PEPS HYDrATA and the Austrian Science Fund (FWF) grant F5004.

†DAVID Laboratory, University of Versailles Saint Quentin en Yvelines. Part of the work of this author was done during a long-term visit at the Institute of Discrete Mathematics and Geometry of the TU Wien. Supported by the P.H.C. Amadeus project *Probabilities and tree representations for Boolean functions* (2013-14) and by the ANR project BOOLE (2009-13).

‡Institute of Discrete Mathematics and Geometry, TU Wien. Supported by the FWF (Austrian Science Foundation), Special Research Program F50, grant F5003-N15, and by the ÖAD, grant Amadée F01/2015.

§Supported by ÖAD, grant Amadée F01/2015.

of our reach for most classical satisfiability problems, there are some problems for which we may still hope to obtain a (partial) description of the probability distribution on the set of Boolean functions. The case of 2-Xor expressions is such a problem, and this paper is devoted to presenting our results in this domain.

Consider random 2-Xor-Sat instances with a large number n of variables, and m of clauses. Creignou and Daudé established that their limit probability of satisfiability goes from positive values to zero when the ratio m/n crosses $1/2$ (see [9]). They then proved that this threshold is coarse (*cf.* [11]). Further work by Daudé and Ravelomanana [14] and by Pittel and Yeum [32] led to a precise understanding of the transition in a window of size $n^{-1/3}$ around $1/2$.

The paper is organized as follows. We present in the next section 2-Xor expressions and the set of Boolean functions that they can represent. Then we give a modelization of these expressions 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 in Section 5.

A preliminary version of our work was presented at the conference Latin'14 [16].

2 Boolean Expressions and Functions and their Relations to Multigraphs

2.1 2-Xor Expressions and Boolean Functions

In this section we will lay out the framework of Boolean expressions which we will investigate. If x is a Boolean variable, we will denote by $\bar{x}$ its negation.

Definition 1. *Let $\{x_1, x_2, \ldots, x_n\}$ be a set of Boolean variables. A 2-Xor expression is a finite conjunction of clauses $l \oplus l'$, where l and l' are literals, i.e. they are elements of $\{x_1, x_2, \ldots, x_n, \bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\}$.*

The clauses as well as the literals within each clause are ordered (i.e. for instance that the clauses $x \oplus y$ and $y \oplus x$ are distinct). From a combinatorial point of view, an expression can be regarded as a sequence *of clauses where each clause is a pair of two literals. Neither the literals of a clause nor the clauses themselves need to be distinct.*

The set of all such expressions is denoted by $\mathcal{E}_n$.

We say that a 2-Xor expression *defines*, or *computes*, the corresonding Boolean function. We shall denote the number of clauses of an expression by m. 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 represented by a 2-Xor expression. We define $\mathcal{X}$ as the set of functions from $\{0,1\}^{\mathbb{N}}$ to $\{0,1\}$ which have at least one representation by a 2-Xor expression in $\bigcup_{n\geq 1} \mathcal{E}_n$. We also define, for each $n \geq 1$, the set $\mathcal{X}_n$ of functions in $\mathcal{X}$ such that there exists an expression in $\mathcal{E}_n$ representing the function. This implies that $\mathcal{X}_{n_1} \subset \mathcal{X}_{n_2}$ for $n_1 \leq n_2$, and that $\mathcal{X} = \cup_{n\geq 1} \mathcal{X}_n$.[1]

Consider now the expressions in $\mathcal{E}_n$. There there are $4n^2$ distinct clauses. We assume that the m clauses are drawn with a uniform probability (and hence with replacement). This framework allows us to define, for each m, a probability distribution on the set $\mathcal{X}_n$:

Definition 2. *Let $E_{m,n} = (4n^2)^m$ be the total number of expressions with m clauses on the variables x_1, …, x_n, and $E_{m,n}(f)$ denote the number of these expressions that compute f. Then, for a Boolean function $f \in \mathcal{X}_n$ we set $\Pr_{[m,n]}(f) = \frac{E_{m,n}(f)}{E_{m,n}}$.*

2.2 The Sets $\mathcal{X}_n$

Rewriting a clause $l_1 \oplus l_2$ as $l_1 \sim \bar{l}_2$ or $\bar{l}_1 \sim l_2$ (i.e., the literals l_1 and l_2 must take opposite values for the clause to evaluate to TRUE), and merging the clauses sharing a common variable, we see that the functions

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

we obtain can be written as a conjunction of equivalence relations on literals: [2]

$$(l_1 \sim \cdots \sim l_{p_1}) \wedge (l_{p_1+1} \sim \cdots \sim l_{p_2}) \wedge \cdots \wedge (l_{p_{r-1}+1} \sim \cdots \sim l_{p_r}).$$

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 $(x_1 \sim \bar{x}_3) \wedge (x_6 \sim x_5) \wedge (x_7 \sim x_7) \wedge (\bar{x}_2 \sim \bar{x}_3)$, or equivalently as $(x_1 \sim \bar{x}_2 \sim \bar{x}_3) \wedge (x_5 \sim x_6)$; furthermore this function partitions the set of Boolean variables $\{x_1, \ldots, x_7\}$ into the subsets $\{x_1, x_2, x_3\}$, $\{x_4\}$, $\{x_5, x_6\}$ and $\{x_7\}$.

If a clause inducing $l \sim \bar{l}$ appears, then the expression simply computes FALSE. In other words:

Proposition 1. *For any $n \geq 1$, the set $\mathcal{X}_n$ of Boolean functions on n variables, such that there exists at least one 2-Xor expression in $\mathcal{E}_n$ that computes the function, comprises exactly the function* FALSE *and those functions f that are specified as follows: Fix a set $Y = \{y_1, y_2, \ldots, y_n\}$ such that $y_i = x_i$ or $y_i = \bar{x}_i$, for all $i = 1, \ldots, n$, and partition the set Y into subsets. Then f attains the value* TRUE *if and only if for each block of the partition all the literals have the same value. A variable which appears in no clause of an expression computing the function, or only as $l \sim l$, is put into a singleton.*

Proof. Given a set of literals $p = \{l_1, \ldots, l_s\}$, let $\bar{p}$ denote the set where each literal is switched

$$\bar{p} = \{\bar{l}_1, \ldots, \bar{l}_s\}.$$

Let us first observe that if a satisfiable expression is specified, in the sens of the proposition, by the partition

$$Y = p_1 \uplus p_2 \uplus \cdots \uplus p_t,$$

where each variable appears in exactly one literal of Y, then it is also specified by the partition where any number of p_i is replaced by $\bar{p}_i$.

We prove the proposition by recurrence on the number of clauses m. For $m = 0$, the Boolean function computed is TRUE, and is specified by the partition

$$\{\{x_1\}, \{x_2\}, \ldots, \{x_n\}\}.$$

of $Y = \{x_1, \ldots, x_n\}$. Let us assume that the proposition is proven for a given m, and consider a 2-Xor expression with $m + 1$ clauses

$$E = \tilde{E} \wedge (l_1 \oplus l_2),$$

where $\tilde{E}$ is a 2-Xor expression with m clauses. If $\tilde{E}$ computes the Boolean function FALSE, then E also computes FALSE and the proposition holds. Otherwise, let

$$Y = p_1 \uplus p_2 \uplus \cdots \uplus p_t$$

denote the partition obtained by application of the proposition to the expression $\tilde{E}$. The last clause of E is $(l_1 \oplus l_2)$, which is equivalent with $l_1 \sim \bar{l}_2$ and is satisfied if and only if l_1 and $\bar{l}_2$ are assigned the same Boolean value. Without loss of generality, we can assume that l_1 belongs to Y. Otherwise, we just replace the set p_i from the partition that contains $\bar{l}_1$ with $\bar{p}_i$.

- If l_2 also belongs to p_i then, according to the proposition, $\tilde{E}$ is satisfied only if l_1 and l_2 take the same Boolean value, so the clause $(l_1 \oplus l_2)$ cannot be satisfied. Therefore, E is not satisfiable, so it computes the Boolean function FALSE.
- If $\bar{l}_2$ belongs to p_i, then the clause $l_1 \oplus l_2$ is satisfied by any assignment satisfying $\tilde{E}$, so E is satisfiable, and the partition built by the proposition for E is $Y = p_1 \uplus p_2 \uplus \cdots \uplus p_t$.
- Otherwise, there is a set p_j from P, distinct from p_i, that contains either l_2 or $\bar{l}_2$. Without loss of generality, we can assume that p_j contains $\bar{l}_2$. Otherwise, we simply replace p_j with $\bar{p}_j$. Then E is satisfiable. The corresponding partition is obtained from $(p_1, \ldots, p_t)$ by replacing the sets p_i and p_j with $p_i \cup p_j$. ■

[2] Note that the relation $\sim$ corresponds to an equivalence relation on the set of variables and therefore induces a partition on the set of variables. But as to the presence of negations, the formal structure is in fact a little bit richer than only a set with an equivalence relation.

We now define an equivalence relation on $\mathcal{X}_n$.

Definition 3. *Two Boolean functions f and g on n variables are* equivalent*, denoted as $f \equiv g$, if g can be obtained from f by permuting the variables and flipping some of the literals. We denote by $\mathcal{C}(f)$ the equivalence class of a function f.*

For example, for $n = 7$ the function f we have defined before is equivalent to the function $g = (x_3 \sim x_5 \sim x_2) \wedge (x_1 \sim \bar{x_6})$. It is easy to check that all the Boolean functions in $\mathcal{C}(f)$ have the same probability $\mathrm{Pr}_{[m,n]}(f)$.

Definition 4. *Let $f \in \mathcal{X}$; we say that a Boolean variable x is an* essential *variable of f if and only if $f|_{x=1} \neq f|_{x=0}$. We set $e(f)$ as the number of the essential variables of f.*

Remark 1. Although writing the constant functions TRUE and FALSE as 2-Xor expressions requires the use of (at least) one variable, these two functions have no essential variable: $e(\text{TRUE}) = e(\text{FALSE}) = 0$.

Note that $g \notin \mathcal{X}_{e(f)-1}$ for all g with $f \equiv g$. But there exists a function g with $f \equiv g$ such that $g \in \mathcal{X}_{e(f)}$. In our running example, $e(f) = 5$ and the essential variables are x_1, x_2, x_3, x_5 and x_6, so we can take, *e.g.*, $g = (x_3 \sim x_5 \sim x_2) \wedge (x_1 \sim \bar{x_6})$.

Again, with the exception of FALSE that forms a class by itself, the classes we have thus defined on $\mathcal{X}_n$ are in bijection with the partitions of the integer n; in our example the class of the function f partitions the integer 7 as $1 + 1 + 2 + 3$.

Notation 1. *Let $\mathcal{P}(n)$ denote the set of partitions of the integer n. For any $\mathbf{i} = (i_\ell)_{\ell \geq 1}$ in $\mathcal{P}(n)$, i_ℓ is the number of parts of size ℓ. Hence the size of $\mathbf{i}$ is $s(\mathbf{i}) := \sum_\ell \ell\, i_\ell = n$, and the total number of parts (or* blocks*) is $\xi(\mathbf{i}) := \sum_\ell i_\ell$. A partition of the type $(0, \ldots, 0, 1, 0, \ldots)$ with the single 1 in position n is denoted by $\mathbf{i_{max(n)}}$.*

We can now express a bijection between classes of Boolean functions and integer partitions.

Proposition 2. *Given an integer partition $\mathbf{i}$ of n, let $\mathcal{C}_\mathbf{i}$ denote the set of Boolean functions from $\mathcal{X}_n \setminus \{\text{FALSE}\}$ with i_ℓ blocks of size ℓ for all $\ell \geq 1$. Then $\{\mathcal{C}_\mathbf{i}\}_{\mathbf{i} \in \mathcal{P}(n)}$ is in bijection with the quotient of the set $\mathcal{X}_n \setminus \{\text{FALSE}\}$ by the equivalence relation "$\equiv$".*

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 $\mathbf{i} = \mathbf{i}(f)$.

Proof. Given a Boolean function f in $\mathcal{X}_n \setminus \{\text{FALSE}\}$, $\mathcal{C}(f)$ denotes the class of f for the equivalence relation "$\equiv$". Therefore, the set of distinct classes $\mathcal{C}(f)$ is in bijection with $(\mathcal{X}_n \setminus \{\text{FALSE}\})/\equiv$. Let $\mathbf{i}$ denote the integer partition matching the block composition of f. The demonstration of the proposition is over once we have proven $\mathcal{C}_\mathbf{i} = \mathcal{C}(f)$.

Let us write the block representation of f, defined in Proposition 1, as

$$\begin{aligned}
&\{\{l_{1,1}\}, \{l_{1,2}\}, \ldots, \{l_{1,i_1}\},\\
&\ \{l_{2,1}, l_{2,2}\}, \{l_{2,3}, l_{2,4}\}, \ldots, \{l_{2,2i_2-1}, l_{2,2i_2}\},\\
&\qquad\qquad \vdots\\
&\ \{l_{t,1}, \ldots, l_{t,t}\}, \ldots, \{l_{t,ti_t-(t-1)}, \ldots, l_{t,ti_t}\}, \ldots\},
\end{aligned}$$

where all $l_{i,j}$ are literals corresponding to distinct variables. Let g be a Boolean function in $\mathcal{C}_\mathbf{i}$, with block representation

$$\begin{aligned}
&\{\{\tilde{l}_{1,1}\}, \{\tilde{l}_{1,2}\}, \ldots, \{\tilde{l}_{1,i_1}\},\\
&\ \{\tilde{l}_{2,1}, \tilde{l}_{2,2}\}, \{\tilde{l}_{2,3}, \tilde{l}_{2,4}\}, \ldots, \{\tilde{l}_{2,2i_2-1}, \tilde{l}_{2,2i_2}\},\\
&\qquad\qquad \vdots\\
&\ \{\tilde{l}_{t,1}, \ldots, \tilde{l}_{t,t}\}, \ldots, \{\tilde{l}_{t,ti_t-(t-1)}, \ldots, \tilde{l}_{t,ti_t}\}, \ldots\}.
\end{aligned}$$

By flipping some of the literals and permuting the variables, the block representation of f can be sent to the block representation of g, so $f \equiv g$ and $\mathcal{C}_{\mathbf{i}}$ is a subset of $\mathcal{C}(f)$.

Reciprocally, let h denote a Boolean function in $\mathcal{C}(f)$. By definition, a block representation of h can be obtained from the block representation of f by flipping some literals and permuting the variables. Therefore, the block representation of h corresponds to the same integer partition $\mathbf{i}$ as f, so h belongs to $\mathcal{C}_{\mathbf{i}}$ and $\mathcal{C}(f)$ is a subset of $\mathcal{C}_{\mathbf{i}}$.

Since we have both $\mathcal{C}_{\mathbf{i}} \subset \mathcal{C}(f)$ and $\mathcal{C}(f) \subset \mathcal{C}_{\mathbf{i}}$, we conclude that those two sets are equal. ■

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 form $l \oplus \bar{l}$.

Proposition 3. *i) Set $p(n)$ as the number of integer partitions of n. Then the number of equivalence classes of computable Boolean functions is $p(n) + 1$.*

ii) The class $C_{\mathbf{i}}$ associated to an integer partition $\mathbf{i} = (i_\ell)$ has cardinality

$$|C_{\mathbf{i}}| = \frac{2^{n-\xi(\mathbf{i})}\, n!}{\prod_{\ell \geq 1} i_\ell!(\ell!)^{i_\ell}}. \tag{1}$$

Remark 2. As an aside, we mention that, as $n \to +\infty$ (see [22, p. 578]),

$$p(n) \sim \frac{1}{4n\sqrt{3}} \exp\left(\pi\sqrt{2n/3}\right).$$

Proof. The number of classes comes from the bijection between classes, with the exception of the one with FALSE, and integer partitions, hence we get i).

To prove ii), note that the number of partitions of the set of the n Boolean variables that lead to $\mathbf{i}$ is

$$\frac{n!}{\prod_{l=1}^{n}(l!)^{i_l} i_l!},$$

cf. [8, p. 205, Theorem B] or [1, Theorem 13.2].

Now observe that there are two possible polarities for each variable and hence 2^n choices. But in this way, each block of variables is counted twice, e.g. $x_1 \sim \bar{x_2} \sim x_3$ defines the same function as $\bar{x_1} \sim x_2 \sim \bar{x_3}$. Hence we have to divide by 2 for each block and therefore the cardinality of the equivalence class $C_{\mathbf{i}}$ is given by (1). ■

Remark 3. The factor $2^{n-\xi(\mathbf{i})}$ can also be arrived at as follows. Choose a variable in each block and then fix the polarities of the other variables in this block as equal or opposite to the chosen variable of the block. This gives $l-1$ decisions for a block of size l and thus in total a contribution of the multiplicative factor $2^{\sum_{l=2}^{n} i_l(l-1)}$.

2.3 2-Xor Expressions as Colored Multigraphs

In their seminal articles on the first cycle in an evolving graph and the birth of the giant component, Flajolet, Knuth and Pittel [20] and Janson, Knuth, Łuczak and Pittel [28] introduced the following notions.

The *multigraph process*, also known as the *uniform graph model*, produces a labelled multigraph G with n vertices and m edges by drawing independently and uniformly $2m$ vertices in $[1, n]$:

$$u_1, v_1, u_2, v_2, \ldots, u_m, v_m.$$

The set of vertices of G is $V(G) = [1, n]$ and its set of edges is

$$E(G) = \{\{u_1, v_1\}, \{u_2, v_2\}, \ldots, \{u_m, v_m\}\}.$$

Different drawings can lead to the same multigraph: The number of ordered sequences of vertices that correspond to a given multigraph G is denoted by $\mathrm{seqv}(G)$ and satisfies

$$\mathrm{seqv}(G) = |\{u_1, v_1, \ldots, u_m, v_m \in [1, n]^{2m} \mid E(G) = \{\{u_1, v_1\}, \ldots, \{u_m, v_m\}\}\}|.$$

A multigraph is *simple* if no edge contains twice the same vertex and all its edges are distinct. Therefore, it contains neither loops nor multiple edges. It follows that the number of sequences of vertices that correspond to a given simple multigraph G with m edges is

$$\mathrm{seqv}(G) = 2^m m!.$$

The *compensation factor* $\kappa(G)$ of a multigraph G is classically defined as

$$\kappa(G) = \frac{\mathrm{seqv}(G)}{2^m m!},$$

so a multigraph is simple if and only if its compensation factor is equal to 1.

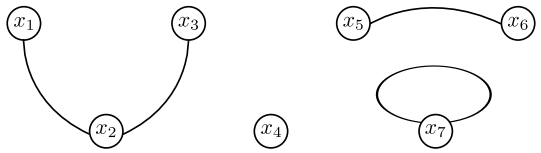

Figure 1: The multigraph underlying our running example.

For example, for $m = 4$ and $n = 7$ the drawings x_2, x_3, x_7, x_7, x_1, x_3, x_6, x_5 and x_7, x_7, x_1, x_3, x_3, x_2, x_5, x_6 both lead to the multigraph of Figure 1; indeed the number of ordered sequences leading to this multigraph is $4!\ 2^3 = 192$ and its compensation factor is $\frac{1}{2}$.

Fact 1. *Let $\mathcal{M}_{m,n}$ denote the set of multigraphs with n vertices and m edges. The probability for the multigraph process to produce a multigraph G among all multigraphs in $\mathcal{M}_{m,n}$ is proportional to its compensation factor $\kappa(G)$*

$$\mathbb{P}(G \mid G \in \mathcal{M}_{m,n}) = \frac{\kappa(G)}{\sum_{H \in \mathcal{M}_{m,n}} \kappa(H)}.$$

The *number* of multigraphs in a family $\mathcal{F}$ is defined as the sum of their compensation factors

$$\sum_{G \in \mathcal{F}} \kappa(G),$$

although this quantity might not be an integer. For example, the total number of multigraphs with n vertices and m edges is

$$M_{m,n} = \frac{n^{2m}}{2^m m!},$$

and the number of cubic multigraphs (*i.e.* multigraphs where all the vertices have degree 3) with $2r$ vertices is

$$\frac{(6r)!}{(3!)^{2r} 2^{3r} (3r)!},$$

because such multigraphs have $3r$ edges. If $\mathcal{F}$ contains only simple multigraphs, its number of multigraphs is equal to its cardinality.

Let $n(G)$ and $m(G)$ denote the number of vertices and number of edges of a multigraph G, respectively. The generating function corresponding to a family $\mathcal{F}$ of multigraphs is

$$\sum_{G \in \mathcal{F}} \kappa(G) z^{m(G)} \frac{v^{n(G)}}{n(G)!}.$$

For example, the generating function of all multigraphs is

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

As already observed by Janson, Knuth, Łuczak and Pittel[28], and Flajolet, Salvy and Schaeffer [21], a multigraph is a set of connected multigraphs, so the generating function for connected multigraphs is

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

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

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

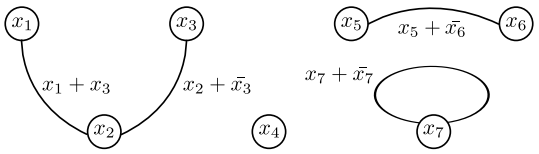

Figure 2: The colored multigraph for our running example.

Proposition 4. *The 2-Xor expressions are in bijection with multigraphs where loops are 4-colored and other edges are 8-colored. This bijection is such that, for all $f \in \mathcal{X}$ the number of connected components of the associated multigraph is $\xi(\mathbf{i}(f))$. Thus the function $M(8z, v)$ is the bi-exponential generating function for 2-Xor expressions, i.e.*

$$M(8z,v) = \sum_{n\geq 0}\sum_{m\geq 0} E_{m,n} \frac{z^n v^m}{n!m!}.$$

Proof. We first present the bijection between a 2-Xor expression of m clauses on n variables, and a colored multigraph on n vertices and with m edges.

- 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 has one of 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 has one of 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.

It is then an easy matter to check that the number of connected components of the multigraph is simply the number of parts in the integer partition associated with the function f computed by the expression.

We next turn to the generating function for 2-Xor expressions and start from the generating function for multigraphs

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

with v marking the vertices and z marking the edges and loops, and $M_{m,n}$ the number of multigraphs on n vertices and with m edges. Consider expressions built on n variables, and set $E_{m,n}$ as the number of such expressions with m clauses. Each vertex contributes a term e^{4z} for the loops: There are 4 possible colors; each vertex x also contributes a term $\prod_{y:x<y\leq n} e^{8z} = e^{8z(n-x)}$ for the edges to a different vertex:

There are 8 possible colors. We order the vertices so as not to count them twice. Taking into account all n vertices gives

$$\sum_m E_{m,n} \frac{z^m}{m!} = \prod_{s=1}^{n} e^{4z}\ e^{8z(n-s)} = e^{4n^2 z},$$

which in turn leads to an expression for the global generating function as

$$\sum_{m,n} E_{m,n} \frac{z^m}{m!}\ \frac{v^n}{n!} = \sum_n e^{4n^2\, z}\ \frac{v^n}{n!} = M(8z, v). \qquad \blacksquare$$

2.4 The Different Ranges

We shall not consider the whole range of values for the parameters n and m when studying the probabilities on $\mathcal{X}_n$, but restrict our investigations to the case where m and n are (roughly) proportional – which is the most interesting part as it includes the domain around the threshold – and set $m \sim \alpha n$ (α is usually assumed to be a constant). It is well known (see, e.g., [14]) 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 several regions according to the value of the ratio $\alpha = m/n$ when $m, n \to +\infty$:

- $\alpha < 1/2$. Here the probability of satisfiability is non-zero, but the attainable functions cannot have more than $n(1-\alpha)$ blocks.
- $\alpha = 1/2$. This is precisely the threshold range.
- $1/2 < \alpha < 1$. 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 the Set of 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 $f \in \mathcal{X}_n$) to $\Pr_{[m(n),n]}(f)$. We shall be interested in its limit when $n \to +\infty$ and f is a function of $\mathcal{X}$. First we will consider the case $f =$ FALSE (which is the usual satisfiability problem) and derive anew the probability of satisfiability in the critical window, before turning to general Boolean functions. We begin with some enumeration results on multigraphs that will be useful in the proofs of our results.

Remark 4. Note that the classical satisfiability problems as well as the above described extension are looking for the limit of the probability $\Pr_{[m(n),n]}(f)$, as $n \to \infty$. This raises the question whether the squence of distributions $\Pr_{[m(n),n]}$ defines a limiting distribution on the set $\mathcal{X}$. We do not know whether this is true or not, but our asymptotic results either concern the limit of the probabilities $\Pr_{[m(n),n]}(f)$ for some *a priori* given function f which is independent of n (lying in some $\mathcal{X}_{n_0}$; then the limit for $n \to \infty$ is taken) or a particular sequence of function which depends on n.

When looking into the literature of quantitative logic, the question for certain limiting probabilities often arises and is settled by means of the Drmota-Lalley-Woods theorem (see [22, p. 489] for the polynomial version and [18, Sec. 2.2.5] for the analytic version). In order to apply this theorem, one has to specify

the problem in terms of a system of functional equations which has certain technical properties, in particular it must not be linear. Usually, for each Boolean function one defines a generating associated with the expressions representing the Boolean function and sets up a sort of a recursive description of the Boolean function in terms of the other Boolean functions. If we do that for 2-Xor formulas, we get a linear system of functional equations, which is therefore not covered by the Drmota-Lalley-Woods framework. Despite linearity, the system is complicated to analyze, and so we decided to approach the problem through a bijection to certain classes of multigraphs and exploit the rich existing knowledge on multigraph generating functions.

3.1 Asymptotics for Multigraphs

3.1.1 Connected Multigraphs

Connected graphs with a large number of vertices have been counted for various ranges of number of edges. The first result is attributed to Cayley, who obtained in 1889 an exact formula for the number of unrooted trees by resolution of a recurrence (see [4, p. 51] for a historical discussion by Biggs, Lloyd and Wilson). Rényi [19] derived an asymptotic formula for the number of unicyclic graphs. Erdős and Rényi obtained in [19] the probability for a random graph with high density of edges to be connected. From their result follows an expression for the asymptotic number of connected graphs with n vertices and m edges when $m - n = \frac{1}{2}n(\log(n) + c)$ for any value c fixed or growing to infinity. Using generating functions, Wright [35] gave the asymptotic number of connected graphs for $m - n$ arbitrary but fixed, and also studied the case $m - n = o(n^{1/3})$ in [36]. Finally, Bender, Canfield, and McKay [3] obtained the asymptotic number of connected graphs for all $n, m-n \to \infty$. Their proof is based on a recursive formula derived by Wright. New proofs were proposed in [31] and [33].

For historical reasons, most of those results were only stated for simple graphs. In the following theorems, we summarize those results and adapt them to multigraphs.

Notation 2. *The number of connected multigraphs with n vertices and m edges is denoted by $C_{m,n}$.*

The exponential generating function of connected multigraphs with excess $r = m - n$ is denoted by

$$C_r(v) = \sum_{n\geq 0} C_{n+r,n} \frac{v^n}{n!}.$$

Theorem 1. *When the excess $r = m - n$ is fixed, then*

$$C_{m,n} \sim K_r n^{n+\frac{3r-1}{2}}, \tag{2}$$

where the value of K_r is

$$K_r = \begin{cases} 1 & \text{if } r = -1, \\ \frac{\sqrt{2\pi}}{4} & \text{if } r = 0, \\ \frac{\sqrt{2\pi}}{2^{3r/2}\Gamma(3r/2)}[v^{2r}] \log\left(\sum_{\ell\geq 0} \frac{(6\ell)!}{288^\ell(3\ell)!}\frac{v^{2\ell}}{(2\ell)!}\right) & \text{if } r > 0. \end{cases}$$

Remark 5. Note that the excess of a connected multigraph is always greater or equal to -1.

Proof.

- For $r = -1$, the connected component is an unrooted tree, $C_{-1}(v) = T(v) - T(v)^2/2$ where $T(v) = \sum_n n^{n-1}\frac{v^n}{n!}$ is the so-called tree function, and [22, p. 132]:

$$n![v^n]C_{-1}(v) = n^{n-2}.$$

- For $r = 0$, the connected component is unicyclic, $C_0(v) = \frac{1}{2}\log\frac{1}{1-T(v)}$ and (again from [22, p. 133]):

$$n![v^n]C_0(v) \sim \frac{1}{4}n^{n-1}\sqrt{2n\pi}.$$

- For $r \geq 1$, we follow the approach of Wright [35]. A *kernel* is a multigraph with minimum degree at least 3. Let us define the *deficiency* of a kernel of excess r with n vertices as $d = 2r - n$. it follows that the number of edges of a kernel is $m = 3r - d$. Let also $C_{r,d}^{(\geq 3)}$ denote the number of connected kernels of excess r and deficiency d. All connected multigraphs of excess $r \geq 1$ can be build from the connected kernels of excess r by replacing the edges with paths and the vertices with rooted trees, so

$$C_r(v) = \sum_{d=0}^{2r-1} \frac{C_{r,d}^{(\geq 3)}}{(2r-d)!} \cdot \frac{T(v)^{2r-d}}{(1-T(v))^{3r-d}},$$

which gives

$$C_{n+r,n} = n![v^n]C_r(v) = \sum_{d=0}^{2r-1} \frac{C_{r,d}^{(\geq 3)}}{(2r-d)!} \cdot [v^n] \frac{T(v)^{2r-d}}{(1-T(v))^{3r-d}}. \tag{3}$$

We must compute the coefficients $[v^n]\frac{T(v)^{2r-d}}{(1-T(v))^{3r-d}}$. We have, for any fixed positive integers a and b,

$$n![v^n]\frac{T(v)^a}{(1-T(v))^b} \sim \frac{2^{-b/2}}{\Gamma(b/2)}\, e^n\, n^{b/2-1}\, n!,$$

which is independent of a. When r is fixed, we see that, of the $2r$ terms in Equation (3), the one for $d = 0$ gives the dominant term and we get, also from [22, p. 134]:

$$n![v^n]C_r(v) \sim \frac{C_{r,0}^{(\geq 3)}}{(2r)!}\, \frac{\sqrt{2\pi}}{2^{3r/2}\,\Gamma(3r/2)}\, n^{n+\frac{3r-1}{2}}.$$

Finally, the constant $C_{r,0}^{(\geq 3)}$ is the number of connected cubic multigraphs (*i.e.* 3-regular multigraphs). Since there are $\frac{(6\ell)!}{(3!)^{2\ell}2^{3\ell}(3\ell)!}$ cubic multigraphs with 2ℓ vertices (see Section 2.3), the generating function of connected cubic multigraphs is

$$\sum_{\ell \geq 1} C_{\ell,0}^{(\geq 3)} \frac{v^{2\ell}}{(2\ell)!} = \log\left(\sum_{\ell \geq 0} \frac{(6\ell)!}{288^\ell (3\ell)!} \frac{v^{2\ell}}{(2\ell)!}\right),$$

and a coefficient extraction leads to

$$\frac{C_{r,0}^{(\geq 3)}}{(2r)!} = [v^{2r}] \log\left(\sum_{\ell \geq 0} \frac{(6\ell)!}{288^\ell (3\ell)!} \frac{v^{2\ell}}{(2\ell)!}\right). \qquad \blacksquare$$

When the excess r goes to infinity, non-cubic kernels cease to be negligible, and a different approach is needed to enumerate the connected multigraphs.

Theorem 2. *When $m-n$ goes to infinity and $\frac{2m}{n} - \log(n)$ tends towards a constant or $-\infty$, the asymptotic number of connected multigraphs is*

$$C_{m,n} = \sqrt{\frac{2(e^\lambda - 1 - \lambda)^2}{\lambda(e^{2\lambda} - 1 - 2\lambda e^\lambda)}}\, \frac{n^m}{\sqrt{2\pi n}}\, \frac{(2\sinh(\lambda/2))^n}{\lambda^m} \left(1 + \mathcal{O}\left((m-n)e^{-2m/n}\right)^{-1/2+\epsilon}\right)$$

for any $\epsilon > 0$, where the value λ is characterized by the relation

$$\frac{\lambda}{2}\coth\frac{\lambda}{2} = \frac{m}{n}.$$

Proof. This asymptotic expression has already been derived for simple graphs. Unfortunately, the corresponding proofs are too long to be reproduced and adapted here for multigraphs. Instead, we follow the proof from Pittel and Wormald [31]. and indicate the necessary changes in order to obtain the same result

for multigraphs. A proof based on analytic combinatorics is also available in [15, Theorem 5.1.3]. It is however restricted to the case where $2m/n$ tends toward a constant.

The proof starts with the enumeration of *cores*, which are multigraphs with minimum degree at least 2. Cores correspond to sequences of vertices

$$u_1, v_1, \ldots, u_m, v_m$$

where each vertex appears at least 2 times. The number of such sequences of length $2m$ with n vertices is

$$\sum_{\substack{d_1,\ldots,d_n \geq 2 \\ d_1+\cdots+d_n=2m}} \binom{2m}{d_1,\ldots,d_n} = (2m)!Q(n,m),$$

where the quantity $Q(n,m)$ is defined in [31, Equation (2.1)] by

$$Q(n,m) = \sum_{\substack{d_1,\ldots,d_n \geq 2 \\ d_1+\cdots+d_n=2m}} \prod_{j=1}^{n} \frac{1}{d_j!}.$$

Therefore, the number of cores with n vertices and m edges, defined as the sum of their compensation factors, is

$$\mathrm{Core}_{m,n} = \frac{(2m)!}{2^m m!} Q(n,m),$$

which replaces Equation (3.9) of [31, Theorem 8, p. 13]. Its asymptotic estimate, given in [31, Equation (3.11), p. 13] is now

$$\mathrm{Core}_{m,n} = (1 + \mathcal{O}((m-n)^{-1} + (m-n)^{1/2} n^{-1+\epsilon})) \frac{(2m-1)!! f(\lambda)^n}{\lambda^{2m}} \frac{1}{\sqrt{2\pi n c(1+\bar{\eta}-c)}}$$

where λ, f, c and $\bar{\eta}$ have the same definition as in [31].

The second step of the proof is the enumeration of cores that contain no isolated cycles. Let $\mathrm{Core}_{m,n}^{(\backslash \mathrm{cycle})}$ denote the number of such multigraphs with n vertices and m edges. The result is stated in [31, p. 4, Theorem 2] and its proof can be found in [31, Section 6]. It relies on the exponential generating function of simple undirected cycles

$$\sum_{\ell \geq 3} \frac{x^\ell}{2\ell} = -\frac{1}{2}\log(1-x) - \frac{x}{2} - \frac{x^2}{4}.$$

In multigraphs, a cycle might also have size 1 (a loop), or size 2 (a double edge), so we replace the previous generating function with

$$\sum_{\ell \geq 1} \frac{x^\ell}{2\ell} = -\frac{1}{2}\log(1-x)$$

and replace the function $h(x)$, defined in [31, p. 4, Equation (2.3)], by

$$h(x) = e^{-\sum_{\ell \geq 1} \frac{x^\ell}{2\ell}} = (1-x)^{1/2}.$$

Theorem 2 of [31, p. 4] becomes for multigraphs: "*when $m-n$ goes to infinity and $m = \mathcal{O}(n\log(n))$, then for any fixed $\epsilon > 0$, the number of cores with n vertices and m edges that contain no isolated cycles is*

$$\mathrm{Core}_{m,n}^{(\backslash \mathit{cycle})} = (1 + \mathcal{O}(n^{-1/2+\epsilon} + (m-n)^{-1})) h\left(\frac{\lambda}{e^\lambda - 1}\right) \mathrm{Core}_{m,n}$$

where λ is the unique positive root of $\frac{\lambda(e^\lambda - 1)}{e^\lambda - 1 - \lambda} = \frac{m}{n}$. "

The last ingredient of the proof is an observation from Erdős and Rényi, that when $m-n$ tends to infinity, almost all graphs or multigraphs that contain neither trees nor unicyclic components are connected. Therefore, $C_{m,n}$ is asymptotically equal to the number of such multigraphs. They correspond to the cores

without isolated cycle, where each vertex is replaced with a rooted tree. Their exact number is derived in [31, Equation (7.1)] and becomes for multigraphs

$$\sum_{\mu=1}^{n} \binom{n}{\mu} \mu n^{n-\mu-1} \operatorname{Core}^{(\setminus \text{cycle})}_{m-n+\mu,\mu}.$$

Borrowing the notation of [31], the summand is estimated in [31, Equation (7.2)] by combining [31, Theorem 2] and [31, Equation (3.11)], which we both have modified

$$\binom{n}{\mu} \mu n^{n-\mu-1} \operatorname{Core}^{(\setminus \text{cycle})}_{m-n+\mu,\mu} = (1+\mathcal{O}(\beta_1)) n^m F_n(y) \exp(nH(y,\lambda)).$$

The adaptation for multigraphs only recquires to change the definition of $F_n(y)$ and replace it with

$$F_n(y) = \frac{1}{2\pi n} \sqrt{\frac{(1-\sigma)y}{u(1+\bar{\eta}-2u/y)(1-y+\rho)}},$$

using again the notations u, c, λ, $\bar{\eta}$, ρ and σ of [31]. The rest of the proof is a Laplace method. The modification we made in the definition of F_n also impacts [31, p. 31, Equation (7.16)] which becomes

$$F_n(\bar{y}) = \frac{\sqrt{2}(e^{\bar{\lambda}} - 1 - \bar{\lambda})^{3/2}}{2\pi n \bar{\lambda} \sqrt{(e^{\bar{\lambda}}-1)^2 - \bar{\lambda}^2 e^{\bar{\lambda}}}}.$$

As a consequence, the definition of the value α of [31, p. 5, Theorem 3] is, for multigraphs,

$$\alpha = \sqrt{\frac{2(e^{\bar{\lambda}} - 1 - \bar{\lambda})^2}{\bar{\lambda}(e^{2\bar{\lambda}} - 1 - 2\bar{\lambda}e^{\bar{\lambda}})}}$$

while the other quantities of the theorem stay unchanged. ■

Remark that the value λ of the previous theorem is a constant only when $\frac{m}{n}$ is fixed.

As observed by Pittel and Wormald in [31], the asymptotic formula of the previous theorem also holds when $\frac{2m}{n} - \log(n)$ tends slowly towards infinity. However, we do not need this extension, because this range of m is already covered by the following theorem.

Theorem 3. *When both $m - n$ and $\frac{2m}{n} - \log(n)$ go to infinity, the asymptotic number of connected multigraphs becomes*

$$C_{m,n} \sim \frac{n^{2m}}{2^m m!}.$$

Proof. Erdős and Rényi proved in [19] that when $2m/n - \log(n)$ goes to infinity, a random multigraph with n vertices and m edges is connected with high probability. Therefore, the number of connected multigraphs is then asymptotically equal to the total number of multigraphs, $\frac{n^{2m}}{2^m m!}$. ■

3.1.2 Weighted Multigraphs

As recalled in the definition of the multigraph process, multigraphs are counted according to their compensation factor, meaning that the number of multigraphs in a family $\mathcal{F}$ is defined as the sum of their compensation factors $\sum_{G\in\mathcal{F}} \kappa(G)$. The proof of Theorems 4 and 5 require a refinement of this definition, involving the number of connected components of the multigraphs. Specifically, we now count the number of multigraphs with n vertices and m edges according to their compensation factor and a factor σ for each connected component

$$\sum_{G\in\mathcal{M}_{m,n}} \kappa(G) \sigma^{c(G)}$$

where σ is a positive real value and $c(G)$ denotes the number of components of G. Since the generating function of connected multigraphs is $\log M(z,v)$ and a multigraph is a set of connected multigraphs, the previous quantity can be expressed by a coefficient extraction

$$\sum_{G\in\mathcal{M}_{m,n}} \kappa(G)\sigma^{c(G)} = n![z^m v^n]e^{\sigma \log M(z,v)} = n![z^m v^n]M^\sigma(z,v).$$

We list asymptotic formulas for those values in the following lemma, which combines Theorems 8, 9 and 10 of [17]. The first part focuses on multigraphs with less edges than half the number of vertices. As proved by Erdős and Rényi, with high probability, they contain only trees and unicyclic components. The second part investigates the critical window where the number of edges is around half the number of vertices. In this range, connected components with fixed excess appear. Higher number of edges seem more technical to analyze. However, the probability of satisfiability of the corresponding 2-Xor formulas has already reached 0 almost surely, and its study is therefore less interesting.

Lemma 1. *Let σ denote a fixed positive value. When $\frac{m}{n}$ is in a fixed closed interval of $]0,1/2[$, then*

$$n![z^m v^n]M^\sigma(z,v) \sim \frac{n^{2m}}{2^m m!}\sigma^{n-m}\left(1-\frac{2m}{n}\right)^{\frac{1-\sigma}{2}}.$$

When $m = \frac{n}{2}(1+\mu n^{-1/3})$ and μ is bounded, then

$$n![z^m v^n]M^\sigma(z,v) \sim \frac{n^{2m}}{2^m m!}\sigma^{n-m}n^{\frac{\sigma-1}{6}}\sum_{r\geq 0}\sigma^r e_r^{(\sigma)}\sqrt{2\pi}A(3r+\sigma/2,\mu),$$

where the value of $e_r^{(\sigma)}$ is

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

and the function A is defined in [28, Lemma 3] by

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

Remark 6. In the rest of the paper, we will only need the cases $\sigma = 1$ and $\sigma = 1/2$.

The function $A(y,\mu)$ is a variation of the classical Airy function which has been thoroughly analyzed in [28, Lemma 3]. For example, as mentioned in [28, Equation (10.28)], for $y = 1$ it satisfies the relation

$$A(1,\mu) = e^{-\mu/12}\,\mathrm{Ai}(\mu^2/4),$$

and for $y = 0$, it holds that

$$A(0,\mu) = -\frac{1}{2}\mu e^{-\mu^3/12}\,\mathrm{Ai}(\mu^2/4) - e^{-\mu^3/12}\,\mathrm{Ai}'(\mu^2/4).$$

It is also close to the function defined in [2, Theorem 11] and [22, Theorem IX.16], denoted by G in the first one, and by S in the second one.

3.2 Probability of Satisfiability

The probability of satisfiability of a random 2-Xor expression has been studied by Creignou and Daudé [9, 11], Daudé and Ravelomanana [14] and Pittel and Yeum [32]. We derive anew their results to give a first application of the link between 2-Xor expressions and colored multigraphs.

Theorem 4. *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 limit for $n \to +\infty$ when $\frac{m}{n}$ is in a fixed closed interval of $]0, \frac{1}{2}[$ is

$$\left(1 - \frac{2m}{n}\right)^{1/4}.$$

When $m = \frac{n}{2}(1 + \mu n^{-1/3})$ and μ is bounded, this becomes

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

with the notations of Lemma 1.

Proof. To obtain the generating function 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 $M(8z, 2v)$ is the generating function associated with the 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 generating function 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 3) 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 generating function 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 before and in the critical window $m = n/2 + \mathcal{O}(n^{2/3})$, we use Lemma 1. ■

The link between the enumeration of 2-Xor expressions and of multigraphs and the knowledge of the asymptotic number of multigraphs can also be combined to investigate the probability for a satisfiable expression to be satisfied by an input.

Theorem 5. *The probability that an input (fixed or random) satisfies a random satisfiable expression with n variables, m clauses and excess $r = m - n$ is*

$$\frac{[z^m v^n]M(4z,2v)}{2^n[z^m v^n]\sqrt{M(4z,2v)}}.$$

When $\frac{m}{n}$ is in a closed interval of $]0, \frac{1}{2}[$, then this is asymptotically equivalent to

$$\frac{1}{2^m}\left(1 - \frac{2m}{n}\right)^{-1/4},$$

and it is

$$\frac{n^{1/12}}{2^m}\frac{1}{\sum_r 2^{-r}e_r^{(2)}A(3r + 1/4, \mu)}.$$

for $m = \frac{n}{2}(1 + \mu n^{-1/3})$ with μ bounded, using the notation of Lemma 1.

Proof. The probability that a random expression is satisfied by a random assignment is equal to the number of pairs {satisfiable expression, satisfying assignment}, divided by the number of satisfiable expressions and by the number 2^n of assignments. The exact value follows from the fact that the generating functions for the number of satisfiable expressions and for the number of pairs {satisfiable expression, satisfying assignment} are respectively $\sqrt{M(4z, 2v)}$ and $M(4z, 2v)$; the asymptotic approximations come again from Lemma 1. ■

3.3 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 5 the generating functions for all Boolean functions (except again FALSE), then use it to provide a general expression for the probability of a Boolean function in Theorem 6, 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 5. *Let f denote a Boolean function in $\mathcal{X}$ and $\mathbf{i}(f)$ the corresponding integer partition. Define $\phi_{\mathbf{i}(f)}(z)$ as the generating function for Boolean expressions that compute f:*

$$\phi_{\mathbf{i}(f)}(z) = \sum_m E_{m,n}(f)\frac{z^m}{m!}.$$

When $\mathbf{i} = \mathbf{i_{max}}(n)$, we set $\phi_n(z) := \phi_{\mathbf{i_{max}}(n)}(z)$. Then

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

Proof. A canonical representant of the class $\mathbf{i_{max}}(n)$ 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 6. *1. The probability that a random expression of m clauses on n variables computes the function $x_1 \sim \cdots \sim x_n$ is*

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

2. Let f be a function of $\mathcal{X}$, with $q = \sum_\ell \mathbf{i}(f)_\ell$, 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

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

where $C_r(v)$ denote the generating function of connected multigraphs of excess r, defined in Notation 2.

Proof. The probability $\Pr_{[m,n]}(f)$ that an expression of m clauses on n variables computes a function f is the quotient of the number of corresponding expressions divided by the total number of expressions

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

For $f = x_1 \sim \cdots \sim x_n$, we obtain the first part of the theorem by substitution of the expression of $\phi_{\mathbf{i}(f)}$, derived in Proposition 5. More generally, we have

$$\phi_{\mathbf{i}(f)}(z) = \prod_{\ell \geq 1} \left(\ell![v^\ell]C(4z, v)\right)^{i(f)_\ell}.$$

By definition, $i(f)_\ell$ is the number of blocks of f of size ℓ, so the previous equation can be rewritten

$$\phi_{\mathbf{i}(f)}(z) = \prod_{j=1}^{q} |B_j|![v^{|B_j|}]C(4z, v),$$

and

$$m![z^m]\phi_{\mathbf{i}(f)}(z) = m! \sum_{m_1+\cdots+m_q=m} \prod_{j=1}^{q} |B_j|![z^{m_j} v^{|B_j|}]C(4z, v). \tag{4}$$

The generating function $C(z, v)$ can be expanded with respect to the excesses

$$C(z, v) = \sum_{r \geq -1} z^r C_r(vz),$$

so

$$|B_j|![z^{m_j} v^{|B_j|}]C(4z, v) = 4^{m_j}|B_j|![v^{|B_j|}]C_{r_j}(v), \tag{5}$$

where $r_j = m_j - |B_j|$. We obtain the second part of the theorem by combination of Equations (4) and (5). ■

4 Explicit Probabilities

We now show on examples how Theorem 6 allows us to compute the asymptotic probability of a specific function. Attempting to give explicit results for each and every case that may appear is not realistic; rather we aim at giving the reader a feeling of the kind of results our method allows to obtain and the kind of technical tools we need for obtaining precise asymptotics.

We consider first a fixed Boolean function f and how its probability varies when $n \to +\infty$ (i.e. when we add non-essential variables), then turn to a family of 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, once it can be obtained, and see how it varies when $n, m \to +\infty$ with fixed ratio α.

Proposition 6. *Let $f \in \mathcal{X}_n$, with $e(f)$ being the number of its essential variables, and $\mathbf{i} = \mathbf{i}(f) = (i_1, i_2, \dots) = (n - e(f), i_2, \dots)$ its associated integer partition. Assume $m = \alpha n \geq n - \xi(\mathbf{i}(f))$; then*

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

Proof. Let $\mathbf{i} = \mathbf{i}(f)$ be an integer partition with $s(\mathbf{i}) = n$ and for all $\ell \geq 2$, let i_ℓ be fixed, independent of n. The number of expressions with n variables and m clauses that correspond to Boolean functions in $\mathcal{C}_\mathbf{i}$ is then (cf: Proposition 5)

$$n!m![z^m]\frac{e^{\mathbf{i}_1 2z}}{\mathbf{i}_1!} \prod_{\ell \geq 2} \frac{\phi_\ell^{\mathbf{i}_\ell}(z)}{\mathbf{i}_\ell!}.$$

We derive an asymptotic equivalent by the saddle point method for a *large power* scheme, assuming that $\alpha = \frac{m}{n}$ is bounded ([22, Th. VIII.8 p. 587]). We get

$$m!\,\frac{n!}{\mathbf{i}_1!}\left(\frac{2en}{m}\right)^m \frac{1}{\sqrt{2\pi m}} e^{-(s(\mathbf{i})-\mathbf{i}_1)m/n} \prod_{\ell\geq 2} \frac{\phi_\ell^{\mathbf{i}_\ell}\left(\frac{m}{2n}\right)}{\mathbf{i}_\ell!}(1+o(1)).$$

Using Stirling's formula, this can be rewritten as

$$\frac{n!}{\mathbf{i}_1!}(2n)^m e^{-(s(\mathbf{i})-\mathbf{i}_1)m/n} \prod_{\ell\geq 2} \frac{\phi_\ell^{\mathbf{i}_\ell}\left(\frac{m}{2n}\right)}{\mathbf{i}_\ell!}(1+o(1)).$$

By division by $|\mathcal{C}_\mathbf{i}| = 2^{n-\xi(\mathbf{i})}\frac{n!}{\prod_{\ell\geq 2}\mathbf{i}_\ell!(\ell!)^{\mathbf{i}_\ell}}$, we obtain the number of expressions that correspond to any given function in $\mathcal{C}_\mathbf{i}$:

$$2^{\xi(\mathbf{i})-n}(2n)^m e^{-(s(\mathbf{i})-\mathbf{i}_1)m/n} \prod_{\ell\geq 2}\left(\ell!\phi_\ell\left(\frac{m}{2n}\right)\right)^{\mathbf{i}_\ell}.$$

We finally divide by the number of (n,m)-expressions, $4^m n^{2m}$, to obtain the asymptotic probability that a random expression with n variables and m clauses corresponds to the given Boolean function f described by the integer partition $\mathbf{i}$:

$$\frac{e^{-(s(\mathbf{i})-i_1)m/n}}{(2n)^m} \prod_{\ell\geq 2}\left(\ell!\phi_\ell\left(\frac{m}{2n}\right)\right)^{\mathbf{i}_\ell}.$$

The final form comes from the fact that $s(\mathbf{i}) - i_1$ is precisely the number of essential variables of f. ■

4.2 Asymptotics for a single-block function

All Boolean variables are in a single block: We consider the class of $x_1 \sim ... \sim x_n$ and the range $m \geq n-1$. From Theorem 6 , we have

$$\Pr_{[m,n]}(x_1 \sim ... \sim x_n) = \frac{m!}{n^{2m}}.n![v^n]C_{m-n}(v).$$

We now specialize this according to the possible values for the excess $r = m - n$ and obtain the

Proposition 7. *1. For $r = -1$, we have $\Pr_{[m,n]}(x_1 \sim ... \sim x_n) = \frac{(n-1)!}{n^n} \sim \sqrt{\frac{2\pi}{n}}\, e^{-n}$.*

2. *For $r = 0$, we get $\Pr_{[m,n]}(x_1 \sim ... \sim x_n) \sim \frac{\pi}{2}\, e^{-n}$.*
3. *For $r \geq 1$ but still fixed, $\Pr_{[m,n]}(x_1 \sim ... \sim x_n) \sim C_r\, e^{-n} n^{r/2}$ where $c_r = \sqrt{2\pi}\, e^{-r}\, K_r$ with K_r as in Theorem 1.*
4. *For $r \to \infty$ and $r = o(\sqrt{n})$, $\Pr_{[m,n]}(x_1 \sim ... \sim x_n) \sim \sqrt{\frac{3}{2}}\frac{e^{r/2}}{(2\sqrt{3})^r} e^{-n}\left(\frac{n}{r}\right)^{r/2}$.*
5. *For $r = (\alpha-1)n$ with $\alpha > 1$, $\Pr_{[m,n]}(x_1 \sim ... \sim x_n) \sim K\left(\frac{\alpha^{\alpha-1}\cosh\zeta}{(2\zeta)^{\alpha-1}e^\alpha}\right)^n$ where $\zeta\coth\zeta = \alpha$ and $K = \sqrt{\alpha}\frac{e^{2\zeta}-1-2\zeta}{\sqrt{\zeta(e^{4\zeta}-1-4\zeta e^{2\zeta})}}$.*
6. *When $r \to +\infty$ and $2m/n - \log(n)$ is bounded, then*

$$\Pr_{[m,n]}(x_1 \sim ... \sim x_n) \sim \frac{K}{(2\zeta)^r}\left(\frac{\sinh\zeta}{\zeta}\right)^n \frac{(1+r/n)^{n+r+1/2}}{e^{n+r}}$$

with ζ the positive solution of $\zeta\coth\zeta = 1 + \frac{r}{n}$ and $K = \frac{e^{2\zeta}-1-2\zeta}{\sqrt{\zeta(e^{4\zeta}-1-4\zeta e^{2\zeta})}}$.

7. *Finally, when $2m/n - \log(n) \to +\infty$ as $n \to +\infty$, $\Pr_{[m,n]}(x_1 \sim ... \sim x_n) \sim \frac{1}{2^m}$ because almost all multigraphs are connected.*

Proof. We simply use the expressions for the coefficients of $C_r(v)$ given in Section 3.1. The first three cases come from Theorem 1, the next two cases are subcases of the sixth case which comes from Theorem 2, and the last case comes from Theorem 3. ■

4.3 Asymptotics for a two-blocks function

We consider a function in the class of $x_1 \sim ... \sim x_p$, $x_{p+1} \sim ... \sim x_n$ (the block sizes are p and $n-p$), which has cardinality $2^{n-2}\frac{n!}{p!(n-p)!}$. We are again in range C: $m \geq n-2$, i.e. $r \geq -2$. Theorem 6 gives the generating function as

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

from which we readily obtain that

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

Its asymptotics varies with the excess $r = m - n$, and the sizes of the two blocks. In the following propositions, we consider several cases, depending of the respective sizes of the blocks and the excess corresponding to the underlying multigraph. The proofs are then presented in Sections 4.3.2 and 4.3.1.

Proposition 8 (Fixed excess and a single large part)**.** *If p and d belong to some fixed, finite set which does not depend on n, then*

$$\Pr_{[m,n]}(f) \sim K_f \,.\, n^{\frac{r+3}{2}-p}\, e^{-n},$$

for some explicitly computable constant K_f.

Proposition 9 (Fixed excess r and two large parts)**.** *Assume that p and $n-p$ both tend to infinity, as $n \to \infty$. W.l.o.g. let $p \leq n-p$. Then we have*

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

for suitable constants K_j. Depending on the actual growth rate of p we can distinguish two cases:

1. *If $p = \gamma n$ for some constant $\gamma > 0$, then $p/(n-p) = \Theta(1)$ and*

$$\Pr_{[m,n]}(f) \sim K\, n^{-\frac{r+1}{2}}\, \beta^{2n}\, e^{-n} \qquad with \qquad \beta = (1-\gamma)^{1-\gamma}\, \gamma^{\gamma}.$$

2. *If $p = \varepsilon_n\, n$ with $\varepsilon_n = o(1)$, then $p/(n-p) = o(1)$ and*

$$\Pr_{[m,n]}(f) \sim K\, e^{-n}\, n^{\frac{r-1}{2}}\, \varepsilon^{n\varepsilon_n - 1}\, (1-\varepsilon_n)^{(1-\varepsilon_n)n}.$$

   A more precise evaluation of probabilities gives for instance

   (a) *If $p = \sqrt{n}$, then $\varepsilon_n = n^{-1/2}$ and the probability of the function has order $\frac{n^{-\frac{r}{2}+\frac{3}{4}}\, e^{-n-2\sqrt{n}}}{n^{\sqrt{n}}}$.*

   (b) *If $p = \log n$, then $\varepsilon_n = \frac{\log n}{n}$ and the probability is of order $\left(\frac{\log n}{n}\right)^{\log n - 1} n^{\frac{r+1}{2}}\, e^{-n}$.*

Proposition 10 (Large excess r)**.** *Assume that $r = cn$ for a fixed positive value c. Again, we distinguish two cases:*

1. **Single large part.** *If p is constant, then*

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

   for some explicitly computable constant K_f, where $\zeta \coth \zeta = 1 + \frac{cn+1}{n-p}$.

2. **Two proportional large parts.** *If $p = \gamma n$ and $r = cn$, then*

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

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

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

where $\zeta_{\frac{ac}{\gamma}}(a)\coth\zeta_{\frac{ac}{\gamma}}(a) = 1+\frac{ac}{\gamma}$ and $\zeta_{\frac{(1-a)c}{1-\gamma}}(a)\coth\zeta_{\frac{(1-a)c}{1-\gamma}}(a) = 1+\frac{(1-a)c}{1-\gamma}$.

Decomposing the two connected multigraphs according to excess gives the generating function for multigraphs with 2 connected components of respective number of vertices p and $n-p$:

$$\begin{aligned}\phi_{\mathbf{j}}(z) &= p![v^p]\sum_{r\geq -1}(4z)^r C_r(4zv)\cdot(n-p)![w^{n-p}]\sum_{s\geq -1}(4z)^s C_s(4zw)\\ &= p!(n-p)![v^p w^{n-p}]\sum_{r,s\geq -1}(4z)^{r+s}\,C_r(4zv)\,C_s(4zw)\end{aligned}$$

and

$$\begin{aligned}[z^m]\phi_{\mathbf{j}}(z) &= p!(n-p)![z^m v^p w^{n-p}]\sum_{r,s\geq -1}(4z)^{r+s}\,C_r(4zv)\,C_s(4zw)\\ &= 4^m\,p!(n-p)![v^p w^{n-p}]\sum_{r,s\geq -1, r+s+n=m} C_r(v)\,C_s(w)\\ &= 4^m\,p!(n-p)![v^p w^{n-p}]\sum_{r=-1}^{m-n+1} C_r(v)\,C_{m-n-r}(w)\\ &= 4^m\sum_{r=-1}^{m-n+1} p![v^p]C_r(v)\cdot(n-p)![w^{n-p}]C_{m-n-r}(w).\end{aligned}$$

Then

$$\begin{aligned}\Pr_{[m,n]}(f) &= \frac{m!}{4^m n^{2m}}[z^m]\phi_{\mathbf{j}}(z)\\ &= \frac{m!}{n^{2m}}\sum_{d=-1}^{r+1} p![v^p]C_d(v)\cdot(n-p)![w^{n-p}]C_{r-d}(w).\end{aligned}$$

4.3.1 Function with two blocks and fixed excess

Single large part We now present the proof of Proposition 8. In the range we are working in, p and d belong to a fixed, finite set; let us define

$$\gamma_{d,p} = p![v^p]C_d(v).$$

Then

$$\Pr_{[m,n]}(f) = \frac{m!}{n^{2m}}\sum_{d=-1}^{r+1}\gamma_{d,p}(n-p)![w^{n-p}]C_{r-d}(w)$$

and the asymptotic value of the coefficient $(n-p)![w^{n-p}]C_{r-d}(w)$ is given by Equation (2), with a suitable constant:

$$(n-p)![w^{n-p}]C_{r-d}(w) \quad\sim\quad K_{r-d}.\,n^{n-p+\frac{3(r-d)-1}{2}}.$$

We see that the dominant term of the sum $\sum_{d=-1}^{r+1}\gamma_{d,p}\,[w^{n-p}]\,C_{r-d}(w)$ will be obtained for $d=-1$, which gives, for some suitable constant K_f that can be explicitly computed

$$\Pr_{[m,n]}(f) \sim K_f\,.\,n^{\frac{r+3}{2}-p}\,e^{-n}.$$

Two large parts This paragraph contains the proof of Proposition 9. By symmetry, we can assume that $p \leq n-p$. Recall that

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

but now the coefficients $[v^p]\, C_d(v)$ and $[w^{n-p}]\, C_{r-d}$ can both be obtained from the expansion (2) (*p and $n-p$* are large); moreover we are dealing with a fixed number of terms:

$$\begin{aligned}
\Pr_{[m,n]}(f) &\sim \frac{m!}{n^{2m}} \sum_{d=-1}^{r+1} K_d\, p^{p+\frac{3d-1}{2}}\, K_{r-d}\, (n-p)^{n-p+\frac{3(r-d)-1}{2}} \\
&\sim \frac{m!}{n^{2m}}\, p^{p-\frac{1}{2}} (n-p)^{n-p+\frac{3r-1}{2}} \sum_{d=-1}^{r+1} K_d\, K_{r-d} \left(\frac{p}{n-p}\right)^{\frac{3d}{2}} \\
&\sim \sqrt{\frac{2\pi\, n}{p(n-p)}}\, \frac{e^{-n}}{n^{n+r}}\, (n-p)^{n+\frac{3r}{2}} \left(\frac{p}{n-p}\right)^{p} \sum_{d=-1}^{r+1} K_d\, K_{r-d} \left(\frac{p}{n-p}\right)^{\frac{3d}{2}}.
\end{aligned}$$

Now we have to find the behaviour of the sum in the above expression, and we see that there are two different cases:

1. If p and n are proportional, then $p/(n-p) = \Theta(1)$ (for simplification we set $p = \gamma n$ and assume γ is constant, but the sequel only requires that $\gamma = \Theta(1)$); all terms $\left(\frac{p}{n-p}\right)^{\frac{3d}{2}}$ contribute to a constant factor, and the sum itself is constant, hence for a suitable constant [3] K we have

$$\Pr_{[m,n]}(f) \sim K\, n^{\frac{r-1}{2}}\, \beta^n\, e^{-n} \qquad with \qquad \beta = (1-\gamma)^{1-\gamma}\, \gamma^{\gamma}.$$

2. If $p/(n-p) = o(1)$ i.e. $p = o(n)$, then the first term of the sum dominates: Up to a constant multiplicative factor, the whole sum is asymptotically equivalent to $\left(\frac{n-p}{p}\right)^{\frac{3}{2}}$. Setting $\varepsilon = p/n$ we get

$$\Pr_{[m,n]}(f) \sim K\, e^{-n}\, n^{\frac{r-1}{2}}\, \varepsilon^{n\varepsilon-1}\, (1-\varepsilon)^{(1-\varepsilon)n}.$$

4.3.2 Large excess

This section contains the proof of Proposition 10. Let $C_{n+r,n}$ denote the number of connected multigraphs with n vertices and excess r. For this proof, we rewrite the asymptotics of $C_{n+r,n}$ when $r \to \infty$ and $(r+n)e^{-2r/n} \to \infty$, already derived in Theorem 2, as

$$C_{n+r,n} = \frac{\alpha(\zeta_{\frac{r}{n}})}{\sqrt{2\pi}(2\zeta_{\frac{r}{n}})^r} \left(\frac{\cosh \zeta_{\frac{r}{n}}}{1+\frac{r}{n}}\right)^n n^{n+r-\frac{1}{2}} \left(1 + \mathcal{O}\left(re^{-2r/n}\right)^{-\frac{1}{2}+\epsilon}\right) \tag{6}$$

for any small $\epsilon > 0$, where $\zeta_{\frac{r}{n}} \coth \zeta_{\frac{r}{n}} = 1 + \frac{r}{n}$ and $\alpha(\zeta) = \frac{e^{2\zeta}-1-2\zeta}{\sqrt{\zeta(e^{4\zeta}-1-4\zeta e^{2\zeta})}}$.

We are interested here in the probability that a random 2-Xor expression with n variables and m clauses compute the Boolean function with two blocks of sizes γn and $(1-\gamma)n$

$$x_1 \sim \ldots \sim x_{\gamma n},\ x_{\gamma n+1} \sim \ldots \sim x_n.$$

[3] Here and in what follows, the constant denoted by K may vary and may depend on r – but it is always possible to get an explicit, though cumbersome, expression for it.

This probability can be expressed as

$$\begin{aligned}\Pr_{[m,n]}(2 \text{ blocks}) &= \frac{m!}{n^{2m}} \sum_{d=-1}^{r+1} C_{\gamma n+d,\gamma n} C_{(1-\gamma)n+r-d,(1-\gamma)n} \\ &= \frac{m!}{n^{2m}} \sum_{d=-1}^{r+1} A_d\end{aligned}$$

where $r = m - n$ is the global excess of the multigraphs representing the random expression and d (resp. $r - d$) the excess of its first (resp. second) connected component.

The main ingredient for the proof of Proposition 10 is the Laplace methode. It involves first a reduction to a problem of real analysis, then the analysis of a real function. Those steps are detailed in the next two paragraphs.

Reduction to a real analysis problem We make the assumption that the excess r increases proportionately to n, so $r = (\alpha - 1)n$ where $\alpha = \frac{m}{n} > 0$ is a constant. In that case,

$$\begin{aligned}\frac{m!}{n^{2m}} &= \frac{(n+r)!}{n^{2(n+r)}} \\ &\sim \frac{(n+r)^{n+r}\sqrt{2\pi(n+r)}}{n^{2(n+r)}e^{n+r}} \\ &\sim \frac{\left(1+\frac{r}{n}\right)^{n+r+1/2}\sqrt{2\pi n}}{n^{n+r}e^{n+r}} \\ &\sim \sqrt{2\pi}\,\frac{\alpha^{\alpha n+1/2}}{e^{\alpha n}} n^{-\alpha n+1/2}\end{aligned}$$

Let us summarize some notations

total number of vertices	n
size of the first and smallest block	$p = \gamma n$
size of the second block	$n - p = (1-\gamma)n$
total excess	$r = cn$
excess of the first block	$d = ar$
excess of the second block	$r - d = (1-a)r$

The expression of A_d is quite complicated, so, in order to avoid forgetting some terms in the product, we write them down in the following array

$C_{\gamma n+ar,\gamma n}$	$C_{(1-\gamma)n+(1-a)r,(1-\gamma)n}$
γn	$(1-\gamma)n$
ar	$(1-a)r$
$\alpha(\zeta_{\frac{ac}{\gamma}})$	$\alpha(\zeta_{\frac{(1-a)c}{1-\gamma}})$
$\cosh(\zeta_{\frac{ac}{\gamma}})^{\gamma n}$	$\cosh(\zeta_{\frac{(1-a)c}{1-\gamma}})^{(1-\gamma)n}$
$\gamma^{(\gamma+ac)n-1/2} n^{(\gamma+ac)n-1/2}$	$(1-\gamma)^{(1-\gamma+(1-a)c)n-1/2} n^{(1-\gamma+(1-a)c)n-1/2}$
$2^{acn}\zeta_{\frac{ac}{\gamma}}^{acn}$	$2^{(1-a)cn}\zeta_{\frac{(1-a)c}{1-\gamma}}^{(1-a)cn}$
$\left(1+\frac{ac}{\gamma}\right)^{\gamma n}$	$\left(1+\frac{(1-a)c}{1-\gamma}\right)^{(1-\gamma)n}$

.

Now write $A_d = C_{\gamma n+ar,\gamma n} C_{(1-\gamma)n+(1-a)r,(1-\gamma)n}$ as

$$\begin{aligned} A_d &\sim \frac{\left(\gamma^\gamma(1-\gamma)^{1-\gamma}\right)^n n^{(c+1)n-1}}{2\pi\sqrt{\gamma(1-\gamma)}2^{cn}}\alpha(\zeta_{\frac{ac}{\gamma}})\alpha(\zeta_{\frac{(1-a)c}{1-\gamma}}) \\ &\quad\times\left(\frac{\cosh(\zeta_{\frac{ac}{\gamma}})^\gamma \cosh(\zeta_{\frac{(1-a)c}{1-\gamma}})^{1-\gamma}}{\left(1+\frac{ac}{\gamma}\right)^\gamma\left(1+\frac{(1-a)c}{1-\gamma}\right)^{1-\gamma}}\left(\frac{\gamma}{\zeta_{\frac{ac}{\gamma}}}\right)^{ac}\left(\frac{1-\gamma}{\zeta_{\frac{(1-a)c}{1-\gamma}}}\right)^{(1-a)c}\right)^n \end{aligned}$$

which gives

$$\begin{aligned} \frac{m!}{n^{2m}}A_d &\sim \frac{\left(\gamma^\gamma(1-\gamma)^{1-\gamma}\right)^n (c+1)^{(c+1)n+\frac{1}{2}}}{\sqrt{2\pi n}\sqrt{\gamma(1-\gamma)}2^{cn}e^{(c+1)n}}\alpha(\zeta_{\frac{ac}{\gamma}})\alpha(\zeta_{\frac{(1-a)c}{1-\gamma}})g(a)^n \\ &\sim \sqrt{\frac{c+1}{\gamma(1-\gamma)}}\left(\frac{\gamma^\gamma(1-\gamma)^{1-\gamma}(c+1)^{(c+1)}}{2^c e^{c+1}}\right)^n \frac{\alpha(\zeta_{\frac{ac}{\gamma}})\alpha(\zeta_{\frac{(1-a)c}{1-\gamma}})}{\sqrt{2\pi n}}g(a)^n \end{aligned}$$

where

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

$$\zeta_{\frac{ac}{\gamma}}\coth\zeta_{\frac{ac}{\gamma}} = 1+\frac{a(\alpha-1)}{\gamma}, \qquad \alpha(\zeta_{\frac{ac}{\gamma}}) = \frac{e^{2\zeta_{\frac{ac}{\gamma}}}-1-2\zeta_{\frac{ac}{\gamma}}}{\sqrt{\zeta_{\frac{ac}{\gamma}}(e^{4\zeta_{\frac{ac}{\gamma}}}-1-4\zeta_{\frac{ac}{\gamma}}e^{2\zeta_{\frac{ac}{\gamma}}})}},$$

$$\zeta_{\frac{(1-a)c}{1-\gamma}}\coth\zeta_{\frac{(1-a)c}{1-\gamma}} = 1+\frac{(1-a)c}{1-\gamma}, \qquad \alpha(\zeta_{\frac{(1-a)c}{1-\gamma}}) = \frac{e^{2\zeta_{\frac{(1-a)c}{1-\gamma}}}-1-2\zeta_{\frac{(1-a)c}{1-\gamma}}}{\sqrt{\zeta_{\frac{(1-a)c}{1-\gamma}}(e^{4\zeta_{\frac{(1-a)c}{1-\gamma}}}-1-4\zeta_{\frac{(1-a)c}{1-\gamma}}e^{2\zeta_{\frac{(1-a)c}{1-\gamma}}})}}.$$

We will see in the next paragraph that the dominant part of the sum $\sum_{d=-1}^{r+1} A_d$ is reached for a compact range of a included in $]0,1[$. This justifies the use of the asymptotic formula 6. Furthermore, the error term of A_d

$$\left(1+\mathcal{O}\left(are^{-ac/\gamma}\right)^{-\frac{1}{2}+\epsilon}\right)\left(1+\mathcal{O}\left((1-a)re^{-(1-a)c/(1-\gamma)}\right)^{-\frac{1}{2}+\epsilon}\right)$$

becomes uniform in a, so

$$\Pr_{[m,n]}(2\text{ blocks}) \sim \sqrt{\frac{c+1}{\gamma(1-\gamma)}}\left(\frac{\gamma^\gamma(1-\gamma)^{1-\gamma}(c+1)^{c+1}}{2^c e^{c+1}}\right)^n \frac{1}{\sqrt{2\pi n}}\sum_{d=0}^{r}\alpha(\zeta_{\frac{ac}{\gamma}})\alpha(\zeta_{\frac{(1-a)c}{1-\gamma}})g\left(\frac{d}{r}\right)^n.$$

Analysis of $g(a)$ We prove here that $g(a)$ has a unique maximum a_0 in $[0,1]$ such that $0<a_0<1$. To do so, we use the concavity of $\log(g(a))$. The *Laplace's method for sums* described in [22] p.761 then leads to

$$\sum_{d=0}^{r}\alpha(\zeta_{\frac{ac}{\gamma}})\alpha(\zeta_{\frac{(1-a)c}{1-\gamma}})g\left(\frac{d}{r}\right)^n \sim \sqrt{\frac{2\pi}{\lambda n}}\alpha(\zeta_{\frac{a_0c}{\gamma}})\alpha(\zeta_{\frac{(1-a_0)c}{1-\gamma}})g(a_0)^n$$

where $\lambda = -\frac{g''(a_0)}{g(a_0)}$, so

$$\Pr_{[m,n]}(2\text{ blocks}) \sim \sqrt{\frac{c+1}{\gamma(1-\gamma)\lambda}}\left(\frac{\gamma^\gamma(1-\gamma)^{1-\gamma}(c+1)^{c+1}}{2^c e^{c+1}}\right)^n \frac{\alpha(\zeta_{\frac{ac}{\gamma}})(a_0)\alpha(\zeta_{\frac{(1-a)c}{1-\gamma}})(a_0)}{n}g(a_0)^n.$$

The proof of the asymptotics is now reduced to a real analysis problem: Proving that

$$\begin{aligned} g(a) &= \left(\frac{\cosh(\zeta_{\frac{ac}{\gamma}})}{1+\frac{ac}{\gamma}}\right)^{\gamma}\left(\frac{\cosh(\zeta_{\frac{(1-a)c}{1-\gamma}})}{1+\frac{(1-a)c}{1-\gamma}}\right)^{1-\gamma}\left(\frac{\gamma}{\zeta_{\frac{ac}{\gamma}}}\right)^{ac}\left(\frac{1-\gamma}{\zeta_{\frac{(1-a)c}{1-\gamma}}}\right)^{(1-a)c} \\ &= \left(\frac{\cosh\zeta_{\frac{ac}{\gamma}}}{\zeta_{\frac{ac}{\gamma}}^{x_1}}\frac{\gamma^{x_1}}{1+x_1}\right)^{\gamma}\left(\frac{\cosh\zeta_{\frac{(1-a)c}{1-\gamma}}}{\zeta_{\frac{(1-a)c}{1-\gamma}}^{x_2}}\frac{(1-\gamma)^{x_2}}{1+x_2}\right)^{1-\gamma}, \end{aligned}$$

where $x_1 = \frac{ac}{\gamma}$ and $x_2 = \frac{(1-a)c}{1-\gamma}$, has a unique maximum in the interior of $]0,1[$ for all $c > 0$ and $\gamma \in]0, 1/2]$. Let $\zeta(x)$ be defined implicitly as

$$\zeta\coth\zeta = 1 + x,$$

then

$$\begin{aligned} \frac{\zeta'}{\zeta} &= \frac{1}{\zeta^2 - x(1+x)}, \\ \zeta'\tanh\zeta &= \frac{\zeta^2}{(\zeta^2 - x(1+x))(1+x)}, \end{aligned}$$

so

$$\begin{aligned} \frac{d}{dx}\log\left(\frac{\cosh\zeta}{\zeta^x}\frac{\gamma^x}{1+x}\right) &= \zeta'\tanh(\zeta) - x\frac{\zeta'}{\zeta} - \frac{1}{1+x} + \log(\gamma) - \log(\zeta) \\ &= \frac{\zeta^2}{(\zeta^2 - x(1+x))(1+x)} - \frac{x}{\zeta^2 - x(1+x)} - \frac{1}{1+x} + \log\left(\frac{\gamma}{\zeta}\right) \\ &= \frac{\zeta^2 - x(1+x)}{(\zeta^2 - x(1+x))(1+x)} - \frac{1}{1+x} + \log\left(\frac{\gamma}{\zeta}\right) \\ &= \log\left(\frac{\gamma}{\zeta}\right) \end{aligned}$$

and

$$\frac{d}{dx}\log\left(\frac{\cosh\zeta}{\zeta^x}\frac{(1-\gamma)^x}{1+x}\right) = \log\left(\frac{1-\gamma}{\zeta}\right).$$

Therefore,

$$\begin{aligned} \frac{d}{da}\log(g(a)) &= \gamma\left(\frac{d}{da}x_1\right)\frac{d}{dx_1}\log\left(\frac{\cosh\zeta_{\frac{ac}{\gamma}}}{\zeta_{\frac{ac}{\gamma}}^{x_1}}\frac{\gamma^{x_1}}{1+x_1}\right) \\ &\qquad + (1-\gamma)\left(\frac{d}{da}x_2\right)\frac{d}{dx_2}\log\left(\frac{\cosh\zeta_{\frac{(1-a)c}{1-\gamma}}}{\zeta_{\frac{(1-a)c}{1-\gamma}}^{x_2}}\frac{(1-\gamma)^{x_2}}{1+x_2}\right) \\ &= c\log\left(\frac{\gamma}{\zeta_{\frac{ac}{\gamma}}}\right) - c\log\left(\frac{1-\gamma}{\zeta_{\frac{(1-a)c}{1-\gamma}}}\right) \\ &= c\log\left(\frac{\gamma}{1-\gamma}\frac{\zeta_{\frac{(1-a)c}{1-\gamma}}}{\zeta_{\frac{ac}{\gamma}}}\right) \end{aligned}$$

and

$$\begin{aligned}\frac{1}{c}\frac{d^2}{(da)^2}\log(g(a)) &= \frac{d}{da}\log\left(\frac{\zeta_{\frac{(1-a)c}{1-\gamma}}}{\zeta_{\frac{ac}{\gamma}}}\right)\\ &= \left(\frac{d}{da}x_2\right)\frac{\zeta'_{\frac{(1-a)c}{1-\gamma}}}{\zeta_{\frac{(1-a)c}{1-\gamma}}} - \left(\frac{d}{da}x_1\right)\frac{\zeta'_{\frac{ac}{\gamma}}}{\zeta_{\frac{ac}{\gamma}}}\\ &= -\frac{c}{1-\gamma}\frac{1}{\zeta^2_{\frac{(1-a)c}{1-\gamma}} - x_2(1+x_2)} - \frac{c}{\gamma}\frac{1}{\zeta^2_{\frac{ac}{\gamma}} - x_1(1+x_1)}\end{aligned}$$

which is negative because for all $x > 0$,

$$\zeta(x) > \sqrt{x(1+x)}.$$

Therefore, $\frac{d}{da}\log(g(a))$ is decreasing on $]0,1[$. Let us summarize some values:

a	0	γ	1
$x_1 = \frac{ac}{\gamma}$	0	c	$\frac{c}{\gamma}$
$x_2 = \frac{(1-a)c}{1-\gamma}$	$\frac{c}{1-\gamma}$	c	0
$\zeta_{\frac{ac}{\gamma}}$	0	$\zeta(c)$	$\zeta\left(\frac{c}{\gamma}\right)$
$\zeta_{\frac{(1-a)c}{1-\gamma}}$	$\zeta\left(\frac{c}{1-\gamma}\right)$	$\zeta(c)$	0
$\frac{\zeta_{\frac{(1-a)c}{1-\gamma}}}{\zeta_{\frac{ac}{\gamma}}}$	$+\infty$		$-\infty$
$\frac{d}{da}\log(g(a))$	$+\infty$		$-\infty$

so $\frac{d}{da}\log(g(a))$ has a zero on $]0,1[$, and $g(a)$ has a unique maximum in $]0,1[$.

4.4 Number of blocks proportional to n

A general approach via Theorem 6 seems difficult, so we assume a certain regularity: Let f denote a Boolean function such that the associated integer partition is of the form $\mathbf{i}(f) = (0,\dots,0,n/g,0,\dots)$, 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 $r = -\frac{n}{g}$ is no more bounded from below as $n \to \infty$. Such functions may now appear even close to the threshold $1/2$. In Proposition 11, we derive an exact result for those functions; an asymptotic result is stated in Proposition 12.

Proposition 11. *The number of expressions $E_{m,n}(f)$ with n variables and m clauses computing a function f with associated integer partition representation of the form $\mathbf{i}(f) = (0,\dots,0,n/g,0,\dots)$, i.e. n/g blocks of size g, is given by*

$$E_{m,n}(f) = m!4^m(g!)^{\frac{n}{g}}[z^m]\Big(\sum_{j=1}^{g}\frac{(-1)^{j-1}}{j}e_{j,g-j}(z)\Big)^{\frac{n}{g}} \tag{7}$$

with

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

Remark 7. One might be tempted to use again Theorem 6. For Boolean functions f having associated integer partition of the form $\mathbf{i}(f) = (0,\dots,0,n/g,0,\dots)$ with $g \geq 2$ this yields

$$E_{m,n}(f) = m!4^m \sum_{\substack{\sum_{k=1}^{q} r_k = m-n\\ r_k \geq -1}} |B_1|!\cdots|B_q|!\left[v^{|B_1|}\dots v^{|B_q|}\right]\prod_{j=1}^{q}C_{r_j}(v_j),$$

where $B_1, \dots B_q$ are the blocks of the set partition, or equivalently the components of the Boolean function, with respective excesses $r_1, \dots, r_q$. Here the number of blocks is $q = \frac{n}{g}$ and all of them have size g; hence

$$E_{m,n}(f) = 4^m (g!)^{\frac{n}{g}} \sum_{\substack{\sum_{k=1}^q r_k = m-n \\ r_k \geq -1}} \prod_{j=1}^{\frac{n}{g}} [v^g] C_{r_j}(v).$$

However, it seems to get enough information on $C_{r_j}(v)$ to derive expression (7) from this formula.

Proof. Instead of analyzing the coefficients $C_r(v)$ of $C(z,v)$ we use directly the relation $C(z,v) = \log M(z,v)$. Since $\mathbf{i}(f) = (0, \dots, 0, n/g, 0, \dots)$, with $g \geq 2$, we have

$$\begin{aligned} E_{m,n}(f) &= m![z^m]\phi_{\mathbf{i}}(z) \\ &= m![z^m] \prod_{\ell \geq 1} \big(\ell![v^\ell] C(4z,v)\big)^{i_\ell} \\ &= m! 4^m (g!)^{\frac{n}{g}} [z^m] \Big([v^g] \log M(z,v) \Big)^{\frac{n}{g}}. \end{aligned}$$

Let

$$\hat{M}(z,v) = (M(z,v) - 1)/v = \sum_{n \geq 0} e^{\frac{(n+1)^2 z}{2}} \frac{v^n}{(n+1)!},$$

such that

$$\log M(z,v) = \sum_{j \geq 1} \frac{(-1)^{j-1}}{j} v^j \hat{M}^j(z,v).$$

We get

$$E_{m,n}(f) = m![z^m] \prod_{\ell \geq 1} \big([\frac{v^\ell}{\ell!}] C(4z,v)\big)^{i_\ell} = m! 4^m (g!)^{\frac{n}{g}} [z^m] \Big([v^g] \log M(z,v) \Big)^{\frac{n}{g}}$$

$$= m! 4^m (g!)^{\frac{n}{g}} [z^m] \Big(\sum_{j=1}^{g} \frac{(-1)^{j-1}}{j} [v^{g-j}] \hat{M}^j(z,v) \Big)^{\frac{n}{g}}.$$

We can expand $\hat{M}^j(z,v)$ in terms of the functions $e_{j,n}(z)$ as defined above using the multinomial theorem:

$$\hat{M}^j(z,v) = \left(\sum_{n \geq 0} e^{\frac{(n+1)^2 z}{2}} \frac{v^n}{(n+1)!} \right)^j = \sum_{n \geq 0} e_{j,n}(z) v^n.$$

Extraction of coefficients then directly leads to the stated result. ∎

Corollary 1. *Under the assumptions of Proposition 11, in the case $g = 2$ we get*

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

and for $g = 3$

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

Proof. Using Proposition 11, Equation (7), we obtain first

$$\begin{aligned} E_{m,n}(f) &= m! 4^m (2!)^{\frac{n}{2}} [z^m] \Big(\frac{1}{2} e^{2z} - \frac{1}{2} e^z \Big)^{\frac{n}{2}} = m! 4^m [z^m] \Big(e^{2z} - e^z \Big)^{\frac{n}{2}} \\ &= m! 4^m [z^m] e^{\frac{zn}{2}} \Big(e^z - 1 \Big)^{\frac{n}{2}}. \end{aligned}$$

The expansion of $\left(e^z-1\right)^{\frac{n}{2}}$ by the binomial theorem and the extraction of coefficients leads then to the stated result after dividing by the total number of expressions $(4n^2)^m$. We proceed for $g=3$ in a similar way:

$$E_{m,n}(f) = m!4^m(3!)^{\frac{n}{3}}[z^m]\left(\frac{1}{6}e^{\frac{9z}{2}} - \frac{1}{2}e^{\frac{5z}{2}} + \frac{1}{3}e^{\frac{3z}{2}}\right)^{\frac{n}{3}}.$$

In order to extract coefficients we use

$$\left(\frac{1}{6}e^{\frac{9z}{2}} - \frac{1}{2}e^{\frac{5z}{2}} + \frac{1}{3}e^{\frac{3z}{2}}\right)^{\frac{n}{3}} = e^{\frac{nz}{2}}\left(\frac{1}{6}e^{3z} - \frac{1}{2}e^{z} + \frac{1}{3}\right)^{\frac{n}{3}},$$

and expand twice using the binomial theorem. This leads to the stated result after a few elementary computations. ■

When considering asymptotics, we observed in the Sections 4.2 and 4.3 that the asymptotic behaviour is different depending on the fact whether the excess is constant or large, *i.e.* tending to infinity. For functions with two blocks there are also several phases in the case of large excess. But this observation is misleading, because in fact is not the excess but rather the distance from the minimal possible excess which determines the behaviour. In the case considered in this section, we will therefore write $m = \frac{g-1}{g}\cdot n + \kappa_n$, with $\kappa_n \geq 0$, because the minimal excess is $-n/g$. Furthermore, it turns out that there is no qualitative difference between constant and large κ_n in the sense that both cases can be covered by one single formula. This holds, however, only up to the interesting range, which has been shown to be $\kappa_n = \Theta(n^{2/3})$ in [11].

The expression for $E_{m,n}(f)$ given in Equation (7) that appears is a fixed function $G(z) = [v^g]\log M(z,v)$ raised to a large power n/g; e.g., for $g=2$ we have $G(z) = \frac{1}{2}e^{2z} - \frac{1}{2}e^{z}$ and for $g=3$ we have $G(z) = \frac{1}{6}e^{\frac{9z}{2}} - \frac{1}{2}e^{\frac{5z}{2}} + \frac{1}{3}e^{\frac{3z}{2}}$. By definition of $\log M(z,v) = \sum_{r\geq -1} z^r C_r(zv)$, the function $G(z)$ is of the form $G(z) = \sum_{\ell \geq g-1} a_\ell z^\ell$ for certain coefficients a_ℓ. Thus, in case of constant κ_n, Equation (7) involves a sum with a bounded range depending only on κ_n:

$$\begin{aligned} E_{m,n}(f) &= m!4^m(g!)^{\frac{n}{g}}[z^m]G(z)^{\frac{n}{g}} = m!4^m(g!)^{\frac{n}{g}}[z^{\frac{g-1}{g}\cdot n+\kappa_n}]\left(\sum_{\ell\geq g-1} a_\ell z^\ell\right)^{\frac{n}{g}} \\ &= m!4^m(g!)^{\frac{n}{g}}[z^{\kappa_n}]\left(\sum_{\ell\geq 0}\tilde{a}_\ell z^\ell\right)^{\frac{n}{g}}, \end{aligned}$$

with $\tilde{a}_\ell = a_{g-1+\ell}$ for $\ell \geq 0$. Using

$$\left(\sum_{\ell\geq 0}\tilde{a}_\ell z^\ell\right)^{\frac{n}{g}} = \sum_{i\geq 0} z^i \sum_{\substack{k_j\geq 0 \\ \sum_{j=1}^{\frac{n}{g}} k_j = i}} \binom{n/g}{k_1,\ldots,k_{n/g}} \prod_{s=1}^{n/g}\tilde{a}_{k_s}$$

we get

$$E_{m,n}(f) = m!4^m(g!)^{\frac{n}{g}} \sum_{\substack{k_j\geq 0 \\ \sum_{j=1}^{\frac{n}{g}} k_j = \kappa_n}} \binom{n/g}{k_1,\ldots,k_{n/g}} \prod_{s=1}^{n/g}\tilde{a}_{k_s},$$

with $\tilde{a}_\ell$ denoting the shifting coefficients of $G(z) = [v^g]\log M(z,v)$.

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

Proposition 12. *The number of expressions $E_{m,n}(f)$ 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 for $m = \frac{n}{2} + \kappa_n$ *with* $\kappa_n = O(n^{2/3})$ *by*

$$E_{m,n}(f) = m! \frac{2^{2m+\frac{n}{2}+1}}{\sqrt{6\pi n s_n}} s_n^{-m+\frac{n}{2}} \exp\left(\frac{3ns_n}{4} + \frac{1}{48} n s_n^2 + O(n s_n^4)\right).$$

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

$$s_n = \frac{4}{3} \cdot \frac{\kappa_n}{n} + \mathcal{O}\left(\frac{\kappa_n^2}{n^2}\right).$$

Proof. In the expression for $E_{m,n}(f)$, Eq. (7), a fixed function

$$G(z) = \sum_{j=1}^{g} \frac{(-1)^{j-1}}{j} e_{j,g-j}(z)$$

raised to a large power appears. Hence, we can apply the saddle-point technique to obtain an asymptotic expansion of $E_{m,n}(f)$ for m and n tending to infinity. In general,

$$\begin{aligned} E_{m,n}(f) &= m! 4^m (g!)^{\frac{n}{g}} [z^m] \big(G(z)\big)^{\frac{n}{g}} \\ &= \frac{m! 4^m (g!)^{\frac{n}{g}}}{2\pi i} \oint_r \frac{G^{\frac{n}{g}}(z)}{z^{m+1}} dz \\ &= \frac{m! 4^m (g!)^{\frac{n}{g}}}{2\pi i} \oint_r \exp\big(\frac{n}{g} \log G(z) - (m+1)\log z\big) dz. \end{aligned}$$

The saddle point equation is given by

$$\frac{zG'(z)}{G(z)} = \frac{m+1}{\frac{n}{g}}.$$

By our previous observation on functions f with associated integer partition representation of the form $\mathbf{i}(f) = (0, \dots, 0, n/g, 0, \dots)$ we must have $m \geq \frac{g-1}{g} \cdot n$ in order to ensure that $E_{m,n}(f) > 0$. Hence, we assume that $m = \frac{g-1}{g} \cdot n + \kappa_n - 1$, with $\kappa_n \geq 1$ and asympotically $\kappa_n = o(n)$.[4] Thus, we obtain further

$$\frac{zG'(z)}{G(z)} = g - 1 + g\frac{\kappa_n}{n}.$$

For every concrete fixed g it should be possible to treat this equation (preferentially using a computer algebra system); we outline the main steps for the simplest case $g = 2$ and the case of $\kappa_n \to \infty$, assuming that $\kappa_n = \mathcal{O}(n^{\frac{2}{3}})$. For $g = 2$ we get $G(z) = \frac{1}{2} e^z \cdot (e^z - 1)$. It is convenient to cancel the factor $\frac{1}{2}$, appearing in $G(z)$ (and which is then raised to the power $\frac{n}{2}$) with $(2!)^{\frac{n}{2}}$. We define $\tilde{G}(z) = e^z \cdot (e^z - 1)$ such that the saddle point equation for $\tilde{G}(z)$ is identical to the previous equation for $G(z)$. We obtain

$$\begin{aligned} E_{m,n}(f) &= \frac{m! 4^m}{2\pi i} \oint_r \exp\big(\frac{n}{2} \log \tilde{G}(z) - (m+1)\log z\big) dz \\ &= \frac{m! 4^m}{2\pi i} \oint_r \exp\big(\frac{n}{2} z + \frac{n}{2} \log(e^z - 1) - (m+1)\log z\big) dz, \end{aligned}$$

and the saddle point equation simplifies to

$$\frac{z(2e^z - 1)}{e^z - 1} = 1 + \frac{2\kappa_n}{n}.$$

Note that for $n \to \infty$ we have $\frac{\kappa_n}{n} \to 0$; $\frac{zG'(z)}{G(z)}$ can be expressed in terms of the Bernoulli numbers, such that

$$\frac{z(2e^z - 1)}{e^z - 1} = \sum_{k \geq 0} B_k\big(2 \cdot (-1)^k - 1\big) \frac{z^k}{k!} = 1 + \frac{3}{2} z + \frac{1}{12} z^2 - \frac{1}{720} z^4 + \mathcal{O}(z^6),$$

[4] It is also possible to extend the analysis to larger m, i.e. $m \sim \alpha \cdot n$ with $\alpha > \frac{g-1}{g}$, or $m \gg n$.

in a neighbourhood of zero. Thus, we obtain the solution s_n of the saddle point equation, with $\lim_{n\to\infty} s_n = 0$, by a bootstrapping procedure. First, we obtain

$$s_n = \frac{4}{3}\cdot\frac{\kappa_n}{n} + \mathcal{O}\left(\frac{\kappa_n^2}{n^2}\right).$$

A second bootstrapping step gives the refinement

$$s_n = \frac{4}{3}\cdot\frac{\kappa_n}{n} - \frac{8}{81}\cdot\frac{\kappa_n^2}{n^2} + \mathcal{O}\left(\frac{\kappa_n^3}{n^3}\right).$$

Changing the integration path to $z = s_n\cdot e^{i\varphi}$, $-\pi \le \varphi < \pi$ gives for $g = 2$

$$\begin{aligned} E_{m,n}(f) &= \frac{m!4^m}{2\pi i}\oint_r \exp\big(\frac{n}{2}z + \frac{n}{2}\log(e^z-1) - (m+1)\log z\big)dz \\ &= \frac{m!4^m}{2\pi}\int_{-\pi}^{\pi} s_n \exp\big(i\varphi + \frac{n}{2}\log\tilde{G}(s_n\cdot e^{i\varphi}) - (m+1)(\log s_n + i\varphi)\big)d\varphi \end{aligned}$$

Since $\tilde{G}(z) = e^z\cdot(e^z-1)$ we obtain further

$$E_{m,n}(f) = \frac{m!4^m s_n^{-m}}{2\pi}\int_{-\pi}^{\pi}\exp\big(\frac{n}{2}s_n e^{i\varphi} + \frac{n}{2}\log(e^{s_n\cdot e^{i\varphi}}-1) - mi\varphi\big)d\varphi.$$

The function $|\tilde{G}(s_n\cdot e^{i\varphi})|$ is maximal at $\varphi = 0$. Thus, we restrict ourselves to a neighbourhood of zero $\varphi\in(-\theta,\theta)$. The expansion of the term $\frac{n}{2}\log\tilde{G}(s_n\cdot e^{i\varphi}) - im\varphi$ at $\varphi = 0$ gives

$$\begin{aligned} &\frac{1}{2}ns_n + \frac{1}{2}n\log(e^{s_n}-1) + \varphi i\left(\frac{1}{2}ns_n + \frac{1}{2}n\frac{s_n e^{s_n}}{e^{s_n}-1} - m\right) \\ &\quad + \varphi^2\frac{ns_n}{4}\left(\frac{e^{2s_n}s_n}{(e^{s_n}-1)^2} - \frac{e^{s_n}(s_n+1)}{e^{s_n}-1} - 1\right) + \mathcal{O}(ns_n\varphi^3). \end{aligned}$$

By definition of s_n as the solution of the saddle point equation the linear term vanishes. We obtain

$$\begin{aligned} E_{m,n}(f) &\sim \frac{4^m(2!)^{\frac{n}{2}}s_n^{-m}e^{\frac{ns_n}{2}}(e^{s_n}-1)^{\frac{n}{2}}}{2\pi} \\ &\quad\times\int_{-\theta}^{\theta}\exp\left(\varphi^2\frac{ns_n}{4}\left(\frac{e^{2s_n}s_n}{(e^{s_n}-1)^2} - \frac{e^{s_n}(s_n+1)}{e^{s_n}-1} - 1\right) + \mathcal{O}(ns_n\varphi^3)\right)d\varphi. \end{aligned}$$

The expansion of $(e^{s_n}-1)^{\frac{n}{2}}$ gives

$$(e^{s_n}-1)^{\frac{n}{2}} = \exp(\frac{n}{2}\log(e^{s_n}-1)) = \exp\Big(\frac{1}{2}n\log s_n + \frac{1}{4}ns_n + \frac{1}{48}ns_n^2 + \mathcal{O}(ns_n^4)\Big).$$

Moreover, using

$$\frac{ns_n}{4}\Big(\frac{e^{2s_n}s_n}{(e^{s_n}-1)^2} - \frac{e^{s_n}(s_n+1)}{e^{s_n}-1} - 1\Big) = -\frac{3}{8}s_n n - \frac{1}{24}s_n^2 n + \frac{1}{720}ns_n^4 + \mathcal{O}(ns_n^6),$$

we get for the integral the asymptotic expansion

$$\int_{-\theta}^{\theta}\exp\Big(\varphi^2\big(-\frac{3}{8}s_n n - \frac{1}{24}s_n^2 n\big) + \mathcal{O}(ns_n\varphi^2(s_n^2+\varphi))\Big)d\varphi.$$

Note that the level of precision of the expansions has to be adapted on the actual growth of κ_n, here $\kappa_n = \mathcal{O}(n^{\frac{2}{3}})$. In the final step we substitute $\varphi = \vartheta/\sqrt{ns_n}$ and complete the tails:

$$\begin{aligned} E_{m,n}(f) &\sim \frac{m!4^m\cdot 2^{\frac{n}{2}}s_n^{-m+\frac{n}{2}}e^{\frac{3ns_n}{4}+\frac{1}{48}ns_n^2}}{2\pi ns_n}\int_{-\theta\sqrt{ns_n}}^{\theta\sqrt{ns_n}}\exp\Big(\vartheta^2\big(-\frac{3}{8} + \mathcal{O}(s_n + \frac{\vartheta}{\sqrt{ns_n}})\big)\Big)d\vartheta \\ &\sim \frac{m!2^{2m+\frac{n}{2}}s_n^{-m+\frac{n}{2}}e^{\frac{3ns_n}{4}+\frac{1}{48}ns_n^2}}{2\pi ns_n}\int_{-\infty}^{\infty}\exp\Big(\vartheta^2\big(-\frac{3}{8}\big)\Big)d\vartheta. \end{aligned}$$

Finally, we use $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}e^{-x^2/2}dx = 1$ to obtain the assertion. ■

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 6). 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, with many blocks of bounded size) were studied and, under some regularity conditions on the block sizes, we were able to get the asymptotic probability.

Apart from extensions of our results to cover, e.g., the extension of Theorem 2 to the supercritical case, or of Proposition 12 to a larger number of edges, 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] G. E. Andrews. *The theory of partitions*. Addison-Wesley Publishing Co., Reading, Mass.-London-Amsterdam, 1976. Encyclopedia of Mathematics and its Applications, Vol. 2.

[2] C. Banderier, P. Flajolet, G. Schaeffer, and M. Soria. Random maps, coalescing saddles, singularity analysis, and Airy phenomena. *Random Struct. Algorithms*, 19(3-4):194–246, 2001.

[3] E. A. Bender, E. R. Canfield, and B. D. McKay. The asymptotic number of labeled connected graphs with a given number of vertices and edges. *Random Struct. Algorithms*, 1(2):127–170, 1990.

[4] N. L. Biggs, E. K. Lloyd, and R. J. Wilson. Graph theory, 1736-1936. *Oxford University Press*, 1974.

[5] A. Brodsky and N. Pippenger. The Boolean functions computed by random Boolean formulas or how to grow the right function. *Random Structures and Algorithms*, 27:490–519, 2005.

[6] B. Chauvin, P. Flajolet, D. Gardy, and B. Gittenberger. And/Or trees revisited. *Combinatorics, Probability and Computing*, 13(4-5):475–497, July-September 2004.

[7] B. Chauvin, D. Gardy, and C. Mailler. The growing tree distribution for Boolean functions. In *8th SIAM Workshop on Analytic and Combinatorics (ANALCO)*, pages 45–56, 2011.

[8] L. Comtet. *Advanced combinatorics*. D. Reidel Publishing Co., Dordrecht, enlarged edition, 1974.

[9] N. Creignou and H. Daudé. Satisfiability threshold for random XOR-CNF formulas. *Discrete Applied Mathematics*, 96-97:41–53, 1999.

[10] N. Creignou and H. Daudé. Smooth and sharp thresholds for random k-XOR-CNF satisfiability. *Theor. Inform. Appl.*, 37(2):127–147, 2003.

[11] N. Creignou and H. Daudé. Coarse and sharp transitions for random generalized satisfiability problems. In *Mathematics and Computer Science III*, Trends Math., pages 507–516. Birkhäuser, Basel, 2004.

[12] N. Creignou, H. Daudé, and O. Dubois. Approximating the satisfiability threshold for random k-XOR-formulas. *Combin. Probab. Comput.*, 12(2):113–126, 2003.

[13] N. Creignou, H. Daudé, and U. Egly. Phase transition for random quantified XOR-formulas. *J. Artif. Intell. Res.*, 29:1–18, 2007.

[14] H. Daudé and V. Ravelomanana. Random 2XorSat phase transition. *Algorithmica*, 59(1):48–65, 2011.

[15] E. de Panafieu. *Analytic Combinatorics of Graphs, Hypergraphs and Inhomogeneous Graphs*. PhD thesis, Université Paris-Diderot, Sorbonne Paris-Cité, 2014.

[16] E. de Panafieu, D. Gardy, B. Gittenberger, and M. Kuba. Probabilities of 2-xor functions. In A. Pardo and A. Viola, editors, *International Conference LATIN*, volume 8392. Springer Lecture Notes in Comput. Sci., 2014.

[17] E. de Panafieu and V. Ravelomanana. Analytic description of the phase transition of inhomogeneous multigraphs. *European Journal of Combinatorics*, 2014. Accepted. Also available at http://arxiv.org/pdf/1409.8424.pdf.

[18] M. Drmota. *Random trees*. Springer, Vienna-New York, 2009.

[19] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[20] P. Flajolet, D. E. Knuth, and B. Pittel. The first cycles in an evolving graph. *Discrete Mathematics*, 75(1-3):167–215, 1989.

[21] P. Flajolet, B. Salvy, and G. Schaeffer. Airy phenomena and analytic combinatorics of connected graphs. *Electr. J. Comb.*, 11(1), 2004.

[22] P. Flajolet and R. Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009.

[23] H. Fournier, D. Gardy, A. Genitrini, and B. Gittenberger. The fraction of large random trees representing a given Boolean function in implicational logic. *Random Structures and Algorithms*, 40(3):317–349, 2012.

[24] H. Fournier, D. Gardy, A. Genitrini, and M. Zaionc. Classical and intuitionistic logic are asymptotically identical. In *16th Annual Conference on Computer Science Logic (EACSL), volume 4646 of Lecture Notes in Computer Science*, pages 177–193, 2007.

[25] A. Genitrini, B. Gittenberger, V. Kraus, and C. Mailler. Associative and commutative tree representations for Boolean functions. 2011. Submitted. Also available at http://arxiv.org/abs/1305.0651.

[26] A. Genitrini, B. Gittenberger, V. Kraus, and C. Mailler. Probabilities of Boolean functions given by random implicational formulas. *Electronic Journal of Combinatorics*, 19(2):P37, 20 pages, (electronic), 2012.

[27] A. Genitrini, J. Kozik, and M. Zaionc. Intuitionistic vs. classical tautologies, quantitative comparison. In *Types for proofs and programs*, volume 4941, pages 100–109. Springer Lecture Notes in Comput. Sci., 2008.

[28] S. Janson, D. Knuth, T. Łuczak, and B. Pittel. The birth of the giant component. *Random Structures and Algorithms*, 4(3):233–358, 1993.

[29] J. Kozik. Subcritical pattern languages for And/Or trees. In *Fifth Colloquium on Mathematics and Computer Science*, Blaubeuren, Germany, september 2008. DMTCS Proceedings.

[30] H. Lefmann and P. Savický. Some typical properties of large And/Or Boolean formulas. *Random Structures and Algorithms*, 10:337–351, 1997.

[31] B. Pittel and N. C. Wormald. Counting connected graphs inside-out. *J. Comb. Theory, Ser. B*, 93(2):127–172, 2005.

[32] B. Pittel and J.-A. Yeum. How frequently is a system of 2-linear equations solvable? *Electronic Journal of Combinatorics*, 17, 2010.

[33] R. van der Hofstad and J. Spencer. Counting connected graphs asymptotically. *European Journal on Combinatorics*, 26(8):1294–1320, 2006.

[34] A. Woods. On the probability of absolute truth for And/Or formulas. *Bulletin of Symbolic Logic*, 12(3), 2005.

[35] E. M. Wright. The number of connected sparsely edged graphs. *Journal of Graph Theory*, 1:317–330, 1977.

[36] E. M. Wright. The number of connected sparsely edged graphs III: Asymptotic results. *Journal of Graph Theory*, 4(4):393–407, 1980.

[37] M. Zaionc. On the asymptotic density of tautologies in logic of implication and negation. *Reports on Mathematical Logic*, 39:67–87, 2005.