Every decision tree has an influential variable

Ryan O'Donnell*
Microsoft Research
odonnell@microsoft.com

Michael Saks†
Rutgers University
saks@math.rutgers.edu

Oded Schramm
Microsoft Research

Rocco A. Servedio‡
Columbia University
rocco@cs.columbia.edu

January 26, 2006

Abstract

We prove that for any decision tree calculating a boolean function $f : \{-1,1\}^n \to \{-1,1\}$,

$$\mathbf{Var}[f] \leq \sum_{i=1}^{n} \delta_i \, \mathbf{Inf}_i(f),$$

where δ_i is the probability that the ith input variable is read and $\mathbf{Inf}_i(f)$ is the influence of the ith variable on f. The variance, influence and probability are taken with respect to an arbitrary product measure on $\{-1,1\}^n$. It follows that the minimum depth of a decision tree calculating a given balanced function is at least the reciprocal of the largest influence of any input variable. Likewise, any balanced boolean function with a decision tree of depth d has a variable with influence at least $\frac{1}{d}$. The only previous nontrivial lower bound known was $\Omega(d2^{-d})$. Our inequality has many generalizations, allowing us to prove influence lower bounds for randomized decision trees, decision trees on arbitrary product probability spaces, and decision trees with non-boolean outputs. As an application of our results we give a very easy proof that the randomized query complexity of nontrivial monotone graph properties is at least $\Omega(v^{4/3}/p^{1/3})$, where v is the number of vertices and $p \leq \frac{1}{2}$ is the critical threshold probability. This supersedes the milestone $\Omega(v^{4/3})$ bound of Hajnal [13] and is sometimes superior to the best known lower bounds of Chakrabarti-Khot [9] and Friedgut-Kahn-Wigderson [11].

*Some of this research was performed while this author was at the Institute for Advanced Study

†Supported in part by NSF grants CCR-9988526 and CCR-0515201.

‡Supported in part by NSF CAREER award CCF-0347282 and a Sloan Foundation Fellowship.

This paper is posted by permission from the IEEE Computer Society. To appear in FOCS 2005.

1 Introduction

1.1 Motivation.

This paper lies at the intersection of two topics within the theory of boolean functions.

The first topic is *decision tree complexity*. A *deterministic decision tree* (DDT) for a boolean function $f : \{-1,1\}^n \to \{-1,1\}$ is a deterministic adaptive strategy for reading variables so as to determine the value of f (a formal definition appears in Section 3.1). The cost of a DDT on a given input is simply the number of input variables that it reads, and the DDT complexity of a function f, $D(f)$, is the minimum over all DDT's for f of the maximum cost of any input. A *randomized decision tree* (RDT) for f is a probability distribution over DDTs for f; such trees are sometimes known as *zero-error* randomized decision trees. The RDT complexity of f, $R(f)$, is the minimum over all RDT's for f of the maximum expected cost of any input. Decision tree complexity has been studied in theoretical computer science for over 30 years and there is now a significant body of research on the subject (for a survey, see e.g., [8]).

The second topic is *variable influences*, introduced to theoretical computer science by Ben-Or and Linial in 1985 [2]. Any n-variate boolean function f has an associated *influence vector* $(\mathbf{Inf}_1(f), \ldots, \mathbf{Inf}_n(f))$ where $\mathbf{Inf}_i(f)$ measures the extent to which the value of f depends on variable i (a precise definition appears in Section 1.2). A number of papers have dealt with properties of this vector and its relation to other properties of boolean functions; perhaps the best known work along these lines is that of Kahn, Kalai and Linial [14] ("KKL") concerning the maximum influence $\mathbf{Inf}_{\max}(f) = \max\{\mathbf{Inf}_i(f) : i \in [n]\}$. Their result implies, for example, that $\mathbf{Inf}_{\max}(f) = \Omega(\frac{\log n}{n})$ for any near-balanced boolean function f (where we say that f is near-balanced if both $|f^{-1}(1)|/2^n$ and $|f^{-1}(-1)|/2^n$ are $\Omega(1)$).

The question that originally motivated this paper was: what is the best lower bound on $\mathbf{Inf}_{\max}(f)$ that holds for all near-balanced boolean functions f satisfying $D(f) \leq d$? It is easy to see that such a function f depends on at most 2^d of its variables and therefore the KKL result implies $\mathbf{Inf}_{\max}(f) \geq \Omega(\frac{d}{2^d})$; prior to this work, this was the best lower bound known. Our main inequality for boolean functions, Theorem 1.1, implies a (tight) lower bound of $\mathbf{Inf}_{\max}(f) \geq \Omega(\frac{1}{d})$ for any near-balanced function f satisfying $D(f) \leq d$.

In fact, Theorem 1.1 provides a lower bound on a weighted average of the influence vector, where $\mathbf{Inf}_i(f)$ is weighted by the probability that a DDT for f queries x_i when x is a randomly chosen *input*. This lets us extend our lower bound on $\mathbf{Inf}_{\max}(f)$ to functions with $R(f) \leq d$ and even to functions with $\Delta(f) \leq d$, where $\Delta(f)$ denotes the expected number of queries made by the best DDT for f on a random input (again, see Section 1.2 for precise definitions).

1.2 The main theorem for boolean functions.

Our main theorem holds in a very general setting, that of functions from product probability spaces into metric spaces. However the case of greatest interest to us is much simpler. Fix some $p \in (0,1)$ and let $\{-1,1\}^n_{(p)}$ denote the discrete cube endowed with the p-biased product measure, $\mu_{(p)}(x) = p^{|\{i : x_i = 1\}|}(1-p)^{|\{i : x_i = -1\}|}$. When we write simply $\{-1,1\}^n$ the uniform measure case $p = \frac{1}{2}$ is implied. Our main interest is in boolean functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$, and in this section we will describe our main theorem in this case.

First we recall a few definitions. We have

$$\mathbf{Var}[f] = \mathbf{E}[f^2] - \mathbf{E}[f]^2 = 4\,\mathbf{Pr}[f = 1]\,\mathbf{Pr}[f = -1].$$

This measures the "balance" of f; if f is equally likely to be 1 as -1, then $\mathbf{Var}[f] = 1$. We also make the following definition for the *influence of the ith coordinate* on f:

$$\mathbf{Inf}_i(f) = 2 \Pr_{x, x^{(i)}}[f(x) \neq f(x^{(i)})],$$

where x is drawn from $\{-1,1\}^n_{(p)}$ and $x^{(i)}$ is formed by *rerandomizing* the ith coordinate of x. Note that our definition agrees with the one introduced in [2] in the uniform measure case $p = \frac{1}{2}$, which was $\mathbf{Inf}_i[f] = \mathbf{Pr}[f(x) \neq f(x \oplus i)]$. (Our definition differs from the p-biased notion of influences used in, e.g., [12] by a factor of $4\,p\,(1-p)$; we prefer rerandomizing the ith coordinate to flipping it, since this makes sense in more general product probability spaces which we will consider later.) We call $\mathbf{Inf}(f) := \sum_{i=1}^n \mathbf{Inf}_i(f)$ the *total influence* of f.

Finally, since the notion of influences involves randomizing over the input domain, it makes sense to introduce a notion of randomizing over inputs for decision trees. Let T be a DDT computing a function $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$. We write

$$\delta_i(T) = \Pr_{x \in \{-1,1\}^n_{(p)}}[T \text{ queries } x_i], \qquad \text{and}$$

$$\Delta(T) = \sum_{i=1}^n \delta_i(T) = \mathop{\mathbf{E}}_{x \in \{-1,1\}^n_{(p)}}[\#\text{ coords } T \text{ queries on } x\,].$$

We also let $\Delta(f)$ denote the minimum of $\Delta(T)$ over all DDTs T computing $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$. It is easy to see that this is equivalent to minimizing over all RDTs computing f; hence $\Delta(f) \leq R(f)$ for all p. Also note that $\Delta(f)$ can be upper-bounded in terms of the *size* (number of leaves) of the smallest DDT T for f: [19] shows $\Delta(f) \leq \log_2(\text{size(T)})/H(p)$, where $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$ is the binary entropy of p.

We may now state our main theorem in the case of functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$:

Theorem 1.1 *Let $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ and let T be a DDT computing f. Then*

$$\mathbf{Var}[f] \le \sum_{i=1}^{n} \delta_i(T)\ \mathbf{Inf}_i(f).$$

As an immediate corollary we obtain the lower bound on $\mathbf{Inf}_{\max}(f)$ mentioned in Section 1.1:

Corollary 1.2 *For every $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ we have*

$$\Delta(f) \ge \frac{\mathbf{Var}(f)}{\mathbf{Inf}_{\max}(f)}\,.$$

Proof: Let T be a DDT computing f. From Theorem 1.1,

$$\mathbf{Var}[f] \le \sum_{i=1}^{n} \delta_i(T)\ \mathbf{Inf}_i(f)$$
$$\le \mathbf{Inf}_{\max}(f) \sum_{i=1}^{n} \delta_i(T) = \mathbf{Inf}_{\max}(f) \cdot \Delta(T)\,. \quad \square$$

Some brief comments on our main theorem:

- It is linear in the $\delta_i(T)$'s. Hence if we allow an RDT $\mathcal{T}$ for f and make the natural definition of $\delta_i(\mathcal{T})$, the result still holds by averaging over the distribution $\mathcal{T}$.
- It can be sharp; see Section 3.5 for cases of equality.
- Other corollaries along the lines of Corollary 1.2 follow; for example, if d is an integer $\ge \Delta(f)$, then the sum of the influences of the d most influential variables is at least $\mathbf{Var}[f]$.
- In Section 3.3 we will give a "two function" version, which yields a lower bound for the randomized decision tree complexity of approximating f.

1.2.1 Influence lower bounds — comparison with previous work.

Proving lower bounds on the influences of boolean functions has had a long history in theoretical computer science, starting with the 1985 paper of Ben-Or and Linial [2] on collective coin flipping. Ben-Or and Linial made the basic observation that if $f : \{-1,1\}^n \to \{-1,1\}$ is balanced (i.e., $\mathbf{E}[f] = 0$), then $\mathbf{Inf}_{\max}(f) \ge \frac{1}{n}$. This follows from the edge isoperimetric inequality on the discrete cube (see, e.g., [6]); however, it is more instructive for us to view it as following from the *Efron-Stein inequality* [10, 26],

$$\mathbf{Var}[f] \le \mathbf{Inf}(f) = \sum_{i=1}^{n} \mathbf{Inf}_i(f), \qquad (1)$$

which holds in the general p-biased case, and also in the much more general setting of $f : \Omega \to \mathbb{R}$, where Ω is a n-wise product probability space and $\mathbf{Inf}_i$ is defined appropriately for real-valued functions (specifically, with the "ρ_2 semimetric" discussed in Section 3.4). Theorem 1.1 is immediately seen to improve the Efron-Stein inequality in the case of functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$.

Ben-Or and Linial constructed a balanced function $f : \{-1,1\}^n \to \{-1,1\}$ ("Tribes") satisfying $\mathbf{Inf}_{\max}(f) = \Theta(\frac{\log n}{n})$ and conjectured that for every balanced function $f : \{-1,1\}^n \to \{-1,1\}$, $\mathbf{Inf}_{\max}$ cannot be smaller. There were small improvements on the simple $\frac{1}{n}$ bound ($\frac{2-\epsilon}{n}$ by Alon, $\frac{3-\epsilon}{n}$ by Chor and Gereb-Graus; see [14]) before the famous KKL paper [14] confirmed the conjecture. Note that our theorem improves upon KKL whenever f has $\Delta(f) = o(n/\log n)$; in particular, whenever f has a DDT of size $2^{o(n/\log n)}$.

The KKL result was subsequently generalized by Talagrand [27, Theorem 1.5] who proved that for any $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$,

$$\mathbf{Var}[f] \le O\Big(\log \frac{1}{p(1-p)}\Big) \sum_{i=1}^{n} \frac{\mathbf{Inf}_i(f)}{\log(1/\mathbf{Inf}_i(f))}\,. \qquad (2)$$

Talagrand's motivation for proving this was that when $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ is monotone, lower bounds on the sum of f's influences imply a "sharp threshold" for f, via the Russo-Margulis lemma [16, 21]. Indeed, this connection with threshold phenomena is one of the chief motivations for studying influences, and it is considered an important problem in the theory of boolean functions and random graphs to provide general conditions under which the total influence is large [7]. Our main inequality provides such a condition: $\mathbf{Inf}(f)$ is large if f has a randomized decision tree $\mathcal{T}$ with $\delta_i(\mathcal{T})$ small for all i. Note that when f is a transitive function, this is equivalent to the natural condition that $\Delta(f)$ is small. (See Section 2 for definitions of monotone and transitive functions, as well as further discussion of random graph properties.)

In particular, ours seems to be the first quantitatively strong influence lower bound that takes into account the "structure" or computational complexity of f. We note that previously achievable lower bounds on influences in terms of some measure of the complexity of f yield quantitatively much weaker results than can be obtained from our inequality. For instance, Nisan and Szegedy [18] showed that if $f : \{-1,1\}^n \to \{-1,1\}$ is computed by a polynomial over $\mathbb{R}$ of degree $\deg(f)$, then every coordinate i with nonzero influence has $\mathbf{Inf}_i(f) \geq 2^{-\deg(f)}$. Since $D(f) \leq O(\deg(f)^4)$ (by a result of Nisan and Smolensky [8]), our Corollary 1.2 implies that the maximum influence in fact satisfies $\mathbf{Inf}_{\max}(f) \geq \Omega(\mathbf{Var}[f]/\deg(f)^4)$. As another example, suppose $f : \{-1,1\}^n \to \{-1,1\}$ is approximately computed by a polynomial over $\mathbb{R}$ of degree $\widetilde{\deg}(f)$ — i.e. there is a polynomial $p(x)$ of degree $\widetilde{\deg}(f)$ such that $|p(x) - f(x)| < 1/3$ for all x. Talagrand's result implies that $\mathbf{Inf}_{\max}(f) \geq \exp(-O(\mathbf{Inf}(f)/\mathbf{Var}[f]))$. Since by [24] we have $\mathbf{Inf}(f) \leq O(\widetilde{\deg}(f))$, one could conclude that $\mathbf{Inf}_{\max}(f) \geq \exp(-O(\widetilde{\deg}(f)/\mathbf{Var}[f]))$. However by contrast, since $D(f) \leq O(\widetilde{\deg}(f)^6)$ by [1], our Corollary 1.2 implies that the maximum influence in fact satisfies $\mathbf{Inf}_{\max}(f) \geq \Omega(\mathbf{Var}[f]/\widetilde{\deg}(f)^6)$.

2 Randomized decision tree complexity lower bounds

In this section we give an application of Theorem 1.1 to the problem of randomized decision tree complexity for monotone graph properties. We prove Theorem 1.1 in a more general setting in Section 3.

2.1 History.

As mentioned in Section 1.1, decision tree complexity has been extensively studied for over three decades. Two special classes of functions have played a prominent role in these investigations. The first is the class of *monotone functions*, those satisfying $f(y) \geq f(x)$ whenever $y \geq x$ under the componentwise partial order. The second is the class of *transitive functions*. An automorphism of the n-variate boolean function f is a permutation σ of $[n]$ satisfying $f(x_1, \dots, x_n) = f(x_{\sigma(1)}, \dots, x_{\sigma(n)})$ for all inputs x. We say that f is *transitive* if for each pair $i, j \in [n]$ there is an automorphism of f that sends i to j. For example, Rivest and Vuillemin [20] proved that for n a prime power, any n-variate monotone transitive function f has $D(f) = n$.

One long studied open question about boolean decision tree complexity is the following: how small can $R(f)$ be in relation to $D(f)$? It is well known [5] that $R(f) \geq \Omega(\sqrt{D(f)})$ for any function f, and this is the best general lower bound known. The largest known separation is given by the following recursively defined function: Let f_0 be the identity function on a single variable and for $k \geq 1$, let f_k be the function on $n = 4^k$ variables given by $(f^1_{k-1} \wedge f^2_{k-1}) \vee (f^3_{k-1} \wedge f^4_{k-1})$, where f^i_{k-1} is the value of f_{k-1} on the ith group of 4^{k-1} variables. The function f_k is monotone and transitive, and so by the above result of Rivest and Vuillemin, $D(f_k) = n$. Snir [25] gave an RDT for f_k establishing $R(f) \leq n^\beta$ where $\beta = \log_2\left(\frac{1+\sqrt{33}}{4}\right) \approx 0.753$. Saks and Wigderson [22] proved that Snir's RDT is optimal for f_k and conjectured that $R(f) \geq \Omega(D(f)^\beta)$ for any boolean function; this is not even known to hold for all monotone transitive functions.

A well studied subclass of transitive boolean functions consists of functions derived from graph properties. A *property* of v-vertex (undirected) graphs is a set of graphs on vertex set $V = \{1, \dots, v\}$ that is invariant under vertex relabellings; e.g., the set of graphs on V that are properly 3-colorable. We restrict attention to properties that are non-trivial; i.e., at least one graph has the property and at least one graph does not have the property.

Let $\binom{V}{2}$ denote the set of 2-elements subsets of V. Each graph G on V can be identified with the boolean vector $x^G \in \{-1,1\}^{\binom{V}{2}}$ where $x^G_{\{i,j\}}$ is 1 if $\{i,j\} \in E(G)$ and is -1 otherwise. A graph property $\mathcal{P}$ is thus naturally identified with a boolean function $f_{\mathcal{P}} : \{-1,1\}^{\binom{V}{2}} \longrightarrow \{-1,1\}$ which maps the vector x^G to 1 if and only if G satisfies $\mathcal{P}$. The invariance of properties under vertex relabellings implies that the associated functions are transitive.

There are examples of graph properties on v vertices that have deterministic decision trees of depth $O(v)$; e.g., the property of being a "scorpion graph" [4]. However, for graph properties that are *monotone* (those whose associated function is monotone), Rivest and Vuillemin [20] proved a lower bound $\Omega(v^2)$ on DDT complexity. A conjecture made by Yao [28] and also attributed to Karp [22] is that this $\Omega(v^2)$ lower bound extends to RDT complexity. This is the problem we make progress on in this section.

Yao observed that an $\Omega(v)$ lower bound for RDT computation of monotone graph properties is easy to prove; this also follows from the general bound $R(f) = \Omega(\sqrt{D(f)})$ mentioned earlier. The first improvement on this naive bound came a decade later from Yao himself, who proved an $\Omega(v \log^{1/12} v)$ lower bound using "graph packing" arguments [29]. These arguments were improved by King [15], yielding an $\Omega(v^{5/4})$ lower bound, and by Hajnal [13], yielding an $\Omega(v^{4/3})$ lower bound. This lower bound stood for a decade before Chakrabarti and Khot [9] gave a small improvement to $\Omega(v^{4/3} \log^{1/3} v)$. Both the Hajnal and Chakrabarti-Khot bounds have rather long and technical proofs based on graph packing.

Fairly recently, Friedgut, Kahn and Wigderson [11] proved a general lower bound of a somewhat different form. Given a nonconstant monotone boolean function $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$, it is easy to see that $\mathbf{E}[f]$ is a continuous increasing function of p; therefore there is a *critical probability* p for which $\mathbf{E}[f] = 0$, i.e., $\mathbf{Var}[f] = 1$. Friedgut, Kahn and Wigderson proved that any nontrivial monotone v-vertex graph property has RDT complexity $\Omega(\min\{\frac{v}{\min(p,1-p)}, \frac{v^2}{\log v}\})$ when p is the critical probability for f. In fact, they show that $\Delta(f)$ is at least this quantity. The FKW bound can improve on Chakrabarti-Khot in cases where the critical probability is sufficiently close to 0 or 1. We remark that the proof in FKW also uses a graph packing argument.

2.2 Our $R(f)$ lower bound.

As a simple consequence of our elementary main inequality Theorem 1.1 and a recent elementary inequality from [19], we obtain the following:

Theorem 2.1 *Let $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ be a nonconstant monotone transitive function, where p is the critical probability for f (i.e., f is balanced). Write $q = 1-p$. Then*

$$R(f) \geq \Delta(f) \geq \frac{n^{2/3}}{(4\,p\,q)^{1/3}}.$$

In particular,

$$R(f) \geq \Delta(f) \geq \frac{(v-1)^{4/3}}{(16\,p\,q)^{1/3}}$$

if f corresponds to a v-vertex graph property.

Proof: The inequality we need from [19] is the following:

For all p, if $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ is monotone then

$$\mathbf{Inf}(f) \leq 2\sqrt{p\,q\,\Delta(f)}\,. \qquad (3)$$

Fix p to be the critical probability of f and let T be a DDT computing f with expected cost $\Delta(f)$. We apply Theorem 1.1, using $\mathbf{Var}[f] = 1$ since p is critical and $\mathbf{Inf}_i(f) = \mathbf{Inf}(f)/n$ since f is transitive (and hence all coordinates have the same influence). This gives $1 \leq (\mathbf{Inf}(f)/n) \cdot \Delta(f)$. Using (3) to bound $\mathbf{Inf}(f)$ we get $1 \leq (2\sqrt{p\,q}/n) \cdot (\Delta(f))^{3/2}$, and this can be rearranged to give the desired result. $\Box$

2.3 Discussion.

In the case of monotone graph properties, our result always improves on Hajnal's $\Omega(v^{4/3})$ lower bound and can be superior to both Chakrabarti-Khot (when $\min\{p,q\}$ is small enough) and to FKW (when $\min\{p,q\}$ is large enough). It is worth noting that unlike all previous lower bounds for monotone graph properties, our proof makes no use of graph packing arguments, instead relying only on elementary probabilistic arguments.

Most interestingly, we obtain a result essentially as good as the best unconditional bound (Chakrabarti-Khot) in the more general context of monotone *transitive* functions, not just graph properties. Further, our bound for monotone transitive functions is known to be essentially tight in the case $p = 1/2$: in [3], a sequence $f_n : \{-1,1\}^n \to \{-1,1\}$ of balanced monotone transitive functions is presented with $\Delta(f_n) \leq O(n^{2/3} \log n)$. Our present Theorem 1.1 is used in [3] to show that $\Delta(f_n) = \Omega(n^{2/3})$, by an argument similar to the proof of Theorem 2.1, but using an inequality from [23] in place of (3).

It is tantalizing that the place where the RDT complexity of monotone graph properties has been stuck for almost 15 years, $v^{4/3}$, is exactly the tight bound for monotone transitive functions. Perhaps this suggests that in some way the argument of Hajnal is not really using the fact that f is a graph property — just that it's transitive. Indeed, one might wonder the same thing about Chakrabarti-Khot, since their $v^{4/3}\log^{1/3} v$ lower bound could also hold for monotone transitive functions — the example of [3] does not rule it out.

3 The main inequality

3.1 Decision trees, variation, influences — general definitions.

The proof of Theorem 1.1 is most naturally carried out in a significantly more general context than that of functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$. Specifically, we will consider functions

$$f : \Omega \longrightarrow Z$$

mapping a *product probability space* into a *metric space*. In this section we give the necessary definitions.

Let us begin with the domain. Here we have an n-wise product probability space $\Omega = (X, \mu)$, meaning that that the underlying set X is a product set $X_1 \times \cdots \times X_n$ and the measure μ is a product probability measure $\mu_1 \times \cdots \times \mu_n$, where μ_i is a probability measure on X_i. For simplicity, we assume that X is finite. We write Ω_i for the probability space (X_i, μ_i). We use the notation $x \leftarrow \Omega$ to mean that x is an element of X randomly selected according to μ.

The range of our functions is a metric space (Z, d). (Actually we can allow a "pseudo-metric", meaning we may omit the requirement that $d(z,z') = 0 \Rightarrow z = z'$.) Useful examples to keep in mind are the following: Z any finite set with $d(z,z') = \mathbf{1}_{z \neq z'}$; and, $Z = \mathbb{R}$ with $d(z,z') = |z - z'|$. Of course, in the special case of boolean-valued functions, $Z = \{-1,1\}$, all metrics are the same up to a constant factor.

The definitions of decision trees in the context of functions mapping a product set domain $X = X_1 \times \cdots \times X_n$ into a set Z are the obvious ones. Briefly, a DDT will be a rooted directed tree T in which each internal node v is labelled by a coordinate $i_v \in [n]$ and each leaf is labelled by an element of the output set Z. Further, the arcs emanating from each internal node v must be in one-to-one correspondence with X_{i_v}. The node labels along every root-leaf path are required to be distinct. T computes a function $f_T : X \to Z$ in the obvious way; we retain the notion of the cost of T on input x as the length of the root-leaf path T follows on input x. Thus, we have the usual notions of $D(T)$ and $D(f)$, and also the (zero-error) randomized decision tree complexities $R(\mathcal{T})$ and $R(f)$. With the product probability measure μ on X, we can also naturally extend our notions of *expected cost* from Section 1.2: given a DDT T computing f,

$$\delta_i^\mu(T) = \Pr_{x \leftarrow \Omega}[T \text{ queries } x_i],$$

and $\Delta^\mu(T)$ and $\Delta^\mu(f)$ are similarly defined. We will henceforth drop the superscript μ when it is clear from context. Note that, as before, we have $\Delta(f) \leq R(f)$.

We now give the definitions of *variation* and *influences* for functions $f : \Omega \to Z$. The *variation* of $f : \Omega \to Z$ is

$$\mathbf{Vr}^{\mu,d}[f] = \mathop{\mathbf{E}}_{(x,y) \leftarrow \Omega \times \Omega}[d(f(x), f(y))].$$

To define influences, first let $\Omega^{(i)}$ denote the probability space given by pairs $(x, x^{(i)})$, where x is chosen from Ω and $x^{(i)}$ is formed by rerandomizing the ith coordinate of x using μ_i. Then the *influence of the ith coordinate* on $f : \Omega \to Z$ is defined to be

$$\mathbf{Inf}_i^{\mu,d}(f) = \mathop{\mathbf{E}}_{(x,x^{(i)}) \leftarrow \Omega^{(i)}}[d(f(x), f(x^{(i)}))].$$

We will usually drop the superscripts μ and d on $\mathbf{Vr}$ and $\mathbf{Inf}_i$ when they are implied by context. Note that if we view the functions $f : \{-1,1\}^n_{(p)} \to \{-1,1\}$ from Section 1 as mapping into the metric space on $\{-1,1\}$ with distance d given by $d(z,z') = |z - z'| = 2 \cdot \mathbf{1}_{z \neq z'}$, then we get agreement in the definitions of $\mathbf{Inf}_i(f)$ and also $\mathbf{Vr}[f] = \mathbf{Var}[f]$.

3.2 Theorem and proof.

We now state and prove our main inequality, which includes Theorem 1.1 as a special case.

Theorem 3.1 *Let $f : \Omega \to (Z, d)$ be a function mapping a finite n-wise product probability space into a metric space, and let T be a DDT computing f. Then*

$$\mathbf{Vr}[f] \leq \sum_{i=1}^n \delta_i(T)\, \mathbf{Inf}_i(f).$$

Proof: Let x and y be random inputs chosen independently from Ω. Given a subset $J \subseteq [n]$ we will write $x_J y$ for the *hybrid input* in X that agrees with x on the coordinates in J and with y on the coordinates in $[n] \setminus J$. Let $i_1, \ldots, i_s$ denote the sequence of variables queried by T on input x (these i's are random variables and s is also a random variable). For $t \geq 0$, let $J[t] = \{i_r : s \geq r > t\}$. Finally, let $u[t] = x_{J[t]}y$. (For example, if $x = (1,-1,1,1), y = (1,1,-1,-1)$, the tree read x_4 followed by x_2 and terminates, then $u[0] = (1,-1,-1,1)$, $u[1] = (1,-1,-1,-1)$ and $u[2] = y$.) All $\mathbf{E}[\,\cdot\,]$'s and $\mathbf{Pr}[\,\cdot\,]$'s in what follows are over all the random variables just described (i.e., x, y, i's, s, $u[\cdot]$'s).

We begin with the simple observation

$$\mathbf{Vr}[f] = \mathbf{E}[d(f(x), f(y))] = \mathbf{E}[d(f(u[0]), f(u[s]))],$$

which follows because $y = u[s]$ and $f(x) = f(u[0])$ (although x does not necessarily equal $u[0]$). This latter equality is the only place in the proof we use the fact that T computes f.

We next make the obvious step

$$\mathbf{E}[d(f(u[0]), f(u[s]))] \leq \mathbf{E}\Big[\sum_{t=1}^{s} d(f(u[t-1]), f(u[t]))\Big] \tag{4}$$

which uses the fact that d is a metric. Set $i_t = \emptyset$ for $t > s$. Linearity of expectation and $1_{\{t \leq s\}} = \sum_{i=1}^{n} 1_{\{i_t = i\}}$ give

$$\mathbf{E}\Big[\sum_{t=1}^{s} d(f(u[t-1]), f(u[t]))\Big] = \sum_{t=1}^{n}\sum_{i=1}^{n} \mathbf{E}\Big[d\big(f(u[t-1]), f(u[t])\big)\, 1_{\{i_t = i\}}\Big]. \tag{5}$$

Let X_t denote the sequence of values seen by the decision tree by time t on input x; that is, $X_t = (x_{i_1}, \ldots, x_{i_{t \wedge s}})$, where $t \wedge s$ denotes the minimum of t and s. Note that X_{t-1} determines i_t. Induction on $t \in [n]$ easily shows that conditional on X_{t-1} the variables y and $(x_j : j \neq i_1, \ldots, i_{(t-1)\wedge s})$ are independent and retain their original distributions. It follows that conditional on X_{t-1} the pair $(u[t-1], u[t])$ has the distribution $\Omega^{(i)}$ if $i_t = i \in [n]$. Consequently, for $i, t \in [n]$,

$$\mathbf{E}\Big[d\big(f(u[t-1]), f(u[t])\big)\, 1_{\{i_t = i\}} \;\Big|\; X_{t-1}\Big] = 1_{\{i_t = i\}}\, \mathbf{Inf}_i(f)\,.$$

Taking expectation gives

$$\mathbf{E}\Big[d\big(f(u[t-1]), f(u[t])\big)\, 1_{\{i_t = i\}}\Big] = \mathbf{Pr}[i_t = i]\, \mathbf{Inf}_i(f)\,.$$

Since $\sum_{t=1}^{n} \mathbf{Pr}[i_t = i] = \delta_i(T)$, an appeal to (5) completes the proof. $\Box$

3.3 Corollaries and two function version.

In this section we treat some immediate corollaries of Theorem 3.1. Certainly the analogue of Corollary 1.2 holds for Theorem 3.1, as do the first and third remarks stated after Corollary 1.2. We now give the promised "two function" version. Define

$$\begin{aligned}&\mathbf{CoVr}[f,g]\\ &= \mathop{\mathbf{E}}_{(x,y)\leftarrow\Omega\times\Omega}[d(f(x), g(y))] - \mathop{\mathbf{E}}_{x\leftarrow\Omega}[d(f(x), g(x))],\end{aligned}$$

so in particular $\mathbf{CoVr}[f,f] = \mathbf{Vr}[f]$. Thus the following theorem generalizes Theorem 3.1:

Theorem 3.2 *Let $f, g : \Omega \to (Z, d)$ be functions mapping a finite n-wise product probability space into a metric space, and let $\mathcal{T}$ be an RDT computing f. Then*

$$\big|\mathbf{CoVr}[f,g]\big| \leq \sum_{i=1}^{n} \delta_i(\mathcal{T})\, \mathbf{Inf}_i(g).$$

Proof: As usual we can assume by averaging that $\mathcal{T}$ is a DDT T computing f. Using the same setup as in the proof of Theorem 3.1, we have

$$\begin{aligned}\mathbf{CoVr}[f,g] &= \mathbf{E}[d(f(x), g(y))] - \mathbf{E}[d(f(u[0]), g(u[0]))]\\ &= \mathbf{E}[d(f(u[0]), g(u[s]))] - \mathbf{E}[d(f(u[0]), g(u[0]))]\end{aligned}$$

where in the first equality we used that $u[0]$ is, in isolation, distributed according to Ω, and in the second equality we used the fact that $f(x) = f(u[0])$ since T computes f (as in the previous proof). Now using the fact that d is a metric we get

$$\begin{aligned}&\mathbf{CoVr}[f,g] =\\ &\quad \mathbf{E}[d(f(u[0]), g(u[s]))] - \mathbf{E}[d(f(u[0]), g(u[0]))]\\ &\qquad \leq \mathbf{E}[d(g(u[0]), g(u[s]))]\end{aligned}$$

and of course this is also true for $-\mathbf{CoVr}[f,g]$. The proof now proceeds exactly as before with g in place of f; note that from this point on in the previous proof we did not use the fact that T computed f. $\square$

As mentioned below Corollary 1.2, Theorem 3.2 can be used to give a lower bound for the randomized decision tree complexity of approximating g. Note that the triangle inequality gives

$$\mathbf{CoVr}[f,g] \geq \mathbf{Vr}[g] - 2\,\mathbf{E}\big[d(f(x),g(x))\big].$$

Consequently, Theorem 3.2 implies that for every $\epsilon > 0$ the expected number of queries required by a randomized decision tree to calculate any approximation f of g satisfying $\mathbf{E}\big[d(f(x),g(x))\big] \leq \epsilon$ is at least

$$\frac{\mathbf{Vr}[g] - 2\,\epsilon}{\mathbf{Inf}_{\max}(g)}.$$

We now describe an alternate version of Theorem 3.2. Let $f : \Omega \to [-1,1]$, $g : \Omega \to \mathbb{R}$, and let $\mathcal{T}$ be a randomized decision tree computing f. Then

$$\big|\mathbf{Cov}[f,g]\big| \leq \sum_{i=1}^{n} \delta_i(\mathcal{T})\,\mathbf{Inf}_i^{\rho_1}[g], \tag{6}$$

where $\rho_1(x,y) = |x-y|$ and $\mathbf{Cov}[f,g] = \mathbf{E}[f(x)\,g(x)] - \mathbf{E}[f(x)]\,\mathbf{E}[g(x)]$ is the covariance of f and g. With the usual definitions of $x, y, u[t]$ and s, we have $f(x) = f[u(0)]$ and $u[s] = y$. Hence, we may write

$$\mathbf{Cov}[f,g] = \mathbf{E}\big[f(u[0])g(u[0]) - f(x)g(u[s])\big]$$
$$= \mathbf{E}\big[f(x)\big(g(u[0]) - g(u[s])\big)\big] \leq \mathbf{E}\big[|g(u[0]) - g(u[s])|\big],$$

and the proof of (6) proceeds as above.

3.4 When d is not a metric.

In this section we generalize our results to the case when f maps into (Z,ρ), where (Z,ρ) is a "semimetric". This just means that ρ need not satisfy the triangle inequality; specifically, all we require of ρ is that $\rho \geq 0$, $\rho(z,z) = 0$, and $\rho(z,z') = \rho(z',z)$. (Again we do not insist that $\rho(z,z') = 0 \Rightarrow z = z'$.) Our main motivation for studying this extension is the case $Z = \mathbb{R}$ with $\rho = \rho_2(z,z') := (z-z')^2/2$. In this case $\mathbf{Vr}^{\rho_2}[f] = \mathbf{Var}[f]$ and $\mathbf{Inf}^{\rho_2}(f)$ has the meaning commonly associated with this notation for functions $f : \Omega \to \mathbb{R}$; that is, the interpretation used in, e.g., the Efron-Stein inequality or in [17].

To study the semimetric case, we simply introduce a quantity measuring the extent to which the triangle inequality fails for ρ on paths of length k. We define the *defect* of a sequence $z_0, z_1, \ldots, z_k \in Z^{k+1}$ to be $\rho(z_0, z_k) \;/\; \big(\sum_{t=1}^{k} \rho(z_{t-1}, z_t)\big)$, where $\frac{0}{0}$ is taken to be 1. We then define the *k-defect of ρ*, denoted $\mathbf{Def}_k(\rho)$, to be the maximum defect of any sequence $z_0, \ldots, z_k$. The following facts are easy to check:

- $\mathbf{Def}_1(\rho) = 1$ and $\mathbf{Def}_k(\rho)$ is nondecreasing with k.
- $\mathbf{Def}_k(\rho) \leq (\sup \rho)/(\inf \rho)$ for all k.
- $\mathbf{Def}_2(\rho) = 1$ implies that ρ satisfies the triangle inequality, which, in turn, implies that $\mathbf{Def}_k(\rho) = 1$ for all k; i.e., ρ is a metric.
- If $\rho^{1/q}$ is a metric for some $q \geq 1$, then $\mathbf{Def}_k(\rho) \leq k^{q-1}$. Thus in our motivating case with $Z = \mathbb{R}$ and $\rho(z,z') = (z-z')^2/2$ we have $\mathbf{Def}_k(\rho) \leq k$.
- If $\rho^{1/q}$ is a metric for some $q \geq 1$, then $\mathbf{Def}_k(\rho) \leq |Z|^{q-1}$ for all k.

It is easy to see how to generalize Theorems 3.1 and 3.2 for semimetrics ρ; since Theorem 3.2 is more general, we will only state its extension:

Theorem 3.3 *Let $f, g : \Omega \to (Z,\rho)$ be functions mapping an n-wise product probability space into a semimetric space, and let $\mathcal{T}$ be an RDT computing f. Let k be the length of the longest path in any DDT in $\mathcal{T}$'s support. Then*

$$\big|\mathbf{CoVr}[f,g]\big| \leq \mathbf{Def}_k(\rho)\;\sum_{i=1}^{n} \delta_i(\mathcal{T})\;\mathbf{Inf}_i(g).$$

This is the most general version of our main inequality that we state. In the semimetric setting we are most interested in, namely that of one function $f : \Omega \to (\mathbb{R}, \rho_2)$, we have the following:

Corollary 3.4 *Let $f : \Omega \to (\mathbb{R}, \rho_2)$ be a function mapping an n-wise product probability space into the real line with semimetric $\rho_2(z,z') = (z-z')^2/2$, and let $\mathcal{T}$ be an*

RDT computing f. Let k be the length of the longest path in any DDT in $\mathcal{T}$'s support. Then

$$\mathbf{Var}[f] \leq k \sum_{i=1}^{n} \delta_i(\mathcal{T}) \, \mathbf{Inf}_i^{\rho_2}(f),$$

and f has a coordinate with ρ_2-influence at least $\mathbf{Var}[f]/k^2$ (since $\sum_{i=1}^{n} \delta_i(\mathcal{T}) \leq k$).

3.5 Tightness of the inequality

Our main Theorem 3.1 can be tight; one class of DDTs for which it is tight are *read-once* decision trees. In fact, it is tight for a broader family of decision trees, which we now describe. Observe that each subtree of a decision tree below any given node can be thought of as a decision tree on the same input Ω (which may ignore some of the input variables). Say that a decision tree is *separated*, if for every two subtrees T' and T'' and every input $x \in \Omega$, if T' and T'' compute different values on x, then the sets of variables they query on input x are disjoint. Clearly, read-once trees are separated. Later, we will see that separated trees are not necessarily read-once.

To prove that Theorem 3.1 is tight for every separated tree, note that the only inequality in the proof of the theorem is (4). Suppose that T is separated, that $f(u[0]) \neq f(u[t])$ and that t is minimal with this property. Since $f(u[t]) \neq f(u[t-1])$, on input $u[t]$ the variable y_{i_t} is inspected by T. Let v' be the node of T arrived at right after reading x_{i_t} on input x, and let v'' be the node of T arrived at right after reading y_{i_t} on input $u[t]$. Then on input $u[t]$ the two subtrees of T rooted at v' and v'' calculate $f(x) = f(u[0])$ and $f(u[t])$, respectively. Since $f(u[t]) \neq f(u[0])$ and T is separated, the sets of variables examined by these two subtrees on input $u[t]$ are disjoint. In particular, $f(u[t]) = f(u[t+1]) = \cdots = f(u[s])$. Since $f(u[0]) = f(u[1]) = \cdots = f(u[t-1])$, this shows that (4) must hold as an equality when T is separated. Thus, the inequality in Theorem 3.1 holds as an equality in this case. Note that this argument shows that equality holds even if $d = \rho$ is just a semimetric.

Simple examples of read-once DDTs are those for AND : $\{-1,1\}^n \to \{-1,1\}$ and OR : $\{-1,1\}^n \to \{-1,1\}$. The simplest nontrivial balanced example is the "selection function" $\mathrm{SEL} : \{-1,1\}^3 \to \{-1,1\}$, which maps (x_1, x_2, x_3) to x_2 if $x_1 = 1$, or x_3 if $x_1 = -1$. To describe a collections of trees that are separated but not read-once, we consider (disjoint) compositions. A *disjoint composition* is a function $F = f(f_1, \ldots, f_m)$ where each f_j acts on a disjoint set of input variables, and the value of each of the input variables x_j of f is the value of f_j. An example is given by Tribes (OR of disjoint ANDs). It should be clear that a representation of a function as a disjoint composition $F = f(f_1, \ldots, f_m)$ together with a DDT for each factor function $f, f_1, \ldots, f_m$ induces a DDT for the composition; one just needs to replace each node of the tree computing f by a corresponding tree computing a function f_j. It is not too hard to check that if each of the original trees is separated, then also the tree calculating $F(f_1, \ldots, f_m)$ is separated. In particular, recursive disjoint compositions of read-once trees are separated. On the other hand, it is easy to see that the simplest nontrivial Tribes function $(x_1 \wedge x_2) \vee (x_3 \wedge x_4)$ cannot be represented by a read-once tree.

Finally, we discuss the necessity of the factor k in Corollary 3.4. Indeed, as far as we know, it may be possible to replace the factor k by an absolute constant. However we *can* show that the factor k cannot be replaced by 1. The $\{-1,1\}^3 \to (\mathbb{R}, \rho_2)$ example shown in Figure 1 demonstrates that a constant slightly greater than 1 is necessary. Except for optimizing the leaf labels in this particular tree, this is the worst example we know.

4 Questions for Future Work

- Is it possible to explain the "coincidence" that our near-tight lower bound on $\Delta(f)$ for monotone transitive functions gives a lower bound for graph properties — about $v^{4/3}$ — that essentially matches the lower bound barrier that has stood since Hajnal '91 [13]? Perhaps either the Hajnal or the Chakrabarti-Khot [9] arguments can be reframed in terms of merely transitive functions (if true of Chakrabarti-Khot, this would be quite interesting); or, perhaps graph-theoretic arguments can augment our elementary probabilistic reasoning to produce a better lower bound.

- Can our inequality in the real-valued, ρ_2 case — Corollary 3.4 — be sharpened? If the factor k could

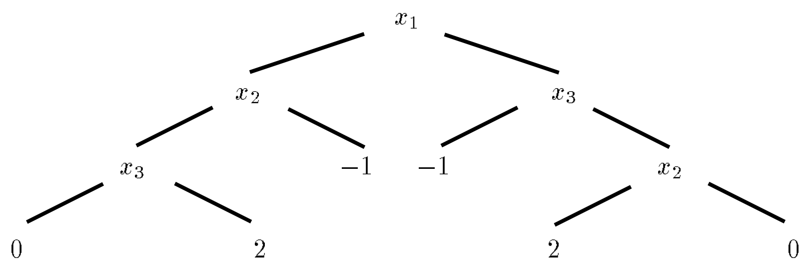

Figure 1: Left edges correspond to input variables with value -1, right edges to value 1. The function $f : \{-1,1\}^3 \to \mathbb{R}$ computed by this DDT has $\mathbf{Var}[f] = \frac{3}{2}$, but $(\delta_1(T), \delta_2(T), \delta_3(T)) = (1, \frac{3}{4}, \frac{3}{4})$ and $(\mathbf{Inf}_1^{\rho_2}(f), \mathbf{Inf}_2^{\rho_2}(f), \mathbf{Inf}_3^{\rho_2}(f)) = (\frac{1}{8}, \frac{7}{8}, \frac{7}{8})$, where $\rho_2(x,y) = (x-y)^2/2$, so $\sum_{i=1}^3 \delta_i(T)\ \mathbf{Inf}_i^{\rho_2}(f) = \frac{23}{16} < \frac{3}{2}$.

be replaced by a universal constant, this would be a very strong variant of the Efron-Stein inequality.

- What other applications might our main inequality have? We suggest there might be applications in computational learning theory or in the theory of random graphs.

5 Acknowledgments

We would like to thank Andris Ambainis, Laci Lovasz, and Avi Wigderson for helpful discussions.

References

[1] R. Beals, H. Buhrman, R. Cleve, M. Mosca, and R. de Wolf. Quantum lower bounds by polynomials. *Journal of the ACM*, 48(4):778–797, 2001.

[2] M. Ben-Or and N. Linial. Collective coin flipping. In *Proceedings of the 26th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 408–416, 1985.

[3] I. Benjamini, O. Schramm, and D. Wilson. Balanced boolean functions that can be evaluated so that every input bit is unlikely to be read. In *Proceedings of the 37th Annual Symposium on Theory of Computing (STOC)*, 2005.

[4] M. Best, P. van Emde Boas, and H. Lenstra. A sharpened version of the Aanderaa-Rosenberg conjecture. Technical Report Report ZW30/74, Mathematisch Centrum Amsterdam, 1974.

[5] M. Blum and R. Impagliazzo. Generic oracles and oracle classes. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, pages 118–126, 1987.

[6] B. Bollobas. *Combinatorics: Set Systems, Hypergraphs, Families of Vectors and Combinatorial Probability*. Cambridge University Press, 1986.

[7] J. Bourgain and G. Kalai. Influences of variables and threshold intervals under group symmetries. *GAFA*, 7:438–461, 1997.

[8] H. Buhrman and R. de Wolf. Complexity measures and decision tree complexity: a survey. *Theoretical Computer Science*, 288(1):21–43, 2002.

[9] A. Chakrabarti and S. Khot. Improved lower bounds on the randomized complexity of graph properties. In *Proceedings of the 28th International Colloquium on Automata, Languages and Programming*, pages 285–296, 2001.

[10] B. Efron and C. Stein. The jackknife estimate of variance. *Annals of Statistics*, 9:586–596, 1981.

[11] E. Friedgut, J. Kahn, and A. Wigderson. Computing graph properties by randomized subcube partitions. In *Proceedings of the 6th International Workshop on Random and Approximation Techniques*, pages 105–113, 2002.

[12] E. Friedgut and G. Kalai. Every monotone graph property has a sharp threshold. *Proceedings of the AMS*, 124:2993–3002, 1996.

[13] A. Hajnal. An $\Omega(n^{4/3})$ lower bound on the randomized complexity of graph properties. *Combinatorica*, 11:131–143, 1991.

[14] J. Kahn, G. Kalai, and N. Linial. The influence of variables on boolean functions. In *Proceedings of the 29th Annual Symposium on Foundations of Computer Science*, pages 68–80, 1988.

[15] V. King. Lower bounds on the complexity of graph properties. In *Proceedings of the 20th Annual Symposium on Theory of Computing*, pages 468–476, 1988.

[16] G. Margulis. Probabilistic characteristics of graphs with large connectivity. *Prob. Peredachi Inform.*, 10:101–108, 1974.

[17] E. Mossel, R. O'Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *Proceedings of the 46th Annual Symposium on Foundations of Computer Science (FOCS)*, 2005.

[18] N. Nisan and M. Szegedy. On the degree of Boolean functions as real polynomials. In *Proceedings of the Twenty-Fourth Annual Symposium on Theory of Computing*, pages 462–467, 1992.

[19] R. O'Donnell and R. Servedio. Learning monotone functions from random examples in polynomial time. Manuscript, 2005.

[20] R. Rivest and J. Vuillemin. On recognizing graph properties from adjacency matrices. *Theoretical Computer Science*, 3:371–384, 1976.

[21] L. Russo. On the critical percolation probabilities. *Z. Wahrsch. werw. Gebiete*, 43:39–48, 1978.

[22] M. Saks and A. Wigderson. Probabilistic boolean decision trees and the complexity of evaluating game trees. In *Proceedings of the 27th Annual Symposium on Foundations of Computer Science*, pages 29–38, 1986.

[23] O. Schramm and J. Steif. Quantitative noise sensitivity and exceptional times for percolation. Manuscript, 2005.

[24] Y. Shi. Lower bounds of quantum black-box complexity and degree of approximating polynomials by influence of boolean variables. *Information Processing Letters*, 75(1-2):79–83, 2000.

[25] M. Snir. Lower bounds for probabilistic linear decision trees. *Theoretical Computer Science*, 38:69–82, 1985.

[26] J. M. Steele. An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics*, (14):753–758, 1986.

[27] M. Talagrand. On Russo's approximate 0-1 law. *The Annals of Probability*, 22(3):1576–1587, 1994.

[28] A. Yao. Probabilistic computations: towards a unified measure of complexity. In *Proceedings of the 18th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 222–227, 1977.

[29] A. Yao. Lower bounds to randomized algorithms for graph properties. In *Proceedings of the 28th Annual Symposium on Foundations of Computer Science*, pages 393–400, 1987.