The Hardness of Being Private

Anil Ada*, Arkadev Chattopadhyay†, Stephen Cook†, Lila Fontes†, Michal Koucký‡, Toniann Pitassi†
† *Department of Computer Science, University of Toronto, Toronto, Canada*
email:{arkadev, sacook, fontes, toni}@cs.toronto.edu
* *Department of Computer Science, McGill University, Montreal, Canada*
email: aada@cs.mcgill.ca
‡ *Institute of Mathematics, Academy of Sciences, Prague, Czech Republic*
email: koucky@math.cas.cz

***Abstract*—In 1989 Kushilevitz [1] initiated the study of information-theoretic privacy within the context of communication complexity. Unfortunately, it has been shown that most interesting functions are not privately computable [1], [2]. The unattainability of perfect privacy for many functions motivated the study of *approximate privacy*. In [5], [7], they define notions of worst-case as well as average-case approximate privacy, and present several interesting upper bounds, and some open problems for further study. In this paper, we obtain asymptotically tight bounds on the tradeoffs between both the worst-case and average-case approximate privacy of protocols and their communication cost for Vickrey-auctions.**

Further, we relate the notion of average-case approximate privacy to other measures based on information cost of protocols. This enables us to prove exponential lower bounds on the subjective approximate privacy of protocols for computing the Intersection function, independent of its communication cost. This proves a conjecture of Feigenbaum et al.

***Keywords*-privacy, communication complexity, Vickrey auctions**

I. Introduction

Privacy in a distributed setting is an increasingly important problem. A key application is the setting of combinatorial auctions where many agents have private information (e.g., their preferences) but would like to compute a function of their inputs without revealing any of their private information. There is a large body of research examining which functions can be computed securely, and how. Many of these results rely on an assumption, such as a computational complexity assumption, or the assumption that more than some fixed fraction m of the players are trustworthy, or the assumption that the auctioneer (a 3^{rd} party) is trustworthy. These assumptions limit the usefulness of such study. As Brandt and Sandholm point out, privacy which is based on an assumption of hardness can easily become outdated as computers become faster and more powerful; security parameters (like key length) need to be continuously updated to cope with increasing computational power [2]. Auctions are a natural setting where we would doubt the trustworthiness of fellow participants or an auctioneer. We nevertheless would like to compute on the internet. In this work, we focus on situations where each player is deterministic and honest but curious. Honest, because they obey the rules of the game. Curious, as they do not miss any opportunity to gain knowledge about others' input.

In 1989, Kushilevitz [1] initiated the study of *information-theoretic* privacy in communication complexity, which is an appealing direction because it does not rely on computational assumptions discussed above. Informally, a multi-player communication protocol for computing a function $f(x, y)$ is private if each player does not learn any additional information (in an information theoretic sense) beyond what follows from knowing his/her private input, and the function value $f(x, y)$. A complete characterization of the privately computable functions was given, but unfortunately, early work ruled out private protocols for most interesting functions [1], [2]. For example, second-price auctions are not possible with more than two participants, and are extremely inefficient even in the setting of two bidders [3], [2].

The unattainability of perfect privacy for many functions motivated the study of *approximate* privacy. Most relevant to our work is the study of Klauck [4] and the more recent work of Feigenbaum, Jaggard and Schapira [5]. The relaxation from perfect to approximate privacy is appealing because it renders more functions computable privately, and more closely mirrors real-world situations in which *some* privacy loss may be acceptable. On the other hand, it is more subtle to capture the notion of approximate privacy. While most reasonable definitions of perfect privacy turn out to be equivalent, this is not quite the case with approximate privacy. In particular, the measures of Klauck [4] and Feigenbaum et al. [5] are different and each has its own advantage and characteristics. Our work here is primarily motivated by recent work [5], [6]. A second motivation is to understand the connections between the two.

In the two player setting, let $f(x, y)$ be a function, and let P be a two-player deterministic communication protocol for f. The privacy loss (or privacy approximation ratio, PAR) on the input (x, y) with respect to P is defined to be the size of the monochromatic region containing (x, y) divided by the size of the protocol-induced rectangle containing (x, y): $\text{PAR}(x, y) = \frac{|f^{-1}(x,y)|}{|P(x,y)|}$. The worst-case privacy loss of

protocol P is the maximum privacy loss over all inputs (x, y), and the worst-case privacy loss of the function f is then the minimum privacy loss over all protocols for f. Perfect privacy of a protocol (as defined in 1989) requires that the privacy approximation ratio (PAR) is 1 for all inputs. (This definition easily extends to the multi-player setting.)

Under this relaxed notion of privacy, things are much more interesting [5], [7], [6]. For example, Feigenbaum et al. study the Vickrey auction problem, and reveal a possible *inherent* tradeoff between privacy and communication complexity: they describe a family of protocols such that the privacy loss approaches 1 (perfect privacy) as the length of the protocol approaches exponential. They also study several prominent boolean functions with respect to approximate privacy.

Feigenbaum et al. consider an *average-case* notion of approximate privacy as well. In this setting, we are interested in the average privacy loss over a distribution on inputs. Here they describe a protocol for Vickrey auction that achieves exponentially smaller average-case PAR than the worst-case PAR achievable by any known protocol. A similar protocol was described by Klauck [4].

Our Contributions

In this paper, we present several new lower bounds on the communication cost for achieving privacy and establish relationships between approximate privacy and several other known measures.

First, we prove that there is an inherent tradeoff between privacy and communication complexity, by proving a privacy/communication complexity tradeoff lower bound for the Vickrey auction problem. This shows that the upper bounds presented in [5] are essentially tight. [5] provided a lower bound only for the special case of bisection-type protocols.

Theorem 1: For all n, for all p, $2 \leq p \leq n/4$, any deterministic protocol for the two-player Vickrey auction problem on inputs of length n obtaining privacy loss (PAR) less than 2^{p-2} has length at least $2^{\frac{n}{4p}}$.

This lower bound is technically interesting as it deals with super-linear communication protocols. The usual communication complexity techniques aim at protocols that are at most linear in their input size.

Our second contribution demonstrates a similar type of tradeoff for the case of average-case approximate privacy. We prove an asymptotically tight lower bound on the average-case approximate privacy of the Vickrey auction problem, showing that the upper bounds from [5] are essentially tight. This generalizes the result of [6] for Vickrey auctions. Again, [5] provided lower bounds only for the special case of bisection-type protocols.

Theorem 2: For all $n, r \geq 1$, any deterministic protocol of length at most r for the two-player 2^n-Vickrey auction problem has average-case PAR at least $\Omega(\frac{n}{\log(r/n)})$ (over the uniform distribution of inputs).[1]

Our lower bounds show that the approximate privacy of any polynomial length protocol is still as large as $\Omega(n/(\log n))$. Indeed, such superlinear protocols have been devised by Klauck [4] who proved upper bounds for his measure of approximate-privacy. To the best of our knowledge, Theorem 2 provides the first (tight) lower bounds on the communication cost of achieving good approximate privacy for Vickrey auctions. The proof of the theorem relates the loss of privacy to a certain Ball Partition Problem that may be of independent interest.

Furthermore, we modify the average-case privacy approximation measure of Feigenbaum et al. Our modification provides a rather natural measure that was disregarded in [5], but coincides with that of Feigenbaum et al. in the case of uniform distribution on the inputs. Our modified measure has several advantages. It allows natural alternative characterizations, and it can be directly related to the privacy measure of Klauck. We can quantitatively connect Klauck's privacy measure to well studied notions of *(internal) information cost* in communication complexity. This allows us to prove a new lower bound on the average-case subjective privacy approximation measure of Feigenbaum et al. [5], and answers affirmatively a conjecture from their paper.

Theorem 3: For all $n \geq 1$, and any protocol P computing the Set Intersection INTERSEC$_n$ the average-case subjective PAR is exponential in n under the uniform distribution:

$$\mathrm{avg}_{\mathcal{U}} \, \mathrm{PAR}^{sub}(P) = 2^{\Omega(n)}.$$

We contend that any of the mentioned measures could serve as a reasonable measure of privacy. Indeed, each of the measures seems to exhibit advantages over the other ones in some scenario so each of the measures captures certain aspect of privacy. For example, the English auction protocol for Vickrey auction achieves perfect privacy (under any measure) but at exponential communication cost. On the other hand, the Bisection protocol achieves linear average-case PAR with merely linear communication cost. However, the difference between these two protocols is not reflected well in Klauck's privacy measure, where both protocols lose constant number of bits on the average.

Outline of Paper

In Section II, we provide our basic notation and background on information theory. In Section III, we review the notion of privacy approximation ratio, and in Section III-A, we review the Vickrey auction problem. In Section III-B, we present our lower bound tradeoff for worst-case privacy of Vickrey auctions. In Section IV, we present our lower bound on average-case PAR for Vickrey auctions, and discuss

[1] Under the original definition [5] or our alternate Definition 11.

the relationship between average-case PAR and information cost, deriving several new results from this relationship.

II. Preliminaries

In this section, we review our basic notations and concepts. For a positive integer k, we let $[k] = \{1, 2, \ldots, k\}$. We assume that the reader is familiar with communication complexity (see [8] for more background.) We will use the following notation. Given $f : X \times Y \to Z$, each input (x, y) is associated with the **region** $R_{x,y}$ of all inputs in the preimage of $f(x, y)$, i.e., $R_{x,y} = \{(x', y') \in X \times Y | f(x, y) = f(x', y')\}$. For any value $z \in Z$ we let $R_z = f^{-1}(z)$ be the preimage of z. The set of all regions of function f is $\mathcal{R}(f) = \{R_{x,y} : (x, y) \in X \times Y\}$. Let P be a communication protocol for the function f. For inputs $(x, y) \in X \times Y$ we let $\Pi_P(x, y)$ denote the transcript of the protocol on input x given to Alice and y given to Bob. We associate the input (x, y) with the **protocol-induced rectangle** $P_{x,y}$ of all inputs which yield the same transcript: $P_{x,y} = \{(x', y') \in X \times Y : \Pi_P(x, y) = \Pi_P(x', y')\}$. Note that $P_{x,y} \subseteq R_{x,y}$ as we assume that P correctly computes f.

A. Information theoretic notions

Information theory provides a highly intuitive and powerful calculus to reason about random variables. We need the following basic notions from this theory whose proofs can be found in any standard textbook on the subject (see for example Cover and Thomas [9]).

For any random variable $\mathbf{X}$, we denote its probability distribution over its range $\mathcal{X}$ by μ_X. The entropy of $\mathbf{X}$, denoted by $H(\mathbf{X})$, is defined as follows:

$$\begin{aligned} H(\mathbf{X}) &= -\sum_{x \in \mathcal{X}} \Pr_{\mu_X}\left[\mathbf{X} = x\right] \log_2 \left(\Pr_{\mu_X}\left[\mathbf{X} = x\right] \right) \\ &= -\mathbb{E}_{\mu_X}\big[\log(\mu_X(x))\big] \end{aligned}$$

Let $\mathbf{Y}$ be another random variable. For any y in the range of $\mathbf{Y}$, $H(\mathbf{X}|\mathbf{Y} = y)$ is defined as just the entropy of $\mathbf{X}$ under the conditional distribution, i.e. $H(\mathbf{X}|\mathbf{Y} = y) \equiv_{\text{def}}$

$$-\sum_{x \in \mathcal{X}} \Pr\left[\mathbf{X} = x \,|\, \mathbf{Y} = y\right] \log \left(\Pr\left[\mathbf{X} = x \,|\, \mathbf{Y} = y\right] \right).$$

Extending the above naturally, we define the notion of conditional entropy $H(\mathbf{X}|\mathbf{Y})$ as

$$H(\mathbf{X}|\mathbf{Y}) \equiv_{\text{def}} \mathbb{E}_{\mu_Y}\big[H(\mathbf{X}|\mathbf{Y} = y)\big].$$

As intuition suggests, conditioning a random variable $\mathbf{X}$ on another random variable $\mathbf{Y}$ cannot increase its uncertainty on the average. Formally,

Fact 4: For any two random variables $\mathbf{X}$ and $\mathbf{Y}$, $H\big(\mathbf{X}|\mathbf{Y}\big) \leq H\big(\mathbf{X}\big)$.

The mutual information between $\mathbf{X}$ and $\mathbf{Y}$, denoted by $I(\mathbf{X} : \mathbf{Y})$, is defined as $H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y})$. It is straightforward to verify that mutual information is a symmetric quantity, i.e. $I(\mathbf{X} : \mathbf{Y}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}|\mathbf{X}) = I(\mathbf{Y} : \mathbf{X})$. Fact 4 implies that mutual information between two random variables is always non-negative. Just like entropy, one can define the conditional mutual information between random variables: let $\mathbf{Z}$ be another random variable with range $\mathcal{Z}$.

$$\begin{aligned} I(\mathbf{X} : \mathbf{Y} \,|\, \mathbf{Z}) &= H(\mathbf{X} \,|\, \mathbf{Z}) - H(\mathbf{X} \,|\, \mathbf{Y}, \mathbf{Z}) \\ &= \mathbb{E}_{\mu_Z}\big[I(\mathbf{X} : \mathbf{Y} \,|\, \mathbf{Z} = z)\big]. \end{aligned}$$

We will also need the following simple claim:

Claim 5: Let $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$ be any random variables. Then, $\big|I(\mathbf{X} : \mathbf{Y}|\mathbf{W}) - I(\mathbf{X} : \mathbf{Y}\,|\mathbf{W}, \mathbf{Z})\big| \leq H(\mathbf{Z})$.

Proof: First, notice that $I(\mathbf{X} : \mathbf{Y} \,|\, \mathbf{W}) - I(\mathbf{X} : \mathbf{Y} \,|\, \mathbf{W}, \mathbf{Z}) = \big(H(\mathbf{X}|\mathbf{W}) - H(\mathbf{X}|\mathbf{W}, \mathbf{Z})\big) - \big(H(\mathbf{X}|\mathbf{W}, \mathbf{Y}) - H(\mathbf{X}|\mathbf{Y}, \mathbf{W}, \mathbf{Z})\big)$. Using symmetry of information, the first bracketed quantity is $I(\mathbf{X} : \mathbf{Z} \mid \mathbf{W}) \leq H(\mathbf{Z})$, and the second bracketed quantity is $\mathbb{E}_{\mu_Y}\big[I(\mathbf{Z} : \mathbf{X} \,|\, \mathbf{W}, \mathbf{Y} = y)\big] \leq H(\mathbf{Z})$. ■

III. Worst-case privacy approximation ratio

In this paper, we are concerned with privacy-preserving communication complexity. A perfectly private communication protocol for f will reveal only the output of f and no additional information. Every two inputs (x, y) and (x', y') such that $f(x, y) = f(x', y')$ should be indistinguishable from each other [1], [10]. Approximate privacy provides a measure of how much indistinguishability has been lost. These notions are formalized as follows.

The following definition captures the privacy loss of a communication protocol with respect to a third party observer (eavesdropper) who overhears the messages sent between the players. This measure is referred to as **objective**.

Definition 6: [5] For an input (x, y), its privacy approximation ratio is $\text{PAR}(P, x, y) = \frac{|R_{x,y}|}{|P_{x,y}|}$. A protocol P for a function f on $X \times Y$ has **worst-case objective privacy approximation ratio** (PAR) defined by

$$\text{PAR}(P) = \max_{(x,y)} \text{PAR}(P, x, y).$$

Often we do not specify the protocol P when it is clear from context.

The PAR measure of privacy can be extended to **subjective PAR**, which measures the privacy that the players lose *to each other*.

Definition 7: [5] A protocol P for a function f on $X \times Y$ has **worst-case subjective privacy approximation ratio** (PAR^{sub}) defined by:

$$\text{PAR}^{sub}(P) = \max\left\{ \max_{(x,y)} \frac{|R_{x,y} \cap X \times \{y\}|}{|P_{x,y} \cap X \times \{y\}|}, \max_{(x,y)} \frac{|R_{x,y} \cap \{x\} \times Y|}{|P_{x,y} \cap \{x\} \times Y|} \right\}.$$

Previous work by Kushilevitz gave a combinatorial characterization of the functions f which are computable with perfect privacy $\mathrm{PAR} = 1$ [1]. This set unfortunately excludes most auctions [2], as well as many basic functions of interest in theoretical computer science, e.g., greater than [11] and set intersection and disjointness [7].

As many functions are not computable with perfect privacy, it is natural to investigate the following general question for a function f: is f privately computable, and how much communication is necessary to achieve PAR less than some number c? In the next section, we focus on the case of Vickrey auctions which is one of the most studied functions in this context.

A. Vickrey auctions

Vickrey auctions (also known known as 2^{nd}-price auctions) arise in mechanism design, and are a canonical example of a *truthful* mechanism: neither player has incentive to cheat, as long as the auction is computed correctly. For a positive integer N, the N-Vickrey auction is defined as $f : X \times Y \to Z \times \{A, B\}$ where $X = Y = Z = \{1, 2, \ldots, N\}$ and

$$f(x,y) = \begin{cases} (x, B), & \text{if } x \leq y \\ (y, A) & \text{if } y < x \end{cases}$$

Two players, Alice and Bob, have private values x and y, respectively. These private values indicate the amount of money that the item is worth to each of them. If $x \leq y$, then Bob wins, and the price that he pays is x. (Thus, $f(x, y) = (x, B)$ means that Bob wins and pays x for the item.) Similarly, if $x > y$, then Alice wins, and the price that she pays is y. This mechanism is also called "2^{nd}-price auction" because the winner's price is the 2^{nd}-highest bid. Vickrey auctions remain truthful for more than two players, but are not computable with perfect privacy ($\mathrm{PAR} = 1$) [2] for more than two players.

The matrix M_f of the 2^n-Vickrey auction looks like:

	1	2	3	4	…	2^n-1	2^n
1	$(1,B)$	$(1,B)$	$(1,B)$	$(1,B)$	…	$(1,B)$	$(1,B)$
2	$(1,A)$	$(2,B)$	$(2,B)$	$(2,B)$	…	$(2,B)$	$(2,B)$
3	$(1,A)$	$(2,A)$	$(3,B)$	$(3,B)$	…	$(3,B)$	$(3,B)$
4	$(1,A)$	$(2,A)$	$(3,A)$	$(4,B)$	…	$(4,B)$	$(4,B)$
⋮	⋮	⋮	⋮	⋮	…	⋮	⋮
2^n-1	$(1,A)$	$(2,A)$	$(3,A)$	$(4,A)$	…	$(2^n-1,B)$	$(2^n-1,B)$
2^n	$(1,A)$	$(2,A)$	$(3,A)$	$(4,A)$	…	$(2^n-1,A)$	$(2^n,B)$

Figure 1. The matrix M_f for 2^n-Vickrey auction.

Perfect privacy for two-player Vickrey auctions is achieved by the successive English bidding protocol, in which bids start at 1 and increase by 1 in each round, and the first player to drop out of bidding reveals his entire private value. (Note that this incurs no loss of privacy, since that value is part of the function output.) The protocol tree for this protocol is given in Figure 2. This protocol takes 2^{n+1} rounds for the 2^n-Vickrey auction, and is known to be the only protocol which obtains perfect privacy $\mathrm{PAR} = 1$ for Vickrey auctions [1].

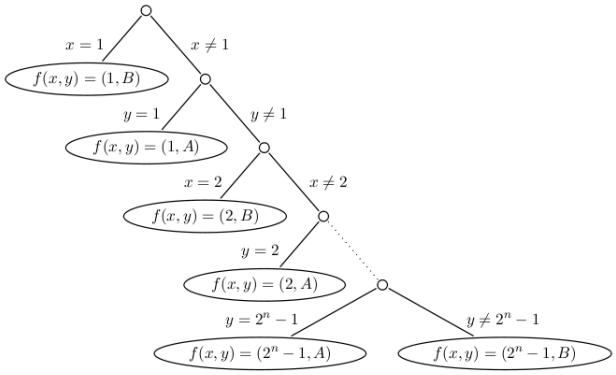

Figure 2. The protocol tree for an English auction computing f.

Notice that the range of f is of size 2^{n+1} and that f is surjective, so that there must be at least 2^{n+1} distinct leaves in any protocol tree for f. Thus any protocol for f requires at least $n + 1$ rounds. An example of such a protocol is the Bisection Protocol that proceeds by binary search on an interval containing the smaller input [5]. Bisection Protocol obtains $\mathrm{PAR} = 2^n$, the worst possible loss of privacy for this function.

These two extremes – on the one hand $\mathrm{PAR} = 1$ at exponential communication cost, and on the other, exponential PAR at linear communication cost – suggest that there is a tradeoff between privacy and communication for Vickrey auctions. The structure of the function itself suggests this tradeoff as well. Any move which differs from the English protocol must divide some monochromatic region into two pieces. Thus inputs in the same monochromatic region are distinguishable by the protocol, and some privacy is lost.

Different PAR is achievable depending on the nature of the protocol. Feigenbaum et al. examine a family of Bisection-type protocols [5] and use average-case PAR to differentiate amongst them. Such protocols obtain worst-case PAR varying from 1 to 2^n, inversely related to their length. This observation inspired the results below.

B. Worst-case lower bound for Vickrey auction

The two algorithms discussed in the previous section suggest that any protocol computing Vickrey auctions should have a tradeoff between length and privacy. Protocol steps which resemble those of the ascending English bidding protocol partition the inputs in an unbalanced way, so that most inputs follow one branch of the protocol tree, and few inputs follow the other branch. Such steps preserve privacy but do not make much progress. (In an imbalanced partition, on the larger side the protocol still has a lot of work to do in order to compute the function.) On the other hand, protocol steps that resemble binary search partition the inputs in a

nearly balanced way. Such steps make good progress, but are bad for privacy. (Dividing the remaining inputs in half increases the PAR by a factor of 2.) This is the intuition behind Theorem 1, stated again below.

Theorem 1: For every n and p, $2 \leq p \leq n/4$, every deterministic protocol for 2-player Vickrey auctions obtaining $\mathrm{PAR} < 2^{p-2}$ must be of length at least $2^{\frac{n}{4p}}$.

Here the variable p serves as a parameter, explicitly linking the protocol length to the achievable PAR. For instance, if we put $p = \sqrt{n}$, then we conclude by Theorem 1 that either the protocol communicates $2^{\Omega(\sqrt{n})}$ bits in the worst case, or the worst-case privacy loss is $2^{\Omega(\sqrt{n})}$. This theorem shows that for Vickrey auctions, there is an inherent tradeoff between communication complexity and privacy.

Proof: We will assume without loss of generality that in the protocol, the players take turns and send one bit per message. (Any protocol can be put into this form by at most doubling the length of the protocol.) Moreover, our protocol is assumed to be deterministic and to have zero error.

The Vickrey auction function has a corresponding matrix M such that entry (x, y) of the matrix is the value of the Vickrey auction on inputs (x, y) (Figure 1). A submatrix of M is called a *rectangle*; a rectangle is "monochromatic" if the matrix is constant on inputs in that submatrix. Every communication protocol can be visualized as a binary decision tree [12]. Each node v of the tree is associated with a rectangle (submatrix) $T(v) = T_A(v) \times T_B(v) \subseteq X \times Y$. The root node r is associated with the entire matrix $T_A(r) \times T_B(r) = X \times Y = M$. Each leaf node l is associated with a monochromatic submatrix $T_A(l) \times T_B(l)$. Each internal node v has two children, v_0 and v_1. If the protocol calls for Alice to speak at node v, then the bit sent by Alice at v induces a partition of $T_A(v)$ into two pieces, $T_A(v_0)$ and $T_A(v_1)$. The submatrix associated with v_0 is $T_A(v_0) \times T_B(v)$, and the submatrix associated with v_1 is $T_A(v_1) \times T_B(v)$. Similarly if Bob speaks at node v, then the submatrix associated with v_0 is $T_A(v) \times T_B(v_0)$ and the submatrix associated with v_1 is $T_A(v) \times T_B(v_1)$.

Traversing the tree from the root to a leaf l generates a transcript of the bits of communication sent for some input $(x, y) \in T_A(l) \times T_B(l)$. The depth of the tree is the worst-case communication cost of the protocol.

Any deterministic protocol consists of a series of partitions of the matrix M into rectangles. The resulting protocol-induced tiling of the matrix M is a partition into monochromatic rectangles, which are precisely the rectangles associated with the leaves of the protocol's decision tree.

For every correct communication protocol, we describe an adversary strategy that follows a path through the protocol tree and finds some input (x, y) such that either: (i) the privacy loss of (x, y) is large i.e., $\mathrm{PAR}(x, y) \geq 2^{p-2}$, or (ii) the communication protocol on (x, y) requires at least $(\ln 2)(\frac{n}{2} - p)2^{n/4p}$ bits to compute.

Let M denote the matrix corresponding to the Vickrey auction problem, as drawn in Figure 1. Bob wins for inputs in the horizontal regions; Alice wins for inputs in the vertical regions. Fix a communication protocol, and corresponding protocol tree, P. For every node v in P that our adversary strategy selects, we will maintain three sets: $S(v), A^L(v), B^L(v) \subseteq [2^n]$. At node v, the adversary will be interested in tracking the privacy loss on the set of inputs $S(v) \times S(v)$. The privacy loss for these inputs will be measured with the help of the two auxiliary sets $A^L(v)$ and $B^L(v)$, respectively.

Initially, at the root r of the protocol tree, $S(r) = [2^{n-p}]$. This initial set of inputs $S(r) \times S(r)$ are the "small" inputs that sit in the upper left submatrix of M. As we move down the protocol tree, we will update $S(v)$ so that it is always a subset of $[2^{n-p}] \cap T_A(v) \cap T_B(v)$. We are interested in these small inputs since the regions that they are contained in are very large, and thus have the potential to incur a large (exponential) privacy loss.

The set $A^L(v)$ is a subset of $T_A(v)$, and similarly $B^L(v)$ is a subset of $T_B(v)$. The sets $A^L(r)$ and $B^L(r)$ are initially $[2^n] \backslash [2^{n-p}]$, the "large" inputs. At vertex v, the set $A^L(v)$ describes the set of large inputs of Alice that have survived so far; thus $A^L(v) = T_A(v) \cap [2^n] \backslash [2^{n-p}]$. Similarly, $B^L(v)$ describes the set of large inputs of Bob that have survived so far; thus $B^L(v) = T_B(v) \cap [2^n] \backslash [2^{n-p}]$. As we traverse the protocol tree, these sets track the loss of privacy for Alice and Bob (respectively) on inputs in $S(v) \times S(v)$.

We can measure the loss of privacy *so far* in the protocol. For any $(x, y) \in T(v)$,

$$\mathrm{PAR}_v(x, y) = \frac{|R_{x,y}|}{|R_{x,y} \cap T(v)|}.$$

If v is a leaf, then for any $(x, y) \in T(v)$, $\mathrm{PAR}_v(x, y) = \mathrm{PAR}(x, y)$. The following simple claim will be useful:

Claim 8: $\forall (x, y) \in T(v)$, $\mathrm{PAR}(x, y) \geq \mathrm{PAR}_v(x, y)$.

In particular the following fact is crucial to our argument. For any (x, y) in $S(r) \times S(r) \cap T(v)$, if (x, y) is in a vertical region ($y < x$, a win for Alice), then

$$\mathrm{PAR}(x, y) = \frac{|R_{x,y}|}{|P_{x,y}|} \geq \mathrm{PAR}_v(x, y) \geq \frac{2^n - 2^{n-p}}{|A^L(v)| + 2^{n-p}}.$$

This holds because $|R_{x,y}| \geq 2^n - 2^{n-p}$ and $|R_{x,y} \cap T(v)| \leq |A^L(v)| + 2^{n-p}$. Similarly, if $(x, y) \in S(r) \times S(r)$ is in a horizontal region ($x \leq y$, a win for Bob), then

$$\mathrm{PAR}(x, y) \geq \mathrm{PAR}_v(x, y) \geq \frac{2^n - 2^{n-p}}{|B^L(v)| + 2^{n-p}}.$$

The above inequality shows how $A^L(v)$ and $B^L(v)$ track the privacy loss of inputs $S(v) \times S(v)$: for those inputs $(x, y) \in S(v) \times S(v)$ where Alice wins, the privacy loss for (x, y) increases as $A^L(v)$ decreases, and similarly for

those inputs where Bob wins, the privacy loss increases as $B^L(v)$ decreases.

Adversary Strategy: We are now ready to describe the adversary strategy. There are two cases, depending on whether it is Alice's or Bob's turn to send a message. We will first describe the case where at node v, it is Alice's turn to speak. Alice sends Bob some bit b which partitions her inputs $T_A(v)$ into two pieces. Since $S(v)$ and $A^L(v)$ are always subsets of $T_A(v)$, this induces a partition of $S(v)$ into $S_0(v)$ and $S_1(v)$ and $A^L(v)$ into $A^L_0(v)$ and $A^L_1(v)$.

Let $\alpha = 1 - 2^{\frac{-n}{4p}}$. We determine if a step made *progress* or was *useless* in the following way:

- If $(1-\alpha)|S(v)| \leq |S_0(v)| \leq \alpha|S(v)|$ (hence $(1-\alpha)|S(v)| \leq |S_1(v)| \leq \alpha|S(v)|$), then we say this step made **progress** on $S(v)$. In this case, the set $S(v)$ is partitioned into roughly balanced pieces. Select i such that $|A^L_i(v)| \leq \frac{1}{2}|A^L(v)|$.
- Otherwise, pick i such that $|S_i(v)| \geq \alpha|S(v)|$. In this case, we call it a **useless** step.

We update sets in the obvious way: if w is the new node in the protocol tree that we traverse to, then $S(w) = S_i(v)$ and $A^L(w) = A^L_i(v)$.

The second case is when it is Bob's turn to speak. Our adversary strategy is entirely symmetric. Now $T_B(v)$ is partitioned into two pieces, inducing a partition of $S(v)$ into $S_0(v)$ and $S_1(v)$, and a partition of $B^L(v)$ into $B^L_0(v)$ and $B^L_1(v)$. We pick i as above, but with A^L_i replaced with B^L_i.

The strategy continues as described above, traversing the protocol tree until one of the two events happens for the first time:

- Alice (or Bob) has made p progress steps, so $A^L(v)$ (or $B^L(v)$) has been halved at least p times.
- The strategy reaches a leaf node, and can go no further.

This completes the description of the strategy.

The following are the two main ideas in analyzing our strategy.

Lemma 9: Let our strategy reach node v and find Alice (or Bob) took p progress steps on the way. Then, for each $(x,y) \in S(v) \times S(v)$ such that $x > y$ (or $x \leq y$) $\mathrm{PAR}_v(x,y) \geq 2^{p-2}$.

We can exit our strategy at this point and invoke Claim 8 to finish the argument. In the other case, we make the following claim:

Lemma 10: If our strategy reaches a leaf node v without Alice or Bob taking p progress steps, then for every $(x,y) \in T(v)$, the protocol communicates at least $2^{n/4p}$ bits.

Thus, we would conclude that in this case the cost of the protocol is larger than $n2^{n/4p}$. Hence, all that remains to finish our argument is to prove Lemma 9 and Lemma 10.

Proof of Lemma 9: Let r be the root node of our protocol tree. For each input $(x,y) \in S(r) \times S(r)$, note that $R_{x,y} \geq 2^n - 2^{n-p}$ and $|A^L(r)| = 2^n - 2^{n-p}$. Let φ be the path in the protocol tree from r to v that our strategy chooses such that Alice takes p progress steps along φ. Consider any pair of adjacent nodes u, w in path φ such that Alice makes progress in going from u to w. Then, by definition of our strategy, $|A^L(w)| \leq \frac{1}{2}|A^L(u)|$. Hence, $|A^L(v)| \leq \frac{1}{2^p}|A^L(r)|$. Thus, for inputs (x,y) in $A^L(v) \times A^L(v) \subseteq S(v) \times S(v)$ on which Alice would win, claim 8 yields:

$$\mathrm{PAR}_v(x,y) \geq \frac{2^n - 2^{n-p}}{\frac{2^n - 2^{n-p}}{2^p} + 2^{n-p}} \geq 2^{p-2}$$

The analysis when Bob makes p progress steps proceeds very similarly. ∎

Proof of Lemma 10: The strategy reaches a leaf node v traversing a path φ, and $|S(v)| = 1$. (If $|S(v)| > 1$, then there is more than one possible answer, and so the computation is not yet finished.) In this case, Alice and Bob each took fewer than p progress steps. Let q be the total number of useless steps followed to get to v. (The protocol is at most $2p + q$ long.) On each progress step (u,w) in path φ, by definition, $|S(w)| \geq (1-\alpha)|S(u)|$. On each useless step (u,w), the updated size of $|S(w)| \geq \alpha|S(u)|$. This gives a lower bound on the size of set $S(v)$. Hence $|S(v)| \geq 2^{n-p}(1-\alpha)^{2p}\alpha^q$.

Assume that $q < 2^{\frac{n}{4p}}$. Then $|S(v)| \geq 2^{n-p}(1-\alpha)^{2p}\alpha^q \geq 1$, contradicting the fact that v is a leaf node where the protocol ends. ∎

Thus the strategy proves Theorem 1, either by finding some large loss of privacy or by finding an input on which the protocol takes exponentially many steps. ∎

Note. The tradeoff of Theorem 1 holds for both the objective PAR and subjective PAR. For Vickrey auctions they coincide, because all regions are rectangles with width or depth one.

IV. Average-case PAR

In this section we consider the average-case privacy approximation ratio. For a probability distribution D on $X \times Y$ and a protocol P for a function $f : X \times Y \to Z$, Feigenbaum et al. [5] define the average-case PAR as follows:

$$\mathrm{avg\,PAR}(P) = \mathbb{E}_D\left[\frac{|R_{x,y}|}{|P_{x,y}|}\right].$$

In this paper we will also consider the following alternative definition.

Definition 11: For a probability distribution D on $X \times Y$ and a protocol P for a function $f : X \times Y \to Z$, let the **average-case objective privacy approximation ratio** of protocol P for function f be:

$$\mathrm{avg}_D\,\mathrm{PAR}(P) = \mathbb{E}_{(x,y)\in D}\left[\frac{|R_{x,y}|_D}{|P_{x,y}|_D}\right],$$

where for $S \subseteq X \times Y$, $|S|_D = \sum_{(x,y)\in S} D(x,y)$. Furthermore, we let the **average-case subjective privacy approximation ratio** of protocol P for function f be:

$$\operatorname{avg}_D \operatorname{PAR}^{sub}(P) = \max\left\{ \mathbb{E}_{(x,y)\in D}\left[\frac{|R_{x,y}\cap X\times\{y\}|_D}{|P_{x,y}\cap X\times\{y\}|_D}\right], \mathbb{E}_{(x,y)\in D}\left[\frac{|R_{x,y}\cap \{x\}\times Y|_D}{|P_{x,y}\cap \{x\}\times Y|_D}\right]\right\}.$$

As opposed to Feigenbaum et al. we measure the size of subsets of $X \times Y$ relative to the measure D. This definition coincides with the definition of Feigenbaum et al. for the uniform distribution [5]. Their paper does not give any results for distributions other than uniform, so our definition is consistent with their results. Similarly, most of our results for concrete functions are for the uniform distribution, so they hold under both definitions.

Definition 11 is motivated by an attempt to prove Theorem 2, and will be convenient and useful in that proof (see Proposition 12). Both measures have advantages and disadvantages; in various scenarios, one may be preferred to the other. However, our definition has interesting mathematical properties and (as we will see in a moment) it is related to other known measures. For further discussion of alternative definitions of average-case PAR, see section 8.1 of [5].

One benefit of Definition 11 is that one can relate average-case PAR to another natural measure on protocols. Consider a protocol P for a function f. For a region $R \in \mathcal{R}(f)$ we let $\operatorname{cut}_P(R) = |\{P_{x,y} \mid (x,y) \in R\}|$ be the number of protocol-induced rectangles contained within R. The following statement is implicit in Feigenbaum et al. [5] for the case of uniform distribution and objective PAR.

Proposition 12: For any function $f : X \times Y \to Z$, protocol P for f and any probability distribution D on $X \times Y$

$$\operatorname{avg}_D \operatorname{PAR}(P) = \sum_{R\in\mathcal{R}(f)} |R|_D \cdot \operatorname{cut}_P(R)$$

and $\operatorname{avg}_D \operatorname{PAR}^{sub}(P) =$

$$\max\left\{ \sum_{y\in Y, R\in\mathcal{R}(f)} |R\cap X\times\{y\}|_D \cdot \operatorname{cut}_P(R\cap X\times\{y\}) \sum_{x\in X, R\in\mathcal{R}(f)} |R\cap \{x\}\times Y|_D \cdot \operatorname{cut}_P(R\cap\{x\}\times Y)\right\}.$$

Proposition 12 holds by easy manipulation of the definitions.

In the setting of our definition, this characterization of average-case PAR provides a simple answer to the conjecture [5] that for any probability distribution D on inputs, there is a protocol that has average-case PAR at most n for the 2^n-Vickrey auction. Recall that the Bisection Protocol for the Vickrey auction proceeds by binary search on the input domain [5].

Proposition 13: For any probability distribution D on $[2^n]\times[2^n]$, the Bisection Protocol for the 2^n-Vickrey auction satisfies:

$$\operatorname{avg}_D \operatorname{PAR}(\text{Bisection Protocol}) \le n+1.$$

Proof: Each region R of the 2^n-Vickrey auction is covered by at most $n+1$ rectangles induced by the Bisection Protocol, i.e., $\operatorname{cut}_{\text{Bisection Protocol}}(R) \le n+1$. The claim follows by the previous proposition. ∎

For the uniform distribution we can prove the following tradeoff between the length and average-case PAR of any protocol. This is one of our main results.

Theorem 2: For all $n, r \ge 1$, any deterministic protocol of length at most r for the two-player 2^n-Vickrey auction problem has average-case PAR at least $\Omega(\frac{n}{\log(r/n)})$ (over the uniform distribution of inputs).

This bound is asymptotically tight for uniform distribution (the n/r-Bisection Protocol achieves asymptotically the same upper-bound). Our lower bound holds only for the uniform distribution on inputs. This is not surprising; if the distribution is concentrated say on a single input one should not expect large loss of privacy.

The rest of this section (up to subsection IV-A) is devoted to the proof of Theorem 2. Proposition 12 characterizes the average-case PAR as the weighted sum of $\operatorname{cut}_P(R)$ over all regions R of the function. We will use this characterization but simplify the calculation a little bit.

- We will sum only over regions $R_{x,y}$ for $x, y \le 2^{n-1}$. Call this collection of regions L. These are the largest regions in $X \times Y$, and together cover $\frac{3}{4}$ the area of $X \times Y$. Hence the loss of privacy on these regions will be significant. Each of the regions is of size between 2^{n-1} and 2^n so up-to a factor of at most 2 they all have the same weight.
- To estimate $\operatorname{cut}_P(R)$ for various regions R we will track only the set of "diagonal" inputs $Diag = \{(x,x) \mid x \in [2^{n-1}]\}$ as they progress in the protocol tree, and count protocol-induced rectangles that intersect regions $R_{x,x}$ and $R_{x,x+1}$.

Combining these two simplifications gives a lower bound on the average-case PAR for the uniform distribution:

$$\frac{2^{n-1}}{4^n} \sum_{R\in L} \operatorname{cut}_P(R). \tag{1}$$

Note that each input pair $(x,x) \in Diag$ must finish the protocol in a separate induced rectangle.

The problem of counting the cuts of interest (in order to get a lower bound) can be abstracted away into the Ball Partition Problem. By Lemma 16, a lower bound on the Ball Partition Problem will yield a lower bound on the average-case PAR for the uniform distribution on Vickrey auctions.

Definition 14 (Ball Partition Problem): For integers N and $r \geq 1$, there are N balls and r rounds. All of the balls begin in one big set. In each round, the balls in each current set are partitioned into (at most) two new sets. The cost of partitioning the balls in any set S into sets S_1 and S_2 is $\min(|S_1|, |S_2|)$. After r rounds, each of the N balls shall be in a singleton set. The total cost of the game is the sum of the cost, over all r rounds, of every partition made during each round. We denote the minimal possible cost by $B(N, r)$.

The interesting values of r lie in a particular range. For $r < \log_2 N$, the game cannot be finished at any cost. For $r > N$, the game can easily be finished with minimal cost $B(N, r) = N - 1$: cut away 1 ball from the largest set at every round. However, for intermediate values $\log N \leq r \leq N$, one might ask: what is the smallest possible cost c achievable in r rounds?

Theorem 15: For the Ball Partition Problem, $B(N, r) \geq \frac{N \log N}{4 \log(\frac{4r}{\log N})}$.

The above lower bound is asymptotically optimal. (A matching upper bound is obtained by splitting $\Theta(n/r)$-fraction of balls from each set at every round.)

Lemma 16 relates a lower bound for the Ball Partition Problem (N balls in r rounds) to a lower bound for the average-case Vickrey auction on the uniform distribution (N possible inputs for each player and r bits of communication).

Lemma 16: Let $N, r \geq 1$ be integers where N is even. Let $B(N, r)$ be the minimal cost of the Ball Partition Problem on N balls in r rounds. Then for any deterministic r-bit protocol P for 2-player N-Vickrey auction, the average-case PAR is $\operatorname{avg}\mathrm{PAR}(P) \geq \frac{B(N,r)}{2N}$ under the uniform distribution.

Proof of Lemma 16: Our goal is to establish that $\sum_{R \in L} \mathrm{cut}_P(R) \geq B(N, r)$. The lemma easily follows from this since each region R in L contains probability mass at least $1/2N$ under the uniform distribution.

The Ball Partition Problem is an abstraction of the calculation of average-case PAR for Vickrey auctions. Recall the following notation used in the proof of Theorem 1. Protocol P is associated with a protocol tree where each node v corresponds to a combinatorial rectangle $T(v) = T_A(v) \times T_B(v) \subseteq X \times Y$. For $t = 0, \ldots, r$, let $\mathcal{R}(P, t)$ be the set of rectangles associated with nodes at level t of the tree, level 0 consisting of the root. For $R \subseteq X \times Y$, let $\mathrm{cut}_P(R, t) = |\{S \in \mathcal{R}(P, t); S \cap R \neq \emptyset\}|$ be the number of rectangles intersecting R after round t of the protocol. Clearly, $\mathrm{cut}_P(R, r) = \mathrm{cut}_P(R)$. We want to estimate from below $\sum \mathrm{cut}_P(R)$ over $R \in L$.

We associate every node v of the protocol tree also with sets $D_v = [N/2] \cap T_A(v) \cap T_B(v)$ and $L_v = \{R_{x,y}; x \leq y, x, y \in D_v\}$. For each leaf node v, $|D_v| \leq 1$ as no two distinct inputs (x, x) and (x', x') can finish in the same protocol-induced rectangle of the leaf. Notice, $L_v \subseteq L$. It is easy to see by induction on the level of the tree that sets D_v associated with nodes at the same level partition $[N/2]$ and hence, sets L_v associated with nodes at the same level are disjoint. Let v be a node at level t, $0 \leq t < r$, with $D_v \neq \emptyset$. Let v_1 and v_2 be its two children. If $D_{v_1} \neq \emptyset \neq D_{v_2}$ then we claim that

$$\sum_{R \in L_v} \mathrm{cut}_P(R, t+1) \quad \geq \quad \sum_{R \in L_v} \mathrm{cut}_P(R, t) + \min(|D_{v_1}|, |D_{v_2}|) - 1.$$

We prove the claim. Assume that v is a node where Alice speaks. Hence, $T_A(v) = T_A(v_1) \dot{\bigcup} T_A(v_2)$ and $T_B(v) = T_B(v_1) = T_B(v_2)$. Clearly, $D_v = D_{v_1} \dot{\bigcup} D_{v_2}$. Let $x_1 = \max(D_{v_1})$ and $x_2 = \max(D_{v_2})$. WLOG, $x_1 < x_2$. For every $y \in D_{v_1}, y \neq x_1$, $(x_1, y) \in R_{y+1,y} \cap T(v_1)$ and also $(x_2, y) \in R_{y+1,y} \cap T(v_2)$, so both are non-empty. Hence, $\mathrm{cut}_P(R_{y+1,y}, t+1) \geq \mathrm{cut}_P(R_{y+1,y}, t) + 1$. As there are $|D_{v_1}| - 1$ such $y's$, the claim follows in this case.

If v is a node where Bob speaks, the argument is similar. Let $y_1 = \max(D_{v_1})$ and $y_2 = \max(D_{v_2})$, and assume WLOG $y_1 < y_2$. Then for every $x \in D_{v_1}$, $(x, y_1) \in R_{x,x} \cap T(v_1)$ and also $(x, y_2) \in R_{x,x} \cap T(v_2)$. Thus in this case one does not even lose the -1 additive term.

Hence, each node v for which D_v is split into two non-empty sets D_{v_1} and D_{v_2} contributes by at least $\min(|D_{v_1}|, |D_{v_2}|) - 1$ to the increase of $\sum_{R \in L} \mathrm{cut}_P(R)$ overall. There are exactly $N-1$ nodes like that as $|D_{\mathrm{root}}| = N$. These sets D_v constitute a solution to the Ball Partition Problem in r rounds, and given the cost function for the Ball Partition Problem it is immediate that the overall increase of $\sum_{R \in L} \mathrm{cut}_P(R)$ is thus at least $B(N, r) - (N-1)$ as the -1 terms add up to $N - 1$. Since $\sum_{R \in L} \mathrm{cut}_P(R, 0) = N - 1$ we get $\sum_{R \in L} \mathrm{cut}_P(R) \geq B(N, r)$. ∎

All that remains to prove the lower bound on average-case PAR for Vickrey auctions (Theorem 2) is to prove the lower bound on the Ball Partition Problem (Theorem 15).

Proof of Theorem 15: We will examine the entropy of the partitions at each round. This permits an abstraction away from a particular ball-partitioning instance, in order to obtain general properties. This will lead to a lower bound on the objective function $B(N, r)$, the cost of the Ball Partition Problem.

It will be useful to associate with the Ball Partition Problem in r rounds a full binary tree of depth r where each set obtained at round t is associated to a distinct node at level t, and remaining nodes are associated with the empty set. The association should be so that a node associated with a set S has its children associated with sets S_1 and S_2 obtained from S during the partitioning. We label each node i, by the size of the associated set N_i, and we label edges by the fraction of balls that travel "over" that edge from the parent to the child node. (See Figure 3: a node labelled N_i with

children labelled c_iN_i and $(1-c_i)N_i$ will have edges to those children labelled c_i and $1-c_i$, respectively.)

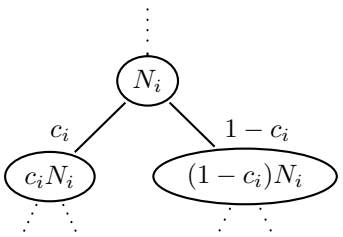

Figure 3. An arbitrary node in the ball-partitioning tree.

The tree's root node is labelled N; each leaf is labelled 1 or 0. (The 0 leaves are a result of assuming the binary tree is full; if some ball is partitioned into a singleton set in round $i < r$, then in each subsequent round it is "partitioned" into two sets: the singleton set and the empty set.)

Remark 17: At each level of the tree, the sum of the node labels $= N$. Thus the sum of labels of all the non-leaf nodes in the tree is rN.

Consider the path followed by any ball b from the root to a leaf. It traverses edges labelled d_1^b, d_2^b, …, d_r^b, where $\prod_{i=1}^{r} d_i^b = \frac{1}{N}$.

Multiplying this number for all balls gives a nice symmetrization which is true for all trees representing solutions to the Ball Partition Problem.

$$\left(\frac{1}{N}\right)^N = \prod_{b \text{ a ball}} \prod_{i=1}^{r} d_i^b \tag{2}$$

Consider some non-leaf node i of the tree, with edges to its children labelled c_i and $1-c_i$ (Figure 3). Together, these edges contribute $(c_i)^{c_iN_i}(1-c_i)^{(1-c_i)N_i}$ to the right-hand side of equation (2). (If $c_i = 0$ this term equals 1 by definition.) WLOG assume each $c_i \le 1/2$. Equation (2) can be rewritten as:

$$\begin{aligned}\left(\frac{1}{N}\right)^N &= \prod_{\text{non-leaf node } i} (c_i)^{c_iN_i}(1-c_i)^{(1-c_i)N_i} \\ -N\log N &= \sum_i N_i(-H(c_i)) \end{aligned} \tag{3}$$

Where $H(x) = x\log\frac{1}{x} + (1-x)\log\frac{1}{1-x}$ is the binary entropy of x.

Since the leaf nodes are not included in the sum, $\sum_{\text{non-leaf node } i} N_i = rN$ (by Remark 17). Let $c = \sum_i \frac{c_iN_i}{rN}$ be the average cost of a cut in the Ball Partition Problem. Then the cost of the entire tree is $B(N,r) = crN$. Since H is concave, $\sum_i \frac{N_i}{rN}H(c_i) \le H(\sum_i \frac{c_iN_i}{rN}) = H(c)$.

$$N\log N = rN\sum_i \frac{N_i}{rN}H(c_i) \le rNH(c) \tag{4}$$

For the sake of contradiction, suppose that the cost of the tree $B(N,r) = crN < \frac{N\log N}{4\log(\frac{4r}{\log N})}$. Then the average cost of a cut is $c < \frac{\log N}{4r\log(\frac{4r}{\log N})}$. This c can be rewritten as $c = \frac{x}{-\log x}$ for $x = \frac{\log N}{4r}$. Combining equation (4) and Lemma 18 (below),

$$\frac{\log N}{r} \le H(c) = H\Big(\frac{x}{-\log x}\Big) < 4x = 4\frac{\log N}{4r} = \frac{\log N}{r}$$

The inequality makes this a contradiction. Therefore every tree of depth $\le r$ must incur cost $\ge \frac{N\log N}{4\log(\frac{4r}{\log N})}$. ∎

Lemma 18: For $0 < x \le \frac{1}{2}$, the binary entropy $H\big(\frac{x}{-\log x}\big) < 4x$.

Proof: For $0 < x \le \frac{1}{2}$, $\log\frac{1}{x} \ge 1$ so clearly $0 < \big(\frac{x}{-\log x}\big) \le \frac{1}{2}$. Let $y = \frac{x}{-\log x}$.

Expanding,

$$H(y) = y\log\frac{1}{y} + (1-y)\log\frac{1}{1-y}$$

For $0 < y \le \frac{1}{2}$, it is not difficult to see that $-\log(1-y) \le 2y$ and $1-y < 1$.

$$H(y) \le y\log\frac{1}{y} + (1-y)2y < y\log\frac{1}{y} + 2y$$

Substituting for y and expanding,

$$H\left(\frac{x}{\log\frac{1}{x}}\right) < x\left(\frac{\log\log\frac{1}{x}}{\log\frac{1}{x}}\right) + x\left(\frac{\log\frac{1}{x}}{\log\frac{1}{x}}\right) + 2x\left(\frac{1}{\log\frac{1}{x}}\right)$$

Examination reveals that for $0 < x \le \frac{1}{2}$, the parenthesized coefficients are each ≤ 1. Hence $H(\frac{x}{\log\frac{1}{x}}) < 4x$. ∎

Note. As in the case of worst-case PAR, for the Vickrey auction problem it is not hard to show that subjective average case PAR and objective average case PAR are equivalent to within a factor of two, and thus our average case lower bound in this section extends to subjective average case PAR as well.

A. Mutual information

The definition of average-case PAR is closely related to previously studied concepts in communication complexity such as information content [13] and (information-theoretic) privacy [4]. The main distinction is that these concepts measure in terms of bits, and PAR does not. Next we recapitulate some of these measures, show their relationship to the average-case PAR, and use this connection to prove new lower bounds for the average-case PAR.

Among these notions, Klauck's privacy measure [4] is most closely related to average-case PAR. Let D be a probability distribution on $X \times Y$. Let $(\mathbf{X}, \mathbf{Y}) \sim D$ be the random variable obtained by sampling according to D. Recall, for a function f on $X \times Y$, its protocol P, and inputs $(x,y) \in X \times Y$, we let $\Pi_P(x,y)$ be the transcript of the protocol on input (x,y). Then $\Pi_P(\mathbf{X}, \mathbf{Y})$ is the random variable obtained by sampling a random input according to

D. Klauck [4] gives the following definition of privacy of a protocol.

$$\mathrm{PRIV}_D(P) = \max\{I(\mathbf{X} : \Pi_P(\mathbf{X},\mathbf{Y})|\mathbf{Y}, f(\mathbf{X},\mathbf{Y})), \\ I(\mathbf{Y} : \Pi_P(\mathbf{X},\mathbf{Y})|\mathbf{X}, f(\mathbf{X},\mathbf{Y}))\}.$$

The relationship between this measure and our average-case PAR is given by the following theorem.

Theorem 19: For a probability distribution D on $X \times Y$ and a protocol P for a function $f : X \times Y \to Z$, the following holds:

$$\mathrm{PRIV}_D(P) \leq \log(\mathrm{avg}_D\, \mathrm{PAR}^{sub}(P)).$$

Proof: By symmetry, it suffices to show that $I(\mathbf{X} : \Pi_P(\mathbf{X},\mathbf{Y})|\mathbf{Y}, f(\mathbf{X},\mathbf{Y})) \leq \log(\mathrm{avg}_D\, \mathrm{PAR}^{sub}(P))$.

$$\begin{aligned} & I(\mathbf{X} : \Pi_P(\mathbf{X},\mathbf{Y})|\mathbf{Y}, f(\mathbf{X},\mathbf{Y})) \\ \leq\ & H(\Pi_P(\mathbf{X},\mathbf{Y})|\mathbf{Y}, f(\mathbf{X},\mathbf{Y})) \\ \leq\ & \sum_{y\in Y, z\in Z} |R_z \cap X \times \{y\}|_D \cdot \log(\mathrm{cut}_P(R_z \cap X \times \{y\})) \\ \leq\ & \log(\mathrm{avg}_D\, \mathrm{PAR}^{sub}(P)), \end{aligned}$$

The first inequality holds by simple algebra. The second inequality holds because, for any $y \in Y$ and $z \in Z$, $\Pr[\mathbf{Y} = y, f(\mathbf{X},\mathbf{Y}) = z] = |R_z \cap X \times \{y\}|_D$ and $H(\Pi_P(\mathbf{X},\mathbf{Y})|\mathbf{Y} = y, f(\mathbf{X},\mathbf{Y}) = z) \leq \log(\mathrm{cut}_P(R_z \cap X \times \{y\}))$. The final inequality follows from concavity of logarithm. ■

Hence, one can use lower bounds on PRIV to derive lower bounds for average-case PAR. For example, consider the function $\mathrm{DISJ}_n : \{0,1\}^n \times \{0,1\}^n \to \{0,1\}$ on inputs $x, y \in \{0,1\}^n$, which is defined to be one if $\{i \in [n]; x_i = y_i = 1\}$ is empty and zero otherwise. Klauck [4] shows that for any protocol P for the disjointness problem, $\mathrm{PRIV}_D(P) \in \Omega(\sqrt{n}/\log n)$, where D is uniform on strings of hamming weight $\sqrt{n}$. Using the above lower bound, we immediately obtain $\mathrm{avg}_D\, \mathrm{PAR}^{sub}(P) \in 2^{\Omega(\sqrt{n}/\log n)}$ for any protocol P for DISJ_n.

There are two other well studied measures that are closely related to our average-case PAR: the *external* and *internal information cost* (IC^{ext} and IC, resp.). The external information cost was defined in [14] where the internal cost was also used implicitly. Later, using this measure, Bar-Yossef et al. [15] obtained $\Omega(n)$ lower bounds on the randomized communication complexity of DISJ_n. The internal information cost was formalized in [13]. For a protocol P for function $f : X \times Y \to Z$ and a distribution D on $X \times Y$, they are defined respectively as follows:

$$\mathrm{IC}^{ext}_D(P) = I(\mathbf{X},\mathbf{Y} : \Pi_P(\mathbf{X},\mathbf{Y}))$$

$$\mathrm{IC}_D(P) = I(\mathbf{X} : \Pi_P(\mathbf{X},\mathbf{Y})|\mathbf{Y}) + I(\mathbf{Y} : \Pi_P(\mathbf{X},\mathbf{Y})|\mathbf{X}).$$

As one can see the internal information cost is closely related to the privacy measure PRIV of Klauck. The only substantial difference is that PRIV is conditioned on the value of the function whereas IC is not. When f is a Boolean function, they are asymptotically identical.

Proposition 20: For any probability distribution D on $X \times Y$ and any protocol P for a function $f : X \times Y \to Z$:

$$\mathrm{PRIV}_D(P) - \log|Z| \leq \mathrm{IC}_D(P) \leq 2\cdot(\mathrm{PRIV}_D(P) + \log|Z|).$$

The proposition follows from Claim 5. This relationship together with the known lower bounds on internal information cost of DISJ_n allow us to prove one of the conjectures of Feigenbaum et al. [5] for the intersection function $\mathrm{INTERSEC}_n$. Function $\mathrm{INTERSEC}_n : \{0,1\}^n \times \{0,1\}^n \to \mathcal{P}([n])$ on inputs $x, y \in \{0,1\}^n$ gives the set $\{i \in [n]; x_i = y_i = 1\}$.

Feigenbaum et al. conjecture that the average-case subjective PAR for the intersection function under the uniform distribution is exponential in n. This can be proven using the above tools and the following result, which strengthens an earlier work by Bar-Yossef et al. [15]. Let ν be the uniform distribution supported on $\{(0,1),(1,0),(0,0)\}$. Let τ be the distribution generated by taking the n-fold product of ν. In other words, τ is the uniform distribution supported on pairs of strings that are disjoint.

Theorem 21: [16] Let P be any randomized protocol that computes disjointness DISJ_n with error probability $< 1/3$. Then, $\mathrm{IC}_\tau\left(P\right) = \Omega(n)$.

Using the above theorem, we show the following bound for Intersection.

Theorem 22: Let P be any deterministic protocol that computes set intersection $\mathrm{INTERSEC}_n$. Then, for $\mathcal{U}$ the uniform distribution, $\mathrm{PRIV}_{\mathcal{U}}\left(P\right) = \Omega(n)$.

Proof of Theorem 22: We prove this by a contradiction. Assume that we have a protocol P to solve $\mathrm{INTERSEC}_m$ on m-bit inputs with little privacy loss under the uniform distribution. The main idea of the argument is to come up with an appropriate reduction from set disjointness DISJ_n on n bits to set intersection $\mathrm{INTERSEC}_m$. This reduction will need to satisfy the following features: solving intersection on the reduced instance should solve set-disjointness on the original input instance. The reduced instance should not blow up too much in size, i.e. $m = \Theta(n)$. Finally, and most importantly, distribution τ on input instances to set-disjointness should generate by our reduction the uniform distribution on Intersection. This last step seems difficult to do via a deterministic reduction. So we aim to get a workaround as follows.

Let $\mathbf{\Pi}$ be the random variable denoting the transcript generated by P. Then, our assumption on P gives the following for some constant β which we fix at the end: $\beta\ m > I_{\mathcal{U}}\big(\mathbf{X} : \mathbf{\Pi}\,|\,\mathbf{Y}, \mathrm{INTERSEC}(\mathbf{X},\mathbf{Y})\big) + I_{\mathcal{U}}\big(\mathbf{Y} : \mathbf{\Pi}\,|\,\mathbf{X}, \mathrm{INTERSEC}(\mathbf{X},\mathbf{Y})\big)$.

The uniformly distributed pairs of m-bit random strings $(\mathbf{X}, \mathbf{Y})$ can be alternatively generated by first selecting a random subset $\mathbf{A}$ of $[m]$ where each element is in the set independently with probability $1/4$. For each $i \in \mathbf{A}$, we set $(\mathbf{X}_i, \mathbf{Y}_i) = (1, 1)$. Then, for each coordinate $i \in \mathbf{A}^c = [m] - \mathbf{A}$, $(\mathbf{X}_i, \mathbf{Y}_i)$ is picked independently according to ν. Let $(\mathbf{X}^A, \mathbf{Y}^A)$ denote pair of random variables that are distributed according to $\mathbf{X}, \mathbf{Y}$ conditioned on $\mathbf{A}$ as above and the underlying distribution on this pair be denoted by τ^A. Thus, our assumption becomes equivalently:

$$\mathbb{E}_{\mu_A}\left[I_{\tau^A}\left(\mathbf{X}^A : \mathbf{\Pi} \,|\, \mathbf{Y}^A\right) + I_{\tau^A}\left(\mathbf{Y}^A : \mathbf{\Pi} \,|\, \mathbf{X}^A\right)\right] < \beta m,$$

where μ_A is the distribution on $\mathbf{A}$. Applying the Chernoff bound on the deviation of $|\mathbf{A}|$ from its expectation, one concludes: $(\beta m)/\big(1 - \exp(-\Omega(m))\big) >$

$$\mathbb{E}_{\mu_A}\left[I_{\tau^A}\left(\mathbf{X}^A : \mathbf{\Pi} \,|\, \mathbf{Y}^A\right) + I_{\tau^A}\left(\mathbf{Y}^A : \mathbf{\Pi} \,|\, \mathbf{X}^A\right) \,\middle|\, |\mathbf{A}| \leq m/2\right]$$

Thus, there exists some fixed set a of size at most $m/2$ such that

$$I_{\tau^a}\left(X^a : \Pi \,|\, Y^a\right) + I_{\tau^a}\left(Y^a : \Pi \,|\, X^a\right) < \beta' m. \qquad (5)$$

This set a is going to provide us with the workaround needed for the deterministic reduction. We define our reduction now w.r.t a. Set $n = m - |a| \geq m/2$. Let P' be a protocol that solves set-disjointness as follows: Given two n-bit strings (u, v), protocol P' first embeds u and v naturally into $a^c = [m] - a$. Let the embedded strings be called $X(u)$ and $Y(v)$ which each player can generate privately on its own. Then, the players run the protocol P on $\big(X(u), Y(v)\big)$. Let J be the intersection set that P returns. Clearly, $\mathrm{DISJ}_n(u, v) = 1$ iff $|J| = |a|$. Finally, note if $(\mathbf{U}, \mathbf{V})$ are generated according to τ, then the mapped strings $\big(\mathbf{X}(\mathbf{U}), \mathbf{Y}(\mathbf{V})\big) \sim (\mathbf{X}^a, \mathbf{Y}^a)$. Hence, (5) implies that $\mathrm{IC}_\tau(P) \leq \beta' m \leq 2\beta' n$. By setting β' to be a small enough constant, we derive a contradiction to Theorem 21. This completes the argument. ∎

By using Theorem 19, this immediately yields the following theorem, conjectured by Feigenbaum et al. [7].

Theorem 3 (Conjectured by Feigenbaum et al.): 3 For all $n \geq 1$, and any protocol P computing the Set Intersection $\mathrm{INTERSEC}_n$ the average-case subjective PAR is exponential in n under the uniform distribution: $\mathrm{avg}_{\mathcal{U}}\, \mathrm{PAR}^{sub}(P) = 2^{\Omega(n)}$.

V. Conclusion

These techniques hold the promise of similar length-privacy tradeoffs for other functions. Further, it seems that one can readily extend this work to include randomized and ϵ-error settings. With the restriction of perfect privacy for two-player functions, [1] shows that the set of functions with deterministic protocols and the set of functions with randomized protocols are the same. Perhaps there is a similar result for any fixed constant PAR, or perhaps as the PAR requirement is relaxed, the two sets gradually differ.

References

[1] E. Kushilevitz, "Privacy and communication complexity," *30th FOCS* pp. 416–421, 1989.

[2] F. Brandt and T. Sandholm, "On the Existence of Unconditionally Privacy-Preserving Auction Protocols," *ACM Transactions on Information and System Security*, vol. 11, pp. 1–21, May 2008.

[3] B. Chor and E. Kushilevitz, "A Zero-One Law for Boolean Privacy (extended abstract)," in *21st ACM STOC*, pp. 62–72, 1989.

[4] H. Klauck, "On quantum and approximate privacy," *Proc. STACS*, 2002.

[5] J. Feigenbaum, A. D. Jaggard, and M. Schapira, "Approximate Privacy: Foundations and Quantification," *11th Conference on Electronic Commerce*, pp. 167–178, 2010.

[6] M. Comi, B. DasGupta, M. Schapira, and V. Srinivasan. "On Communication Protocols That Compute Almost Privately," Symposium on Algorithmic Game Theory, pp. 44-56, 2011.

[7] J. Feigenbaum, A. D. Jaggard, and M. Schapira, "Approximate Privacy: PARs for Set Problems," *DIMACS Technical Report 2010-01*, pp. 1–34, 2010.

[8] E. Kushilevitz and N. Nisan, *Communication Complexity*. Cambridge University Press, 1997.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley-Interscience, 1991.

[10] B. Chor, M. Geréb-Graus, and E. Kushilevitz, "On the structure of the privacy hierarchy," *Journal of Cryptology*, vol. 7, no. 1, pp. 53–60, 1994.

[11] A. C.-c. Yao, "Protocols for Secure Computations," *Proceedings of the 23rd FOCS*, pp. 160–164, 1982.

[12] A. C.-c. Yao, "Some Complexity Questions Related to Distributive Computing," *11th ACM STOC*, pp. 209–213, 1979.

[13] B. Barak, M. Braverman, X. Chen, and A. Rao, "How to compress interactive communication," *42nd ACM STOC*, 2010.

[14] A. Chakrabarti, A. Wirth, A. Yao, and Y. Shi, "Informational complexity and the direct sum problem for simultaneous message complexity," *42nd IEEE FOCS*, pp. 270–278, 2001.

[15] Z. Bar-Yossef, T. Jayram, R. Kumar, and D. Sivakumar, "An information statistics approach to data stream and communication complexity," *JCSS*, vol. 68, no. 4, pp. 702–732, 2004.

[16] M. Braverman, "Interactive information complexity," *ECCC* 18: 123, 2011.