CSCI 5520: Foundations of Data Privacy **Lecture 8**
The Chinese University of Hong Kong, Spring 2015 3 March 2015

In this lecture we show strong evidence that it is in general computationally hard to answer polynomially many (in the number of rows) counting queries in a differentially private manner.

More precisely, we will show that for every efficient non-interactive mechanism there exist a collection of counting queries whose answers can be easily computed from the database itself, but the mechanism cannot produce both differentially private and accurate answers to these queries, under the assumption that pseudorandom generators exist.

Before we state and prove the main result, we take a detour to introduce fingerprinting codes, an object that will play a main role in the construction of the "hard" queries Q.

1 Fingerprinting codes

A string $f \in \{0,1\}^m$ is a *fingerprint* of matrix $W \in \{0,1\}^{n\times m}$ if for every index $j \in [m]$ and bit $b \in \{0,1\}$, if all the entries of the j-th column of W are equal to b, then f_j is also equal to b. For example, the strings $(0,0,1)$, $(0,1,1)$ are fingerprints of the matrix

$$W = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

but the string $(1,1,1)$ is not a fingerprint of this matrix.

Given a value of n, we would like to design a matrix W such that for any submatrix W_{-i} obtained by erasing a single row of W, a fingerprint of W_{-i} identifies at least one of its rows. For $n = 2$, the above matrix W is such an example. However, even for $n = 3$ no such matrix W exists: Given any $W \in \{0,1\}^{3\times m}$, let $f \in \{0,1\}^m$ be the string that in position j contains the majority value of the 3 bits present in the j-th column of W. Then f is a fingerprint for any one of the three relevant submatrices of W. Fingreprinting codes bypass this obstacle by allowing for a probabilistic choice of the matrix W.

A $n \times m$ *fingerprinting code* is a distribution $\mathcal{D}$ over pairs (μ, W) consisting of a private key μ and an $n \times m$ matrix W and a pointing algorithm P that takes as inputs μ, W, and a fingerprint $f \in \{0,1\}^m$ and outputs an index $i \in [n]$ (it points to a row of the matrix) or the special symbol $\perp$.

Definition 1. The fingerprinting code $(\mathcal{D}, P)$ has *completeness gap c* if for every algorithm F that on input W outputs a fingerprint of W, $\Pr_{(\mu,W)\sim\mathcal{D}}[P(\mu, W, F(W)) = \perp] \leq c$.

The fingerprinting code $(\mathcal{D}, P)$ has *soundness gap s* if for every $i \in [n]$ and every algorithm A, $\Pr_{(\mu,W)\sim\mathcal{D}}[P(\mu, W, A(W_{-i})) = i] \leq s$, where W_{-i} is the matrix obtained by erasing the i-th row of W.

A soundness gap of $1/n$ is trivial to achieve if the fingerprint checker outputs a uniformly random index i. We now show a clever construction that does (much) better.

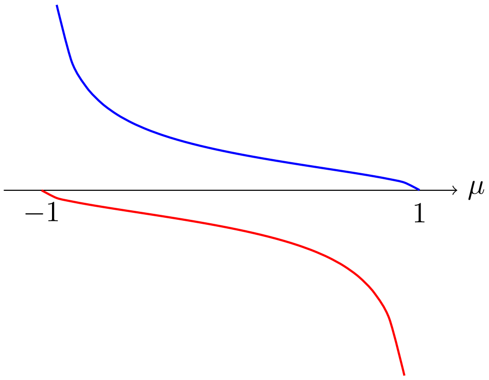
To state and analyze the construction we need a couple of concepts from probability and analysis. It will be easier to work with matrices whose entries are $1, -1$ instead of $0, 1$. For $\mu \in [-1, 1]$, a μ-biased random variable $X \sim \{-1, 1\}_\mu$ takes values $1, -1$ and has expected value μ, namely

$$X = \begin{cases} 1 & \text{with probability } (1+\mu)/2 \\ -1 & \text{with probability } (1-\mu)/2. \end{cases}$$

The *character* of X is the function $\phi_\mu \colon \{-1, 1\} \to \mathbb{R}$ given by

$$\phi_\mu(x) = \frac{x - \mathrm{E}[X]}{\sqrt{\mathrm{Var}[X]}} = \frac{x - \mu}{\sqrt{1-\mu^2}} = \begin{cases} \sqrt{(1-\mu)/(1+\mu)}, & \text{if } x = 1, \\ -\sqrt{(1+\mu)/(1-\mu)}, & \text{if } x = -1. \end{cases}$$

The random variable $\phi_\mu(X)$ has mean 0 and variance 1, i.e., $\mathrm{E}[\phi_\mu(X)] = 0$ and $\mathrm{E}[\phi_\mu(X)^2] = 1$. Here are the graphs of $\phi_\mu(1)$ (top, blue) and $\phi_\mu(-1)$ (bottom, red) as functions of μ.



Construction of fingerprinting codes Let $\mathcal{P}$ be a probability distribution over the interval $[-1, 1]$ to be specified later.

- The private key μ is a vector $(\mu_1, \dots, \mu_m) \in [-1, 1]^n$, one for each column of W, where each μ_j is a random sample from $\mathcal{P}$, independent of the others.
- The matrix W consists of independent $\{0, 1\}$ entries, where the entries W_{ij} in column j come from the distribution $\{-1, 1\}_{\mu_j}$.
- The pointing algorithm $P(\mu, W, f)$ outputs any i such that

  $$\sum_{j=1}^{m} f_j \cdot \phi_{\mu_j}(W_{ij}) \geq t\sqrt{m}$$

  if such an i exists and $\perp$ otherwise.

Analysis We now show that for $t = 10\sqrt{n}$, this fingerprinting code has length $m = O(n^3)$, completeness gap 1/4, and soundness gap $1/100n$. (In fact it is possible to improve the analysis so that the same completeness and soundness are achievable with $m = O(n^2 \text{poly} \log n)$.)

Theorem 2 (Soundness)**.** *For any $\mathcal{P}$, $(\mathcal{D}, P)$ has soundness gap at most $1/t^2$.*

Proof. Since $A(W_{-i})$ does not see the i-th row of W, the random variables W_{ij} are independent of the string $f = A(W_{-i})$. For any fixed p and $f \in \{-1, 1\}^n$, we then have

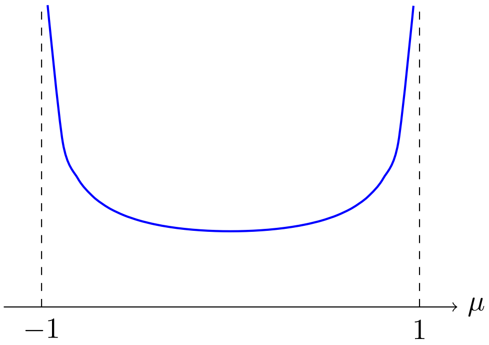
$$\mathrm{E}_{W_i} \sum_{j=1}^{m} f_j \cdot \phi_{p_j}(W_{ij}) = \sum_{j=1}^{m} f_j \cdot \mathrm{E}[\phi_{p_j}(W_{ij})] = 0$$

and because W_{ij} and $W_{ij'}$ are independent when $j \neq j'$,

$$\mathrm{E}_{W_i} \Big(\sum_{j=1}^{m} f_j \cdot \phi_{p_j}(W_{ij}) \Big)^2 = \sum_{j=1}^{m} f_j^2 \cdot \mathrm{E}[\phi_{p_j}(W_{ij})^2] = m.$$

Soundness follows by Chebyshev's inequality applied to the random variable $\sum_{j=1}^{m} f_j \cdot \phi_{p_j}(W_{ij})$. □

To prove completeness we set $\mathcal{P}$ to the following distribution: First, choose a uniformly random $\theta \sim [0, \pi]$, then output $\mu = \cos\theta$. Here is the graph of the probability density function of μ. The distribution favors values of μ that are close to -1 or 1, in which case the corresponding column of W is more likely to leave a fingerprint.



Theorem 3 (Completeness)**.** *If $m \geq n^2t^2/2$ then $(\mathcal{D}, P)$ has completeness gap at most 1/4.*

Proof. Let f_j be the j-th bit of the fingerprint $F(W_{-i})$. We will upper bound the probability that

$$Y = \sum_{i=1}^{n} \sum_{j=1}^{m} f_j \phi_{\mu_j}(W_{ij}) < nt\sqrt{m}$$

by 1/4. If this condition fails, there must exists at least one row i for which $\sum_{j=1}^{m} f_j \phi_{\mu_j}(W_{ij}) \geq t\sqrt{m}$, establishing completeness.

By linearity of expectation,

$$\mathrm{E}[Y] = \sum_{j=1}^{m} \mathrm{E}\Big[f_j \sum\nolimits_{i=1}^{n} \phi_{\mu_j}(W_{ij})\Big].$$

If we fix j and all the entries of p and W except for their j-th components, then f_j is a boolean-valued function of the bits $W_{1j}, W_{2j}, \ldots, W_{nj}$ comprising the j-th column of W. Since F is a fingerprinting algorithm, this function must evaluate to -1 when all these bits are -1 and to 1 when all these bits are 1. Lemma 4 below shows that under such restrictions, the expectation in question evaluates to 2, so the sum equals $2m$.

We now upper bound $\mathrm{Var}[Y]$. First, for any fixed j, $f_j \in \{-1, 1\}$ so,

$$\mathrm{Var}\Big[f_j \sum\nolimits_{i=1}^{n} \phi_{\mu_j}(W_{ij})\Big] \leq \mathrm{E}\Big[\Big(f_j \sum\nolimits_{i=1}^{n} \phi_{\mu_j}(W_{ij})\Big)^2\Big] = \mathrm{E}\Big[\Big(\sum\nolimits_{i=1}^{n} \phi_{\mu_j}(W_{ij})\Big)^2\Big] = n$$

because the random variables $\phi_{\mu_j}(W_{ij})$ are independent with mean 1. When $j \neq j'$, the covariance of any two random variables $f_j \sum_{i=1}^{n} \phi_{\mu_j}(W_{ij})$ and $f_{j'} \sum_{i=1}^{n} \phi_{\mu_{j'}}(W_{ij'})$ equals

$$\begin{aligned}\mathrm{E}\Big[\Big(f_j \sum\nolimits_{i=1}^{n} \phi_{\mu_j}(W_{ij})\Big)\Big(f_{j'} \sum\nolimits_{i=1}^{n} \phi_{\mu_{j'}}(W_{ij'})\Big)\Big] - \mathrm{E}\Big[f_j \sum\nolimits_{i=1}^{n} \phi_{\mu_j}(W_{ij})\Big]\Big[f_{j'} \sum\nolimits_{i=1}^{n} \phi_{\mu_{j'}}(W_{ij'})\Big] \\ = \mathrm{E}_W\left[f_j f_{j'}\, \mathrm{E}_{\mu_j}\left[\sum\nolimits_{i=1}^{n} \phi_{\mu_j}(W_{ij})\Big| W\right] \mathrm{E}_{\mu_{j'}}\left[\sum\nolimits_{i=1}^{n} \phi_{\mu_{j'}}(W_{ij'})\Big| W\right]\right] - 4.\end{aligned}$$

Among all functions f_j, f_j' that map W to -1 or 1, the outer expectation is maximized by choosing f_j to equal the sign of $\mathrm{E}[\sum_{i=1}^{n} \phi_{\mu_j}(W_{ij}) \mid W]$ and $f_{j'}$ to equal the sign of $\mathrm{E}[\sum_{i=1}^{n} \phi_{\mu_j}(W_{ij'}) \mid W]$. Then f_j and $f_{j'}$ are independent random variables and the desired covariance equals zero. So all covariances are zero or negative.

It follows that $\mathrm{Var}[Y] \leq m \cdot n$. By Chebyshev's inequality,

$$\Pr[Y < nt\sqrt{m}] \leq \Pr[Y - 2m < 2\sqrt{mn}] \leq \Pr[Y - \mathrm{E}[Y] < 2\sqrt{\mathrm{Var}[Y]}] < \frac{1}{4}$$

as desired. □

Lemma 4. *For any function* $f\colon \{-1, 1\}^n \to \{-1, 1\}$ *such that* $f(-1^n) = -1$ *and* $f(1^n) = 1$,

$$\mathrm{E}_{\mu\sim\mathcal{P}, X\sim\{-1,1\}^n_p}\Big[\Big(\sum\nolimits_{i=1}^{n} \phi_\mu(X_i)\Big) f(X_1, \ldots, X_n)\Big] = 2.$$

The proof of this Lemma uses the *Margulis-Russo formula* which says that for any function $f\colon \{-1, 1\}^n \to \mathbb{R}$, any μ and $X \sim \{-1, 1\}^n_\mu$

$$\frac{d}{d\mu}\, \mathrm{E}[f(X_1, \ldots, X_n)] = \frac{1}{\sqrt{1 - \mu^2}}\, \mathrm{E}\Big[\Big(\sum\nolimits_{i=1}^{n} \phi_\mu(X_i)\Big) f(X_1, \ldots, X_n)\Big].$$

A proof of this formula using a bit of Fourier analysis is given in the appendix.

Proof of Lemma 4. Let $g(\mu) = \frac{d}{d\mu}\mathrm{E}[f(X_1, \dots, X_n)]$ where $X_1, \dots, X_n$ are independent $\{-1, 1\}_p$ random variables. By the fundamental theorem of calculus,

$$\int_{-1}^{1} g(\mu)d\mu = f(1^n) - f(-1^n) = 1 - (-1) = 2.$$

By the Margulis-Russo formula the expression of interest equals

$$\mathrm{E}_{\mu\sim\mathcal{P}}\big[\sqrt{1-\mu^2}g(\mu)\big] = \int_{0}^{\pi} (\sin\theta)g(\cos\theta)d\theta = -\int_{1}^{-1} g(\mu)d\mu = 2. \qquad \square$$

2 Private data release requires short rows

As a warmup towards the main result, we prove a *statistical* limitation of private data release: Private and accurate data release is impossible for databases with long rows.

Theorem 5. *If a (n, m)-fingerprinting code with completeness gap $1/4$ and soundness gap $1/6n$ exists, then no mechanism for n-row databases over a domain of size 2^m for m counting queries is better than $n/2$-accurate and $(1, 0.1)$-differentially private.*

We will consider databases with n rows over domain $D = \{0, 1\}^m$; we can then view the database as a matrix $K \in \{0, 1\}^{n\times m}$.

To each matrix $Z \in \{0, 1\}^{n\times m}$ we associate m counting queries $Q_Z = (q_1, \dots, q_m)$, where the j-th query q_j counts the number of entries in which the j-th column of W and the j-th column of K differ, i.e.

$$q_j(X) = |\{i\colon K_{ij} \neq Z_{ij}\}|.$$

Suppose M is a $(1, 1/4n)$-differentially private mechanism for counting queries. We instantiate M on database X and queries $Q_{K\oplus W}$, where $K \sim \{0, 1\}^{n\times m}$ is a uniformly random matrix, $W \sim \mathcal{D}$ is a sample from the fingerprinting code $(\mathcal{D}, F)$, and $K \oplus W$ is the bitwise XOR of the entries of K and W.

The true answer to q_j is the number of rows i such that $K_{ij} \neq K_{ij} \oplus W_{ij}$, that is the number of one entries of the j-th column of W. This number is n for a column of ones and 0 for a column of zeros.

Let *Round* be an algorithm that "rounds" the corresponding answers of M: It takes as its input answers $a_1, \dots, a_m$ to $q_1, \dots, q_m$ and outputs a vector in $\{0, 1\}^m$ that has 1 in position j if $a_j > n/2$ and 0 if $a_j < n/2$. If M has accuracy better than $n/2$, then $\mathit{Round}(M(K, Q_{K\oplus W}))$ must be a fingerprint of W. By the completeness of the fingerprinting code,

$$\mathrm{Pr}_{(\mu,W)\sim\mathcal{D};K\sim\{0,1\}^{n\times m}}[P(\mu, W, \mathit{Round}(M(K, Q_{K\oplus W}))) \neq \bot] \geq \frac{3}{4}.$$

If i^* be a uniformly random chosen index from the set $[n]$, then

$$\mathrm{Pr}_{(\mu,W)\sim\mathcal{D};K}[P(\mu, W, \mathit{Round}(M(K, Q_{K\oplus W}))) = i^*] \geq \frac{3}{4n}.$$

Now let K^* be the database obtained by replacing the i^*-th row of K by a uniformly random row. By the differential privacy of M

$$\frac{3}{4n} \leq e \cdot \Pr_{(\mu,W)\sim\mathcal{D};K,K^*}[P(\mu, W, \mathit{Round}(M(K^*, Q_{K\oplus W}))) = i^*] + \frac{1}{4n}$$

from where

$$\Pr_{(\mu,W)\sim\mathcal{D};K,K^*}[P(\mu, W, \mathit{Round}(M(K^*, Q_{K\oplus W}))) = i^*] \geq \frac{1}{2en}. \tag{1}$$

The i^*-th row of the matrix $K \oplus W$ is now statistically independent of the database K^*, so if W_{-i^*} is the matrix obtained by zeroing out the i^*-th row of W, then

$$(K^*, K \oplus W) \text{ and } (K^*, K \oplus W_{-i^*}) \text{ are identically distributed.}$$

Let A be an adversary that on input a matrix of the form W_{-i^*} generates a random pair K, K^* that differ in row i^* and outputs $\mathit{Round}(M(K^*, Q_{K\oplus W_{-i^*}}))$. Then

$$\Pr_{(\mu,W)\sim\mathcal{D};A}[P(\mu, W, A(W_{-i^*})) = i^*] \geq \frac{1}{2en}$$

violating the $1/6n$-soundness gap of the fingerprinting code $(\mathcal{D}, P)$. Therefore M could not have been $(1, 1/4n)$-differentially private.

3 Counting queries are hard to answer privately

To reduce the size of the domain, we replace the random rows of the matrix K by pseudorandom strings. This should not affect the analysis as long as all algorithms are efficient, so they do not distinguish random strings from pseudorandom ones. However, pseudorandom strings have a much shorter description, so much less information will need to be encoded in the database.

To state the construction formally and explain the proof we need to make use of pseudorandom generators. The definition here is a bit informal.

Definition 6. An efficient deterministic algorithm $G\colon \{0,1\}^k \to \{0,1\}^m$, where $m > k$, is an *ε-pseudorandom generator* if for every efficient decision procedure D,

$$\left|\Pr_{K\sim\{0,1\}^k}[D(G(X)) \text{ accepts}] - \Pr_{Y\sim\{0,1\}^m}[D(Y) \text{ accepts}]\right| \leq \varepsilon.$$

In words, no efficient adversary D can distinguish outputs of $G(X)$ — which are statistically far from uniformly random, as they contain only $n < m$ bits of information — from uniformly random strings of the same length.

We will model an efficient mechanism as an efficient randomized algorithm that takes as input a database x and a sequence of queries $Q = (q_1, \dots, q_m)$ and outputs a sequence of answers $(a_1, \dots, a_m)$. The mechanism M is *computationally (ε, δ)-differentially private* if for every efficient, randomized, decision procedure T and every pair of adjacent databases x and x',

$$\Pr[T(M(x)) \text{ accepts}] \leq e^{\varepsilon} \Pr[T(M(x)) \text{ accepts}] + \delta.$$

Theorem 7. *Assume there exist a $1/20n$-pseudorandom generator $G\colon \{0,1\}^k \to \{0,1\}^m$ and a (n,m)-fingerprinting code with completeness gap $1/4$ and soundness gap $1/20n$. Then no computationally efficient mechanism for n-row databases over a domain of size 2^k for m counting queries is better than $n/2$-accurate and computationally $(1,0.1)$-differentially private.*

The difference from Theorem 5 is that the database size is now 2^k, which can be much less than 2^m. If our notion of efficiency is "polynomial time" and m is polynomial in n then it is believed that pseudorandom generators exist for $k = \omega(\log n)$ and n sufficiently large.

The proof of Theorem 7 is very similar to the proof of Theorem 5. We explain the differences only. As before, the database K will be chosen at random from $\{0,1\}^{n\times k}$, the matrix W will be chosen at random from $\mathcal{D}$, but now the j-th query will ask for the number of rows i such that $G(K_i)_j \neq Z_{ij}$. The matrix Z is instantiated by $G(K) \oplus W$, where $G(K)$ is the matrix with rows $G(K_1), \ldots, G(K_n)$.

Apart from a change in notation, the reasoning is now exactly the same up to equation (1), which now gives

$$\Pr_{(\mu,W)\sim\mathcal{D};K,K^*}[P(\mu, W, \mathit{Round}(M(K^*, Q_{G(K)\oplus W}))) = i^*] \geq \frac{1}{2en}.$$

However, it is no longer true that $(K^*, G(K)\oplus W)$ and $(K^*, G(K)\oplus W_{-i^*})$ are identically distributed because $G(K_{i^*})$ is not uniformly random. By the pseudorandomness of G, and a bit of manipulation, it still holds that for any efficent D,

$$\big|\Pr[D(K^*, G(K) \oplus W) \text{ accepts}] - \Pr[D(K^*, G(K) \oplus W_{-i^*}) \text{ accepts}]\big| \leq \frac{1}{10n}.$$

Since the condition $P(\mu, W, \mathit{Round}(M(\star, Q_\star))) = i^*$ is efficiently checkable, we can conclude that

$$\Pr_{(\mu,W)\sim\mathcal{D};K,K^*}[P(\mu, W, \mathit{Round}(M(K^*, Q_{G(K)\oplus W_{-i^*}}))) = i^*] \geq \frac{1}{2en} - \frac{1}{10n} \geq \frac{1}{20n}$$

which is the same as

$$\Pr_{(\mu,W)\sim\mathcal{D};A}[P(\mu, W, A(W_{-i^*})) = i^*] \geq \frac{1}{20n}$$

violating the soundness of the fingerprinting code.

A Proof of the Margulis-Russo formula

We first prove the formula for the case $n = 1$ and $f(x) = x$. Then

$$\frac{d}{dp}\,\mathrm{E}[X] = \frac{d}{d\mu}\mu = 1$$

and

$$\mathrm{E}[\phi_\mu(X)\cdot X] = \mathrm{E}\left[\frac{X - \mathrm{E}[X]}{\sqrt{\mathrm{Var}[X]}}\cdot X\right] = \frac{\mathrm{E}[X^2] - \mathrm{E}[X]^2}{\sqrt{\mathrm{Var}[X]}} = \sqrt{\mathrm{Var}[X]} = \sqrt{1-\mu^2}$$

so the formula holds in this case.

Next, we consider the case $f(x_1, \ldots, x_n) = \prod_{i \in S} x_i$ for some $S \subseteq [n]$. Then

$$\begin{aligned}
\frac{d}{d\mu} \mathrm{E}\Big[\textstyle\prod_{i \in S} X_i\Big] &= \frac{d}{d\mu} \prod_{i \in S} \mathrm{E}[X_i] \\
&= \sum_{i \in S} \Big(\textstyle\prod_{j \in S - \{i\}} \mathrm{E}[X_j]\Big) \frac{d}{d\mu} \mathrm{E}[X_i] \\
&= \sum_{i \in S} \Big(\textstyle\prod_{j \in S - \{i\}} \mathrm{E}[X_j]\Big) \frac{1}{\sqrt{1 - \mu^2}} \mathrm{E}[\phi_\mu(X_i) \cdot X_i] \\
&= \frac{1}{\sqrt{1 - \mu^2}} \sum_{i \in S} \mathrm{E}\Big[\phi_\mu(X_i) \textstyle\prod_{j \in S} X_j\Big] \\
&= \frac{1}{\sqrt{1 - \mu^2}} \mathrm{E}\Big[\Big(\textstyle\sum_{i=1}^n \phi_\mu(X_i)\Big) \textstyle\prod_{j \in S} X_j\Big].
\end{aligned}$$

Finally, by Fourier expansion any function $f\colon \{0,1\}^n \to \mathbb{R}$ can be written as a linear combination of the functions χ_S as S ranges over all subsets of $[n]$, so the Margulis-Russo formula for general functions follows by linearity of the derivative $d/d\mu$ and expectation $\mathrm{E}[\cdot]$.

References

The construction of fingerprinting codes is from the work *Optimal probabilistic fingerprint codes* by Gábor Tardos, but his analysis is somewhat different. (His paper does not explain the miraculous choice of the distribution $\mathcal{P}$.) The presentation here is essentially Fourier-analytic; for more on the characters ϕ_μ and the Margulis-Russo formula see Chapter 8 of Ryan O'Donnell's book *Analysis of Boolean Functions.*

The connection between fingerprinting codes and differential privacy is from the works *On the complexity of differentially private data release* by Dwork, Naor, Reingold, Rothblum, and Vadhan and *Answering* $n^{2+o(1)}$ *counting queries with differential privacy is hard* by Jonathan Ullman. (I am not sure if Theorem 5 has appeared before.)

For more on pseudorandom generators see my lecture notes on cryptography or the book *Introduction to Modern Cryptography* by Jonathan Katz and Yehuda Lindell.