CODING STRINGS BY PAIRS OF STRINGS*

F. R. K. CHUNG†, R. E. TARJAN‡, W. J. PAUL† AND R. REISCHUK§

Abstract. Let $X, Y \subseteq \{0, 1\}^n$. We say $Y$ codes $X$ if every $x \in X$ can be obtained by applying a short program to some $y \in Y$. We are interested in sets $Y$ that code $X$ robustly in the sense that even if we delete an arbitrary subset $Y' \subseteq Y$ of size $k$, say, the remaining set of strings $Y \setminus Y'$ still codes $X$. In general, this can be achieved only by making in some sense more than $k$ copies of each $x \in X$ and distributing these copies on different strings $Y$. Thus if the strings in $X$ and $Y$ have the same length, then $k \leq (k+1) \cdot n$. If we allow coding of $X$ by $Y$ in a way that every $x \in X$ is obtained from strings $x, z \in Y$ by application of a short program, then we can do better. Let $Y = \{x \in X : \sum_{i=1}^n x_i \in Y\}$ where $\oplus$ denotes bitwise sum mod 2. Then $\# Y = 2^n$. Yet $Y$ codes $X$ robustly for $k = 2^{n-k}-1$. This paper explores the limitations of coding schemes of this nature.

1. Robust coding of strings by strings. For strings $x, y \in \{0, 1\}^*$, we denote by $K(x|y)$ the Kolmogorov complexity of $x$ given $y$ [P], [ZL]. We say $y$ codes $x$ if $K(x|y) = \Omega(\log |x|)$. We deliberately leave the implicit constant in the $\Omega$-notation undefined. Let $X, Y \subseteq \{0, 1\}^n$. We say $Y$ codes $X$ if for all $x \in X$ there is $y \in Y$ such that $y$ codes $x$. We say $Y$ codes $X$ $k$-robustly if for all $Y' \subseteq Y$ with $\# Y' \leq k$ the set of strings $Y \setminus Y'$ still codes $X$.

Assume that the strings $x \in X$ are of the same length and sufficiently irregular, that the strings in $Y$ are longer than the strings in $X$ by a factor $\alpha$, and that there are $\beta$ times more strings in $Y$ than in $X$. Then one would intuitively expect every $y \in Y$ to code at most $\alpha$ strings $x \in X$, and most strings $x \in X$ are coded by at most $\alpha \beta$ strings $y \in Y$. This is more or less confirmed by Lemma 1.

Lemma 1. Let $p = \alpha \log \frac{n}{p}$. Let $X = \{x_1, \ldots, x_n\} \subseteq \{0, 1\}^p$, $Y = \{y_1, \ldots, y_{jn}\} \subseteq \{0, 1\}^n$ and $K(x_1 \ldots x_n) \geq np$ (i.e. $x_1 \ldots x_n$ is a random string). Then

(a) Each of $y \in Y$ codes at most $\alpha$ strings $x \in X$.

(b) Each of at least $n/2$ strings $x \in X$ is coded by at most $2\alpha \beta$ strings $y \in Y$.

Proof. Let $\{i_1, \ldots, i_t\} \subseteq \{1, \ldots, n\}$. Then

(1) $s p - O(s \log n) \leq K(x_{i_1} \ldots x_{i_t})$ because $x_{i_1} \ldots x_{i_t}$ is random [P, fact 5].

Suppose $y \in Y$ codes $x_{i_1} \ldots x_{i_t}$. Then

(2) $K(x_{i_1} \ldots x_{i_t}) \geq \sum K(x_{i_j}|y) + O(\log K(x_{i_j}|y)) + K(y) \geq O(s \log p) + \alpha p$.

For $s = \alpha + 1$, (1) and (2) imply $p - O(\alpha \log \frac{n}{p}) \leq 0$. This proves (a).

Suppose (b) is false. Then

$\alpha \beta n \geq \sum_{j} \# (x_j y_j \text{codes} x_j) \quad \text{by (a)}$

$= \sum_{i} \# (y_i \text{codes} x_i)$

$>(\log n/2)2\alpha \beta = \alpha \beta n \quad \text{by assumption}.

\square$

Clearly, it makes sense to say that, for every $x \in X$, certain strings $y \in Y$ carry specific information about $x$—namely those strings $y$ that code $x$. By Lemma 1, if the strings in $X$ are messy, then every string $y$ carries specific information about a small number of strings in $X$. Moreover, if one deletes from $Y$ all strings carrying specific

---

* Received by the editors September 22, 1983 and in final form April 4, 1984.
† AT&T Bell Laboratories, Murray Hill, New Jersey 07974.
‡ Part of this research was done while the second author was visiting the University of Bielefeld.
§ IBM Research Laboratory San Jose, California 95193.
information about a particular string \( x \in X \), then the resulting set of strings does not 1-code \( \{x\} \) any more. Thus we have:

**Corollary 1.** If under the hypotheses of Lemma 1, \( Y \) 1-codes \( X \) \( k \)-robustly, then \( 2 \alpha \beta > k \).

2. **Simple coding of strings by pairs of strings.** For \( y, z \in \{0, 1\}^p \), let \( y \oplus z \in \{0, 1\}^p \) be the string whose \( i \)-th bit is the mod 2 sum of the \( i \)-th bits of \( y \) and \( z \) for \( 1 \leq i \leq p \). For \( 1 \leq i \leq p \), let \( e_i \in \{0, 1\}^p \) be the string which has 1 in the \( i \)-th position and 0’s in all other positions. Let \( E = \{e_1, \cdots, e_p\} \). Let \( 0 \in \{0, 1\}^p \) be the string consisting of \( p \) 0’s.

Let \( X, Y \subseteq \{0, 1\}^p \). We say \( Y \) simply 2-codes \( X \) if for all \( x \in X \) there are two strings \( y, z \in Y \) such that \( x = y \oplus z \). We say \( Y \) simply 2-codes \( X \) \( k \)-robustly if for all \( Y' \subseteq Y \) with \( \# Y' \leq k \), the set of strings \( Y' \setminus Y \) simply 2-codes \( X \).

**Example 1.** \( X = E_p \), \( Y = \{y_1, \cdots, y_{p+1}\} \), with \( y_i = e_i \) for \( i \leq p \) and \( Y_{p+1} = 0 \).

**Intuition** suggests that in this example for \( i \leq p \), the string \( y_i \) carries specific information about \( e_i \) and about no other strings in \( X \).

**Example 2.** \( X = E_p \), \( Y = \{y_1, \cdots, y_{p+1}\} \), with \( y_i = \bigoplus_{x \in Y} x_i \) for \( i \leq p \) and \( y_{p+1} = \bigoplus_{x \in X} x \).

Is there still a reasonable way to attribute to every string \( y \in Y \) specific information about a small number of strings \( x \in X \)? Motivated by this question, we consider for arbitrary \( X, Y \subseteq \{0, 1\}^p \) the following edge-labelled graph \( G(X, Y) = (V, E, L) \). \( V = Y \) is the vertex set. For all \( x, z \in Y \), there is an edge \( \{x, z\} \in E \) iff \( y \oplus z = x \) for some \( x \in X \).

\( L : E \to X \) is a mapping that labels every edge \( e = \{y, z\} \) with \( L(e) = y \oplus z \). For \( X, Y \), as in Examples 1 and 2, we get the graph of Fig. 1.

![Fig. 1](image)

Transform the edge labelling \( L : E \to X \) into a node labelling by the following rule:

\((*)\) For every edge \( e = \{y, z\} \), put label \( L(e) \) on node \( y \) or on node \( z \).

There are many ways to do this, and in general, nodes may get more than one label. Thus the resulting node labelling is a mapping from \( Y \) to the power set of \( X \). We will use the letter \( L \) both for edge and node labellings.

If an edge labelling \( L \) has been transformed by Rule \( (*) \) into a node labelling \( L' \), then for every \( x \in X \), the set of strings \( Y' = \{y \mid x \in L(y)\} \) has the property of a set of strings each of which carries specific information about \( x \). \( Y' \notin Y \) does not simply 2-code \( x \). In analogy with the case of 1-coding, we want each \( y \in Y \) to carry specific information about only a small number of strings \( x \in X \). Thus we are interested in node labellings \( L \) that minimize

\[
\max_{y \in Y} L(y).
\]

For edge-labelled graphs, \( G = (V, E, L) \), let

\[
l(G) = \min_{L'} \max_{v \in V} L'(v),
\]

where the minimum is taken over all node labellings \( L' \) that can be obtained from \( L \) by Rule \( (*) \). For the graph \( G \) of Fig. 1, we have \( l(G) = 1 \), which is obtained by the node labelling \( L'(y_i) = \{x_i\} \) for \( i \leq p \) and \( L'(y_{p+1}) = \emptyset \).
3. Transformation of labellings for simple 2-coding. We define the labelled p-dimensional cube $C_p := G(E_p, \{0, 1\}^p)$. If $Y \subseteq \{0, 1\}^n$, then $G(E_p, Y)$ is a subgraph of $C_p$.

Any node labelling of $C_p$ must have at least $2^{p-1}$ occurrences of labels among $2^p$ nodes. As for every node $v$ in $C_p$, different edges incident with $v$ have different labels, we find $l(C_p) \geq p/2$. This shows that the function $l(\cdot)$ is unbounded. As pointed out in the abstract, coding $X = \{x_1, \ldots, x_n\}$ by $Y = \{Y_{i,j} = 1 \mid i \in [1, \ldots, n]\}$ works for arbitrary $X \subseteq \{0, 1\}^p$. Thus in the case of simple 2-coding, Lemma 1 and Corollary 1 do not hold, and one has better robust coding schemes than in the case of 1-coding.

However, we have

**Lemma 2.** For all $p$ and $m$, if $y \subseteq \{0, 1\}^p$ and $y \leq m$, then $l(G(E_p, Y)) \leq \log m$.

**Proof.** The proof is by induction on $p$. For $p = 1$, this is easily verified. Suppose the lemma holds for $p$. Let $Y \subseteq \{0, 1\}^{p+1}$. For $i = 0, 1$, let $Y_i = \{y \in Y \mid y_{p+1} = i\}$ and $m_i = \# Y_i$. Then $l(G(E_{p+1}, Y_i)) \leq \log m$, for $i = 0, 1$ by the induction hypothesis. Assume $m_0 \leq m_1$. For any edge $(y, z)$ with $y \in Y_0$, $z \in Y_1$, put the edge label $e_{p+1}$ of edge $(y, z)$ on $y$. This gives

$$l(G(E_{p+1}, Y)) \leq \max\{1 + \log m, 1 + \log m, \log m\}.$$  

**Corollary 2.** Let $Y \subseteq \{0, 1\}^p$, $Y = m$. For at least $p/2$ strings $e_i \in E_m$, there is a set $Y_i \subseteq Y$ such that $\# Y_i \leq (2m \log m)/p$ and $Y \setminus Y_i$ does not simply 2-code $e_i$.

**Proof.** Assume the corollary is false. Let $L$ be the node labelling of $G(E_p, Y)$ constructed in the proof of Lemma 2. Then

$$m \log m \geq \sum_{y \in Y} \# L(y) = \sum_i \# \{y \mid e_i \in L(y)\} > (p/2)(2m \log m)/p.$$

**Corollary 3.** Let $Y \subseteq \{0, 1\}^p$, $Y = m$, and let $Y$ simply 2-code $E_p$ k-robustly. Then $(2m \log m)/p > k$.

4. General 2-coding and the associated graphs. Let $x, y, z \subseteq \{0, 1\}^n$. We say $y$ and $z$ 2-code $x$ if $K(x; yz) = (\log |x|)$. Let $X, Y \subseteq \{0, 1\}^n$. We say $Y$ 2-codes $X$ if for all $x \in X$, there are $y, z \in Y$ such that $y$ and $z$ 2-code $x$. We say $Y$ 2-codes $X$ k-robustly if for all $Y' \subseteq Y$ with $\# Y' \leq k$, the set of strings $Y \setminus Y'$ 2-codes $X$.

With $X, Y \subseteq \{0, 1\}^n$, we associate again an edge-labelled graph $G(X, Y) = (Y, E, L)$: for every $x \in X$, $y \in Y$, there is an edge $(x, y)$ if $y$ and $z$ 2-code some $x \in X$.

For each edge $e = (x, y) \in E$, we set $L(e) = \{x \in X \mid y \text{ 2-code } x\}$. Thus $L$ is now a mapping from $E$ into the power set of $X$. For $E' \subseteq E$, let

$$L(E') = \bigcup_{e \in E'} L(e).$$

The following lemma exhibits a graph theoretic property of the graphs $G(x, y)$ and their subgraphs.

**Lemma 3.** Let $X = \{x_1, \ldots, x_n\} \subseteq \{0, 1\}^n$, $Y = \{Y_1, \ldots, Y_m\} \subseteq \{0, 1\}^n$ and $G(X, Y) = (Y, E, L)$. Let $K(x_1, \ldots, x_n) \leq np$. Then

$$\# L(E) \leq \# Y \frac{a}{1 - O(p \log np)/p}.$$

**Proof.** Let $d = \# L(E)$ and let $L(E) = \{x_1, \ldots, x_d\}$. Then

$$dp - O(d \log n) \leq K(x_1, \ldots, x_d).$$
The string \( x_1 \cdots x_n \) can be specified in the following way:
- The binary representations of \( n \) and \( b \).
- For each \( j \in \{1, \cdots, d\} \) the binary representation of two indices \( k \) and \( l \) such that \( K(x_j | y_k, y_l) = O(\log p) \) and a program that produces \( x_j \) from \( y_k, y_l \).
- The bits of \( y_1 \cdots y_{\text{lp}} \).

Thus

(4) \( K(x_1 \cdots x_n) \leq O(d \log bn) + O(d \log p) + abnp \).

(3) and (4) imply the lemma. \( \Box \)

Two cases are particularly simple:
(i) \( O(\log bnp)/p < c < 1 \) for some fixed \( c \). Then \( \# L(E) = O(\# Y) \).
(ii) \( a = 1 \) and \( \# Y / (1 - O(\log p + Y)/p) < \# Y + 1 \). Then \( \# L(E) \leq \# Y \).

We now give an example of an edge-labelled graph \( G \) such that \( \# L(E) \leq \# Y \) holds for all subgraphs \( (Y, E, L) \) of \( G \), yet \( G \neq G(X, Y) \) for any \( X, Y \), to which case (ii) applies (if \( p \) is large enough).

Let \( G_i \) be a single edge with label \( x_i \). For \( i \equiv 1 \), let \( G_i^1, G_i^2 \) be two copies of \( G_i \). Connect every vertex of \( G_i^1 \) with every vertex of \( G_i^2 \) with an edge labelled \( x_i \). Call the resulting graph \( G_{i+1} \). By induction on \( i \), one easily verifies that \( \# L(E) \leq \# Y - 1 \) for any subgraph \( (V, E, L) \) of \( G \).

Suppose \( G_b \) is a subgraph of \( G(X, Y) \). Consider any node \( y \) in \( G_b \). Then \( K(x_i | y) > 2p/3 - O(\log p) \) for some \( i \in \{5, \cdots, 8\} \). Otherwise one gets the contradiction

\[
4p - O(\log p) \leq K(x_1 \cdots x_b) \leq \sum_{i=5}^8 (K(x_i | y) + O(\log p)) + K(y) \\
\leq 11p + O(\log p).
\]

Consider in \( G_b \) the subgraph drawn in Fig. 2. For all \( j \in \{1, \cdots, 5\} \), we have

\[
K(z_j | yx_i) \leq K(yx_i, z_j) - K(yx_i) + O(\log p) \\
\leq K(yz_j) + K(x_j | yz_j) - K(yx_i) + O(\log p) \\
\leq K(y) + K(z_j | y) - K(z_j | z_i) - K(x_i | y) + O(\log p) \\
\leq p - 2p/3 + O(\log p).
\]

This gives the contradiction

\[
4p - O(\log p) \leq K(x_1 \cdots x_n) \\
\leq K(yx_i) + \sum_{j=1}^5 K(z_j | yx_i) + \sum_{j=1}^4 K(x_i | z_j x_i) + O(\log p) \\
\leq \left(2 + \frac{5}{3}\right)p + O(\log p).
\]

Thus (f1), \( 1 \leq f(n) \).

For \( 16 \leq m \leq 2/3 \),

\[
\text{Since } m \geq 16, \\
[3m/2] \leq 3m/2 \leq 2/3.
\]

Hence

\[
(2)^m \geq 2/3.
\]

This proves

\[
(2)
\]

[1] C. Hylten-Cotter


MIT Press.
5. Transforming edge labellings into node labellings. For sets $V, V'$, let $V \otimes V' = \{(v, v') | v \in V, v' \in V'\}$.

**Theorem 1.** Let $G = (V, E, L)$ be an edge-labelled graph, let $\# V = n$ and for all $V' \subseteq V$, let $\# L((V \otimes V') \cap E) \leq \# V'$. Then $l(G) \leq \alpha \sqrt{n}$ where $\alpha = 2\sqrt{6}$.

**Proof.** The proof is by induction on $n$. The theorem is true for $n \leq 2$. Let $n > 2$. Find a node $u \in V$ such that $\# L((u \otimes V) \cap E) \geq \alpha \sqrt{n}$ (if no such node exists, the nodes of $G$ can be trivially labelled in the desired way). Let $E_u$ be a smallest set of edges adjacent to $u$ such that $\# L(E_u) \leq \alpha \sqrt{n}$. By hypothesis we have $\alpha \sqrt{n} - 1 \leq \# E_u \leq \alpha \sqrt{n}$. Let $V_u$ be the set of end points of edges in $E_u$ other than $u$.

Let $V_2 = V \setminus (V_1 \cup \{u\})$. Let $E_2 = (V_1 \otimes V_2) \cap E$ and $E_3 = (u \otimes V_2) \cap E$ (see Fig. 3).

Ignoring labels on edges in $E_1$ and $E_2$, we can label the nodes in $V_1$ with

$$\alpha \sqrt{n} - \alpha \sqrt{n} - 2$$

labels per node. By hypothesis, every edge in $E$ has at most 2 labels. Thus putting labels on edges in $E_1$ on the endpoints of these edges in $V_1$ gives at most 2 extra labels per node in $V_1$.

![Fig. 3](image_url)

Ignoring labels on edges in $E_2$ and $E_3$, we can label the nodes of $V_2$ with

$$\alpha \sqrt{n} - \alpha \sqrt{n} - 1 \leq \alpha \left(\sqrt{n} - \frac{\alpha \sqrt{n} - 1}{2\sqrt{n}}\right)$$

$$\leq \alpha \sqrt{n} - \frac{\alpha^2}{2} + 1 \leq \alpha \sqrt{n} - 11$$

labels per node. Putting labels on edges in $E_3$ to the endpoints of these edges in $V_2$ gives at most 2 extra labels per node in $V_2$.

Now for every label $x$ on an edge $e$ in $V_1 \otimes V_2$ that has already been put by the above operations on the endpoints of $e$ in $V_1$, delete label $x$ from edge $e$. We continue to use the letter $L$ for the modified edge labelling.

The theorem follows if we establish

**Lemma 4.** For every node $w \in V_2$, we have

$\# L((w \otimes V_1) \cap E) \leq 9$.

**Proof.** Assume the lemma is false for node $w$. Let $V_6 \subseteq V_2$ be a smallest set of nodes such that $\# L((w \otimes V_3) \cap E) \leq 10$ (Fig. 4). We make three observations:

(i) $\# V_5 \geq 9$.

(ii) Let $V_4 \subseteq V_5$ and $x \in V_5 \setminus V_4$. Then

$L((z, u)) \setminus L(V_4 \otimes u) \neq \emptyset$,

$L((z, w)) \setminus L(V_4 \otimes w) \neq \emptyset$

by the minimality of $V_1$ and $V_5$. 

\[ \text{\hspace{10cm}} \]
(iii) Let $V_4 \subset V_3$, $\# V_4 \leq 2$. Then $\# L(u \otimes V_4) \leq 3$ and $\# L(w \otimes V_4) \leq 3$. By (ii) we have

(5) $L(\{w, z\}) \subset L(\{w, u \otimes V_4\})$ for at most 3 nodes $z \in V_3 \setminus V_4$. Similarly

(6) $L(\{u, z\}) \subset L(\{w, u \otimes V_4\})$ for at most 3 nodes $z \in V_3 \setminus V_4$.

By (i) there is $z \in V_3 \setminus V_4$ such that (5) and (6) both do not hold for $z$.

But $L(\{u, z\}) \cap L(\{w, z\}) = \emptyset$, because labels from this intersection have already been deleted from the edge $\{w, z\}$. Thus

$$\# L(\{u \cup V_4 \otimes \{u, w\}) \geq \# L(V_3 \otimes \{u, w\}) + 2.$$ 

Starting with $V_4 = \emptyset$ and carrying out this construction 3 times gives a set of 3 nodes $z_1, z_2, z_3$ such that

$$6 \leq \# L(\{z_1, z_2, z_3 \otimes \{u, w\}) \leq 5.$$ 

\[ \square \]

**Corollary 4.** Let $X = \{x_1, \ldots, x_n\} \subset \{0, 1\}^p$, $Y = \{y_1, \ldots, y_m\} \subset \{0, 1\}^p$, let $K(x_1, \ldots, x_n) \geq np$, let $(1 - O(\log p + Y)/p) \leq 1 + 1/\# Y$, and suppose $Y$ 2-codes $X$ k-robustly. Then $4/5 \geq Y > k$.

**Theorem 2.** Let $G = (V, E, L)$ be an edge-labelled graph, $\# V = n \gg 1$, and for all $V' \subset V$, let $\# L((V' \otimes V') \cap E) \leq c \# V'$. Then $l(G) \leq 4 c n^{1-\varepsilon}$ where $\varepsilon < 1/(12c)$.

**Proof.** By hypothesis, every edge has at most 2c labels. We show $l(G) \leq 2 n^{1-\varepsilon}$ if every edge has at most 1 label.

For every node $v \in V$ and any edge label $I$ that occurs on at least $n^\varepsilon + 1$ of the edges adjacent to $v$, put label $I$ on $v$ and delete it from the edges adjacent to $v$. By this at most $n^{1-\varepsilon}$ labels are put on every node.

Next, for each $v \in V$, partition the edges adjacent to $v$ into $\tau = n^\varepsilon$ classes $E^1_v, \ldots, E^{\tau}_v$ such that in every class $E^i_v$, every label occurs on at most one edge of $E^i_v$. Partition $E$ into classes $E^I, I \leq i \leq \tau \leq n^\varepsilon$, by $(u, v) \in E^I$ if $[u, v] \in E^I \cap E^I$. For all $i, j$, let $G^{ij} = (V, E^{ij}, L^{ij})$ where $L^{ij}$ is $L$ restricted to $E^{ij}$. Then in $G^{ij}$ for every node $v$, all edges adjacent to $v$ have different labels. We will show $l(G^{ij}) \leq n^{1-3\varepsilon}$.

For every vertex $v$ that is adjacent to at most $n^{1-3\varepsilon}$ edges, put all labels occurring on these edges on $v$. Delete $v$ and its adjacent edges from $G^{ij}$. Continue this process as long as possible. If finally all of $G^{ij}$ is deleted, we are done. Otherwise we are left with an edge-labelled graph $G' = (V', E', L')$ with at most $n$ nodes. Every node $v$ has at least $n^{-1-3\varepsilon}$ neighbors and the edges joining $v$ with its neighbors have all different labels. We will derive a contradiction from this.

We consider the adjacency matrix $A'$ of $G'$ and use the following fact [H].

For natural numbers $m, n, j, k$, let $z(m, n, j, k)$ be the smallest number $z'$ such that every $m \times n$ matrix with $z'$ ones contains a $j \times k$ minor $\mu$ that consists of ones
only. Then \( z(m, n, j, k) \leq 1 + km + (j-1)^{1/k} m^{1-1/k} n \). In particular,
\[
z(n, n, 9c^3, 2c) \leq 1 + 2cn + (9c^3)^{1/(2c)} n^{2-1/(2c)} \leq n^{2-6c}
\]
if \( 6c < 1/(2c) \) and \( n \) is large enough.

Let \( \mu' \) be a \( 9c^3 \times (2c) \) minor of \( A' \) that consists only of ones. Every one in \( \mu' \) corresponds to an edge \( e \in E \). Replace each one in \( \mu' \) by the label \( L^j(e) \) of the corresponding edge. Call the resulting matrix \( \mu \). Every label occurs in each row and column of \( \mu \) at most once. We make the following observation.

If \( R \) is a set of at most \( 2c \) rows of \( \mu \), then at most \( 4c^2 \) different labels occur in \( R \). Each label occurs in at most \( 2c \) more rows of \( \mu \). Thus there is a row \( r \) of \( R \) consisting only of labels that are not yet in \( R \). Starting with \( R \) an arbitrary row of \( M \) and repeating this process \( 2c \) times gives a \( (2c+1) \times (2c) \) minor \( M' \) of \( \mu \) that contains \( 4c^2 + 2c \) different labels. The rows and columns of \( \mu \) correspond to a set \( V' \) of \( 4c + 1 \) vertices of \( \mu \). Thus
\[
2c(2c+1) \leq \# L((V' \otimes V') \cap E) \leq c(4c+1).
\]

**Corollary 5.** Let \( X = \{ x_1, \ldots, x_n \} \subset \{0, 1\}^p \), \( Y = \{ Y_1, \ldots, Y_m \} \subset \{0, 1\}^q \), \( p > \log \log n \), and \( \alpha \) be such that \( K(x_1, \ldots, x_n) \equiv pn \) and \( Y \) 2-codes \( X \) \( k \)-robustly. Then \( 16a(bn)^{1-\alpha} > k \) for \( \alpha = 1/(24a) \) and \( n \) large enough.

6. A lower bound. We want to establish lower bounds for \( l(G) \) where \( G \) satisfies the property in Theorem 2.

**Theorem.** For \( c \leq 2 \), there exists \( G = (V, E, L) \), \( \# V = n \), having an edge labelling satisfying
\[
\# L((V' \otimes V') \cap E) \leq c \# V' \quad \text{for all } V' \subset V
\]
and
\[
l(G) \geq c \cdot n^a \quad \text{where } a > \frac{1}{2} \left( 1 - \frac{1}{4c - 2} \right).
\]

**Proof.** Let \( \delta = (\frac{1}{2} - 1/(4c-2)) \cdot \frac{1}{2a} \). Let \( \alpha = \frac{1}{2} - 1/(4c-2) - \delta = (c-1)/(2c-1) - \delta \) and let \( G = (V, E) \) be a random graph with \( n \) nodes and \( n^{1+\alpha} \) edges, where all such graphs are equally likely. We first show that with high probability,

\[
\#(V' \otimes V') \cap E \leq c \# V' \quad \text{for all } V' \subset V \quad \text{with } \# V' \leq n^a.
\]

The probability that for any set \( V' \) of cardinality \( j \leq n^a \), (***) does not hold, is at most
\[
W_j = \binom{n}{j} \left( \frac{1}{2} \right)^j \left( \frac{n-j}{2} \right) \left( \frac{n^1+n-aj}{2} \right) \left( \frac{n}{2} \right)^{j-1}
\]
\[
\leq \left( \frac{ne^j}{j} \right) \left( \frac{aj}{2j} \right) \left( \frac{n^1+n-aj+1}{2} \right) \left( \frac{n}{2} \right)^{j-1}
\]
\[
\leq \left( \frac{ne^j}{j} \right) \left( \frac{aj}{2j} \right) \left( 2.1 n^{-1+\alpha} \right) \left( \frac{n}{2} \right)^{j-1}
\]
\[
\leq (a^j)^{1-1/a} n^{-1+a+1/j} \left( \frac{n}{2} \right)^{j-1}.
\]
For $2c+2 \leq j \leq \log n$, we estimate
\[
W_j \leq (\log^2 n \cdot n^{-1/2-1/(4c+2)-\delta+1/\epsilon})^j \leq (n^{-1/2-1/6+1/2})^2 \delta = n^{-2}.
\]
For $\log n < j \leq n^{a}$, we have
\[
W_j \leq (c_2 n^{a(1-1/c)-1/a+1/\epsilon})^{j} = (c_2 n^{(2c-1)/c+1/a})^{j} = (c_2 n^{-\delta(2c-1)/\epsilon})^{j} \leq (n^{-\delta})^{j} \leq n^{-2}.
\]
Hence the probability that (***) does not hold is at most
\[
\sum_{j=2c+2}^{n^{a}} W_j \leq n^{-a} n^{-2} \leq n^{-1}.
\]
Next we make use of the fact that with probability $1-o(n^{-1})$, the degree of every node in $G$ is bounded by $3n^{a}$ [ER]. Therefore there exists a graph $G$ with $n$ nodes, $n^{1+a}$ edges, such that the degree of every node in $G$ is bounded by $3n^{a}$ and (***) holds for $G$.

Let $L$ be any edge labelling of $G$, which labels every edge with exactly 1 label $l \in [1, \cdots, n^{a}]$. Let $V' \subseteq V$. Then $|L((V' \cap V') \cap E)| \leq \min \{n^{a}, \#((V' \cap V') \cap E)\} \leq c \neq V'$.

Suppose we choose $L$ randomly in such a way that edges are labelled independently, and such that for each edge, each label is equally likely. Let $v$ be any node of $G$, let $d$ be the degree of $v$ and let $l$ be any label. Then the probability that $j$ or more edges adjacent to $v$ have label $l$ is at most
\[
\left(\frac{d}{j}\right) \left(\frac{1}{n^{a}}\right)^{j} \leq \left(\frac{3n^{a}d}{j}\right) \left(\frac{1}{n^{a}}\right)^{j} = \left(\frac{3n^{a}}{j}\right) \left(\frac{1}{n^{a}}\right)^{j} = O(n^{-1})
\]
if $j \geq \log n$. Therefore the probability that log $n$ or more edges adjacent to the same node as $G$ have the same label is at most $n \cdot n^{a} \cdot O(n^{-3}) = O(n^{-1})$. Hence there is a labelling $L$ such that for every $l$ and $V$ label, $l$ occurs on at most $\log n$ edges adjacent to node $V$. No matter how we transform $L$ into a node labelling $L'$, we have $\sum_{v \neq L'(v)} \leq n^{1+a}/\log n$. This proves the theorem.

7. Simple 2-coding revisited. If $Y$ is a subset of $\{0, 1\}^{p}$ of size $m$, then $G(E_{p}, Y)$ may have at least $m \log m$ labels. This means the number of pairs in $Y$ that code some $e_{i}$ grows faster than the size of $Y$. But at least for the obvious example to demonstrate this, the $(\log m)$-dimensional subcube, one notices that for $\log m \ll p$, only a small subset of the $e_{i}$ can be coded by many pairs. Thus there is hope that, disregarding a small subset of $\{e_{1}, \cdots, e_{p}\}$, the remaining $e_{i}$ have a much smaller number of pairs which simply 2-code them.

For $X \subseteq \{0, 1\}^{p}$ and $1 \leq i \leq p$, define $r(X, i)$ as the number of edges in $G(E_{p}, X)$ with label $i$.

For $1 \leq k \leq p$ define
\[
r_{k}(X) = \min_{D \subseteq \{1, \cdots, p\} \setminus k} \sum_{i \in D} r(X, i)
\]
and
\[
p_{k}(X) = \min_{D \subseteq \{1, \cdots, p\} \setminus k} \max_{i \in D} r(X, i),
\]
and for $m \in \mathbb{N}$,
\[
r_k(m) = \max_{|X| \leq m} r_k(X)
\]
and
\[
\rho_k(m) = \max_{|X| \leq m} \rho_k(X).
\]

One checks easily that for $m \leq 2^n$, $\rho_p(m) = [m/2]$, whereas from Lemma 2 it follows that $r_k(m) = \Theta(m \log m)$.

Let $\ln(x)$ denote the natural logarithm of $x$ and $\ln^k(x) = [\ln(x)]^k$.

**Theorem 4.** There are constants $\alpha > 0$ and $C, h \geq 1$ such that for all $1 \leq k < p$ and $m/(p-k) \geq C/\alpha$,
\[
\rho_k(m) \leq \alpha \frac{m}{p-k} \ln^3 \left( \alpha \frac{m}{p-k} + h \right).
\]

**Corollary 6.** For any $\varepsilon > 0$ and $m = O(p)$,
\[
\rho_{(1-\varepsilon)p}(m) = O(1).
\]

Theorem 4 follows from the following:

**Lemma 5.** There are constants $\beta > 0$, $h \geq 1$ such that for any $X \subset \{0, 1\}^p$,
\[
\frac{1}{|X|} \sum_{i=1}^p \ln^3(r(X, i) + h) \leq \beta |X|.
\]

**Proof of Theorem 4.** Assume
\[
\rho_k(m) > r := \alpha \frac{m}{p-k} \ln^3 \left( \alpha \frac{m}{p-k} + h \right).
\]

Then there exists $X \subset \{0, 1\}^p$ of size $m$ such that $r(X, i) := r_i > r$ holds for more than $p-k$ labels $i \in \{1, \ldots, p\}$. Define
\[
F(x) = \frac{x}{\ln^3(x + h)}.
\]

Later it will be shown that for appropriate $h \geq \varepsilon^2$, $F(x)$ is monotonically increasing for $x \geq 0$. Hence,
\[
\sum_{i=1}^p F(r_i) \equiv \sum_{i, r_i > r} F(r_i) > (p-k)F(r)
\]  
\[= \alpha m \frac{\ln^3(\alpha(m/(p-k)) + h)}{\ln^3(\alpha(m/(p-k)) \ln^3(\alpha(m/(p-k)) + h))}
\]

For an appropriate $C \geq 1$,
\[
x + h \geq \ln^3(x + h)
\]
holds for all $x \geq C$. Thus if $\alpha m/(p-k) \geq C$, then
\[
\ln^3 \left( \frac{\alpha m}{p-k} \ln^3 \left( \frac{\alpha m}{p-k} + h \right) + h \right) \leq \ln^3 \left( \frac{\alpha m}{p-k} \ln^3 \left( \frac{\alpha m}{p-k} + h \right) + h \right)
\]  
\[\leq \ln^3 \left( \left( \frac{\alpha m}{p-k} + h \right)^2 \right) = 8 \ln^3 \left( \frac{\alpha m}{p-k} + h \right).
\]

Therefore, $\sum_{i=1}^p F(r_i) > \alpha m/8$. But this contradicts Lemma 5 if $\alpha/8 \geq \beta$. \qed
Proof of Lemma 5. Define \( h = e^2 \approx 7.389 \) and \( \gamma = 0.16 \).
\[
g(x) = \frac{x}{\ln^3(x+h)}, \quad \text{for } x \geq 0
\]
and
\[
f(n) = 1 + \sum_{m=2}^{n} \frac{1}{m \ln^3(m)}, \quad \text{for } n \in \mathbb{N}, \quad n \geq 1.
\]
We will show in the Appendix:

1. \( g(x) \leq 0.16x \) for all \( x \geq 0 \),
2. \( g'(x) \geq 0 \) for all \( x \geq 0 \),
3. \( g''(x) \leq 0 \) for all \( x \geq 0 \),
4. \( 1 \leq f(n) \leq f(n+1) \leq 5 \) for all \( n \geq 1 \).

Now let \( X \subseteq \{0,1\}^p \). Lemma 5 follows from the following:

**Proposition.** If \( n = \left\lfloor \frac{r}{r_i} \right\rfloor > 0 \), then
\[
\sum_{i=1}^{r} g(r_i) \leq f(n)|X|.
\]

**Proof.** The proof is by induction on \( n \). Define \( r = \max_{i \leq n} r_i \). For each \( i \), the edges with label \( i \) are a matching. Hence, \( |X| \leq 2r \). For all \( 1 \leq h \leq n_0 = 98 \), we get
\[
\sum_{i=1}^{r} g(r_i) \leq n g(r) \leq 98 \gamma \frac{r}{\ln^3(r+h)} \leq \frac{49 \gamma |X|}{\ln^3(r+h)} \leq \frac{49 \gamma |X|}{2^3} \leq f(n)|X|.
\]
Thus, the claim holds for all \( n \leq n_0 \). Now assume
\[
n + 1 > n_0 = 98,
\]
and the claim is true for all \( n \leq n_0 \). We may assume that \( r_1 \equiv r_2 \equiv \cdots \equiv r_{n+1} > r_{n+2} = \cdots = r_p = 0 \). Define for \( l \in \{0,1\} \),
\[
X' = \{ x \in X | x_{n+1} = l \},
\]
and for \( 1 \leq i \leq n \), the number of edges in \( G(E \subseteq X') \) with label \( i \), this means we cut \( X \) in dimension \( n + 1 \). Obviously,
\[
X = X^0 \cup X^1, \quad r_i = r_i^0 + r_i^1 \quad \text{for } 1 \leq i \leq n.
\]
\( D = D^0 \cap D^1 \) and \( d' = |D^1| \). One can check easily
\[
|X'| \geq \max \{ r_{n+1}, d' + 1 \} \quad \text{for } l = 0, 1.
\]
Define \( \Delta g(x,y) = g(x) + g(y) - g(x+y) \). Now
\[
\sum_{i=1}^{p} g(r_i) = \sum_{i=1}^{n} g(r_i) + \sum_{i=1}^{n} g(r_i^0) + \sum_{i=1}^{n} g(r_i^1) - \sum_{i \in D} \Delta g(r_i^0, r_i^1) + g(r_{n+1}).
\]
Applying the induction hypothesis to \( X^0 \) and \( X^1 \) gives
\[
\sum_{i=1}^{p} g(r_i) \leq |X^0|f(d^0) + |X^1|f(d^1) - \sum_{i \in D} \Delta g(r_i^0, r_i^1) + g(r_{n+1}).
\]
The idea of the proof is as follows: if \( D \) is large, then \( \sum_{i \in D} \Delta g(r_i^0, r_i^1) \) is large enough to compensate the term \( g(r_{n+1}) \); otherwise one of the \( d' \) must be relatively small, such
that the difference between $|X'| f(n+1)$ and $|X'| f(d')$ is bigger than $g(r_{n+1})$. We have to distinguish several cases. First, we state some more properties of $f$ and $g$ which will be proved in the appendix.

\[(g4)\quad \Delta g(x, y) \equiv 0 \quad \text{for all } x, y \equiv 0,
\]

\[(g5)\quad \Delta g(x, y) \equiv \Delta g(x, z) \quad \text{for all } 0 \leq x \text{ and } 0 \leq y \leq z,
\]

\[(g6)\quad \Delta g(1, 1) \equiv 0.0298 \gamma,
\]

\[(g7)\quad \Delta g(x, y) \equiv 1.4 \frac{g(x)}{\ln (x + h)} \quad \text{for all } 0 \leq x \leq y \text{ and } y \geq 3 h.
\]

Define $\delta f(n, m) = f(n) - f(m)$ for $1 \leq m \leq n$. Then

\[(f2)\quad \delta f(n, m) \equiv \frac{1}{4 \ln^2 (m + h)} \quad \text{for all } 16 \leq m \leq \frac{3}{2} n.
\]

Case 1. $\exists l$ with $d' \equiv 2/3 n$. Assume $l = 1$. Then (7.4) yields

\[
\sum_{i=1}^{n} g(r_i) \equiv |X'| |f(d')| + |X'| |f(d')| + g(r_{n+1})
\]

\[
\equiv (|X'| + |X'|) f(n + 1) + g(r_{n+1}) - |X'| \delta f(n + 1, d').
\]

If $d' \equiv 16$ and $d' \equiv r_{n+1}$, we get

\[
g(r_{n+1}) - |X'| \delta f(n + 1, d') \equiv g(r_{n+1}) - d' \frac{1}{4 \ln^2 (d' + h)} \quad \text{by (7.3) and (f2)}
\]

\[
\quad \leq g(r_{n+1}) - \gamma \frac{d'}{\ln^2 (d' + h)} \quad \text{since } \frac{1}{4} \geq \frac{\gamma}{\ln^2 (d' + h)}
\]

\[
= g(r_{n+1}) - g(d') \equiv 0 \quad \text{by (g2).}
\]

If $15 \leq d' \leq r_{n+1}$, we get

\[
g(r_{n+1}) - |X'| \delta f(n + 1, d') \equiv g(r_{n+1}) - d' \sum_{j=d', \cdots, 15} \frac{1}{\ln^2 j} \equiv g(15) - \frac{15}{16 \ln^2 16} < 0.
\]

If $16 \leq d' \leq r_{n+1}$, we have

\[
g(r_{n+1}) - |X'| \delta f(n + 1, d') \equiv g(r_{n+1}) - \frac{1}{4 \ln^2 (d' + h)} \equiv g(r_{n+1}) - \gamma \frac{r_{n+1}}{\ln^3 (r_{n+1} + h)} = 0.
\]

If $1 \leq d' \leq 15$ and $d' \equiv r_{n+1}$, we have

\[
g(r_{n+1}) - |X'| \delta f(n + 1, d') \equiv \gamma \frac{r_{n+1}}{\ln^3 (r_{n+1} + h)} - \frac{1}{r_{n+1} (d' + 1) \ln^2 (d' + 1)}
\]

\[
\quad \leq \frac{r_{n+1}}{\ln^2 (d' + 1)} \left( \gamma \frac{1}{\ln (d' + 1)} - \frac{1}{d' + 1} \right) < 0,
\]

because $\ln (d' + 1)/(d' + 1) > \gamma$ for all $d' \in \{1, \cdots, 15\}$. Finally, if $d' = 0$, then

\[
g(r_{n+1}) - |X'| \delta f(n + 1, d') \equiv \gamma \frac{r_{n+1}}{\ln^3 (r_{n+1} + h)} - \frac{1}{2 \ln^2 2}
\]

\[
\quad \leq r_{n+1} \left( \gamma \frac{1}{8 \ln^2 2} \right) < 0.
\]
Thus, \( \sum_{i=1}^{d} g(r_i) \leq \vert X \vert f(n+1) \). We now assume

\[
d^4 \geq \frac{3}{2} n \quad \text{for } l = 0, 1.
\]

**Case 2.** \( r_{n+1} \leq c_1 n \ln^3 (n+h) \), where \( c_1 = 0.0099 \). By (7.1), \( n \geq 98 \geq e^{c_1 - 1/3} - h \). Thus \( \ln (n+h) \geq c_1^{-1/3} \) and

\[
c_1 n \ln^3 (n+h) \geq n.
\]

This implies

\[
g(r_{n+1}) \leq g(c_1 n \ln^3 (n+h)) = \frac{c_1 n \ln^3 (n+h)}{\ln^3 (c_1 n \ln^3 (n+h) + h)} \leq \gamma c_1 n.
\]

From (7.5) follows \( |D| \geq n/3 \). Thus

\[
\sum_{i=1}^{D} g(r_i) \leq \vert X \vert^0 f(d^0) + \vert X \vert^1 f(d^1) - \sum_{i \in D} \Delta g(r^0_i, r^1_i) + g(r_{n+1}) \\
\leq \vert X \vert f(n+1) - \sum_{i \in D} \Delta g(1, 1) + g(r_{n+1}) \quad \text{by (g5)}
\leq \vert X \vert f(n+1) - \frac{n}{3} 0.0298 \gamma + \gamma c_1 n \quad \text{by (g6)}
\leq \vert X \vert f(n+1) - \frac{0.0298}{3} \quad \text{since } 0.0298 \leq c_1.
\]

Let us now assume

\[
r_{n+1} \geq c_1 n \ln^3 (n+h).
\]

From (7.1) it follows that

\[
r_{n+1} \geq n \geq n_0 \geq 98 \geq 6h.
\]

For \( 1 \leq i \leq h \), define \( z_i = \min \{ r^0_i, r^1_i \} \) and \( v_i = \max \{ r^0_i, r^1_i \} \). We have

\[
v_i \geq \frac{r^0_i + r^1_i}{2} \geq 3h.
\]

**Case 3.**

\[
\sum_{i \in D} g(z_i) \geq \frac{1}{8} \frac{r_{n+1}}{\ln^2 (r_{n+1} + h)}.
\]

Then

\[
\sum_{i \in D} g(r^0_i, r^1_i) = \sum_{i \in D} \Delta g(z_i, v_i)
\]

\[
\leq \sum \frac{g(z_i)}{\ln (z_i + h)} \quad \text{by (7.9) and (g7)}
\leq 1.4 \frac{\sum g(z_i)}{\ln (\sum g(z_i) + h)} \geq 1.4 \frac{(1/8) \ln^2 (r_{n+1}/(r_{n+1} + h))}{\ln ((1/8)(r_{n+1}/\ln^2 (r_{n+1} + h) + h))}
\geq 0.175 \frac{r_{n+1}}{\ln^2 (r_{n+1} + h)} \geq g(r_{n+1}), \quad \text{since } 0.175 \geq \gamma.
\]

\( \)
Hence in (7.4),
\[ \sum_{i=1}^{n} g(r_i) \geq (|X|^\gamma + |X|)|f(n+1) + g(r_{n+1}) - \sum_{i=0}^{n} \Delta g(r_i, r_i') \geq |X|f(n+1). \]

It remains the case that
\[ \sum_{i \in D} g(z_i) \leq \frac{1}{8 \ln^2 (r_{n+1} + h)}. \]

Define for \( l = 0, 1 \), \( B^l = \{ i | r_i' > r_i \} \) and \( b^1 = |B^1| \). Since \( b^0 + b^1 \leq n \), we may assume \( b^1 \leq n/2 \).

If we remove from \( G(E_\mu, X^1) \) edges with labels not in \( B^1 \), the remaining graph consists of some connected components \( G(E_\mu, Y^1), \cdots, G(E_\mu, Y^n) \) where \( \cup_{i \in B^1} Y^i = X^1 \). Let us denote by \( y^i \) the number of edges in \( G(E_\mu, Y^i) \) with label \( i \). Each such graph contains only labels from \( B^1 \). Hence by the induction hypothesis,
\[ \sum_{i=1}^{n} g(y^i) \leq |Y|^\gamma f \left( \frac{n}{2} \right) \]

and
\[ \sum_{i \in B^1} g(r_i') \leq \sum_{j=1}^{n} g(y^i) \text{ since } r_i' = \sum_{j=1}^{n} y^i \]
\[ \leq \sum_{j=1}^{n} |Y|^\gamma f \left( \frac{n}{2} \right) = |X|^\gamma f \left( \frac{n}{2} \right). \]

Thus we can conclude
\[ \sum_{j=1}^{p} g(r_j) \leq \sum_{i=1}^{n} g(r_i') + \sum_{i \in B^1} g(r_i') + \sum_{i \in B^1} g(r_i') + g(r_{n+1}) \]
\[ \leq |X|^\gamma f(n+1) + |X|^\gamma f(n+1) - |X|^\gamma f \left( \frac{n}{2} \right) \]
\[ + \sum_{i=1}^{n} g(z_i) + g(r_{n+1}) \quad \text{since } r_i' = z_i \text{ for } i \notin B^1 \]
\[ \leq |X|^\gamma f(n+1) - \frac{1}{4 \ln^2 (n/2 + h)} + \frac{1}{8 \ln^2 (r_{n+1} + h)} \]
\[ + \frac{r_{n+1}}{\ln^3 (r_{n+1} + h)} \quad \text{by (f2)} \]
\[ = \left( 1 - \frac{1}{4 \ln (r_{n+1} + h)} \right) \left( 1 - \frac{1}{8 \ln (r_{n+1} + h)} \right) \]
\[ \leq |X|^\gamma f(n+1). \]

This completes the proof of the Proposition and Theorem 4. \( \Box \)

For \( Y = \{0, 1\}^p \) and \( Q = \{1, \cdots, p\} \), let \( G^{Q}(E_\mu, Y) \) denote the subgraph of \( G(E_\mu, Y) \) that has the same set of nodes, but only edges with labels in \( Q \).

The previous result can then be stated as follows. For any \( \varepsilon, \mu > 0 \), there is a constant \( A(\varepsilon, \mu) \) such that for any \( Y = \{0, 1\}^p \) of size at most \( \mu p \), one can find a set \( Q = \{1, \cdots, p\} \) of size at least \( (1 - \varepsilon)p \) such that in \( G^{Q}(E_\mu, Y) \) the occurrence of each label is bounded by \( A(\varepsilon, \mu) \), and hence \( G^{Q}(E_\mu, Y) \) has less than \( A(\varepsilon, \mu)p \) edges.
This does not necessarily imply that in \( G^Q(E_p, Y) \) the labelled edges are distributed in a nice uniform manner such that every node gets about the same number of labels. There might exist a neighborhood of nodes in \( G^Q(E_p, Y) \) where each node has a high degree (increasing with \( p \)), and some of them might have to accept many labels. It will be shown that the structure of the cube excludes such cases. Define

\[
I_k(Y) = \min_{Q \subseteq \{1, \ldots, p\}} \min_{\text{transformation } L \in G^Q(E_p, Y)} \max_{\|Q\| \leq k} \#L(v)
\]

and

\[
I_k(m) = \max_{\|Y\| = m} I_k(Y).
\]

Obviously, for \( n - (\log p)/2 \leq k \leq n \), it holds that \( I_k(p) = \Theta(\log p) \).

**Theorem 5.** For any \( e, \mu > 0 \) there exists a constant \( R(e, \mu) \) such that

\[
I_{(1-\varepsilon)p}(\mu p) \leq R(e, \mu), \quad \text{for any } p.
\]

**Proof.** From Corollary 6, we know that there is a constant \( A = A(e/2, \mu) \) such that \( I_{(1-\varepsilon)p}(\mu p), (\mu p) \leq A \) for all \( p \).

Let \( R = R(e, \mu) > 10A/\epsilon g(1) \). If the theorem is false, then there exists \( p \in \mathbb{N} \) and \( Y \subseteq \{0, 1\}^p, |Y|=\mu p \) such that for any \( Q \subseteq \{1, \ldots, p\} \) of size at least \( (1-\varepsilon)p \) and any transformation \( L \) of labels to nodes for \( G^Q(E_p, Y) \), we find a node \( v \) with \( \#L(v) > R \).

By Corollary 6, for the given \( Y \) there exists a set \( U \subseteq \{1, \ldots, p\} \) of size \( (1-\varepsilon/2)p \) such that \( G^U(E_p, Y) \) has less than \( Ap \) edges. Among all transformations of labels in \( G^U(E_p, Y) \), choose \( L \) that minimizes the function

\[
F(L) := \sum_{v \in Y} \max\{0, \#L(v) - R\}.
\]

By assumption, for \( L \) and also any restriction \( \bar{L} \) of \( L \) to a graph \( G^Q(E_p, Y) \) where \( Q \) is a subset of \( U \) of size \( (1-\varepsilon)p \), \( F(L) \) and \( F(\bar{L}) \) are positive. \( L \) defines an orientation of the edges in \( G^U(E_p, Y) \): edge \( \{v, v'\} \) is changed into the directed edge \( (v, v') \) iff \( L \) assigns the label of \( (v, v') \) to \( v' \). Let us call this directed graph \( H \).

Let \( Z \subseteq Y \) be the set of all nodes from which there is a path of length \( \Omega 0 \) in \( H \) to a node \( v \) with \( \#L(v) > R \), and let \( \bar{H} \) be the subgraph of \( H \) induced by \( Z \). By assumption, \( Z \) is nonempty, since there is at least one node that gets more than \( R \) labels. Notice that for \( z \in Z, \#L(z) \) equals the indegree of \( z \) in \( \bar{H} \).

**Claim 1.** Each node of \( Z \) has indegree at least \( R \) in \( \bar{H} \).

**Proof.** Assume \( z \in Z \) has indegree less than \( R \), and let \( z = z_0, z_1, \ldots, z_i \) be a path in \( \bar{H} \) from \( z \) to a node \( z_i \) with indegree bigger than \( R \). By definition of \( Z \), such a path must exist.

Change \( L \) into \( \bar{L} \) by assigning for \( 0 \leq i < l \) the label on edge \( \{z_i, z_{i+1}\} \) to node \( z_i \) instead of \( z_{i+1} \). Since in a cube all edges adjacent to a node have different labels, we have \( \#L(z_i) = R, \#L(z_i) = \#L(z) = 1 \leq R \) and \( \#L(z) = \#L(z) \) for all remaining \( z \in Y \). Hence

\[
F(L) > F(\bar{L}),
\]

which contradicts the minimality of \( L \). \( \Box \)

Therefore, we now conclude that \( \bar{H} \) has at least \( R|Z| \) edges.

Since \( \bar{H} \) is a subgraph of \( H \), and \( H \) has the same number of edges as \( G^U(E_p, Y) \), we know that \( R|Z| \leq Ap \). Hence

\[
|Z| \leq \frac{Ap}{R}.
\]
On the other hand, $G(E_p, Z)$ must have at least $\epsilon p/2$ different labels; otherwise, deleting this set of labels from $U$ would yield a subset $Q$ of $\{1, \cdots, p\}$ of size at least $(1-\epsilon)p$ such that $L$ restricted to $G^Q(E_p, Y)$ does not assign more than $R$ labels to any node.

From the Proposition in the proof of Lemma 5, it follows that

$$|Z| \geq \frac{1}{f(n)} \sum_{r_i} g(r_i),$$

where $r_i =$ number of edges in $G(E_p, Z)$ with label $i$ and $n =$ number of $r_i > 0$.

Since $g$ is monotonic and $f$ is bounded by five, we get

$$|Z| \geq \frac{1}{5} \cdot \frac{\epsilon}{2^p} \cdot g(1) = \frac{\epsilon}{10} g(1)p.$$

Combining the two inequalities for $|Z|$ gives

$$\frac{\epsilon}{10} g(1) \geq \frac{A}{R}.$$

Hence

$$R \leq \frac{10A}{\epsilon g(1)}.$$

This contradicts the definition of $R$. $\square$

**Corollary 7.** If $Y \subseteq \{0, 1\}^p$, $\# Y = O(p)$ and $Y$ simply 2-codes $E_p$, then $Y$ 2-codes $E_p$, $O(1)$-robustly.

8. Problems. (i) How good are the bounds of Theorems 1 and 2?

(ii) Consider 3-coding or more general $r$-coding for $r \geq 3$. Now $G(x, y)$ becomes a hypergraph, and a result analogous to Lemma 3 holds. Are there, even in the case of simple 3-coding, any nontrivial bounds on $I(G(x, y))$?

9. Appendix. Proof of Properties (g1)-(g7) and (f1)-(f2). Let $h = e^2$, let $\gamma = 0.16$ and for $x \approx 0$ let

$$g(x) = \gamma \frac{x}{\ln^3(x + h)}.$$

(g1) is obvious. To prove (g2) we get

$$g'(x) = \gamma \frac{\ln^3(x + h) - 3x \ln^2(x + h)/(x + h)}{\ln^6(x + h)} = \gamma \frac{1}{\ln^3(x + h)} \left[ 1 - \frac{3x}{(x + h) \ln(x + h)} \right].$$

Let $\varphi(x) := (x + h) \ln(x + h) - 3x$.

Then for $x \approx 0$, $g'(x) \approx 0 \Rightarrow \varphi(x) \approx 0$. We have $\varphi'(x) = \ln(x + h) - 2$ and $\lim_{x \to 0} \varphi(x) = \infty$, and hence $x = 0$ is the only minimum of $\varphi$ for $x \approx 0$. Since $\varphi(0) = 2e^2$, we get $\varphi(x) \approx 0$ for all $x \approx 0$, and $g'(x) \approx 0$ for all $x \approx 0$.

$$g''(x) = \gamma \left( \frac{-3}{\ln^4(x + h)} \frac{1}{x + h} \left[ 1 - \frac{3x}{(x + h) \ln(x + h)} \right] - \frac{1}{\ln^4(x + h)} \left[ \frac{3(x + h) \ln(x + h) - 3x(\ln(x + h) + 1)}{(x + h)^2 \ln^2(x + h)} \right] \right)$$

$$= -\gamma \frac{3}{\ln^5(x + h)(x + h)^2} [(x + h) \ln(x + h) - 3x + h \ln(x + h) - x]$$

$$= -\gamma \frac{3}{\ln^5(x + h)(x + h)^2} [(x + 2h) \ln(x + h) - 4x].$$
Let \( \varphi(x) := (x + 2h) \ln(x + h) - 4x \). Then for \( x \geq 0 \),
\[
\varphi'(x) = \ln(x + h) + \frac{x + 2h}{x + h} - 4,
\]
\[
\varphi''(x) = \frac{1}{x + h} + \frac{(x + h) - (x + 2h)}{(x + h)^2} - \frac{x}{(x + h)^2}.
\]

Since \( \varphi'(0) = 0 \), \( \varphi'(x) \geq 0 \) for \( x \geq 0 \) and \( \lim_{x \to \infty} \varphi(x) = \infty \), \( x = 0 \) is the only minimum of \( \varphi(x) \). From \( \varphi(0) = 4h \geq 0 \) it follows that

\[
(3) \quad g''(x) \leq 0 \quad \text{for all} \quad x \geq 0.
\]

Define \( \Delta g(x, y) = g(x + y) - g(x) - g(y) \). Calculation proves (6):
\[
\Delta g(1, 1) = 2 \gamma \left[ \frac{1}{\ln^3(1 + h)} - \frac{1}{\ln^3(2 + h)} \right] \geq 0.0298 \gamma.
\]

Assume \( 0 \leq x \) and \( 0 \leq y \leq z \). Since for all \( t \in [y, z] \), \( g'(x + t) \leq g'(t) \) by (3), we can conclude that \( g(x + z) - g(x + y) \geq g(z) - g(y) \). This yields \( g(x) + g(y) - g(x + y) \leq g(x) + g(z) - g(x + z) \), or

\[
(5) \quad \Delta g(x, y) \leq \Delta g(x, z) \quad \text{for all} \quad 0 \leq x \leq y \leq z.
\]

For \( \Delta g \) we can show the bound for \( 0 \leq x \leq y \):
\[
\Delta g(x, y) = g(x) + g(y) - g(x + y) \geq g(x) - x \sup_{z \in [x, x+y]} g'(z) = g(x) - x g'(y).
\]

This yields
\[
\Delta g(x, y) \geq g(x) - xy \frac{3y}{\ln^3(x + h)} \left( 1 - \frac{3y}{(y + h) \ln(y + h)} \right)
\]
\[
= g(x) \left[ 1 - \ln^3(x + h) \left( 1 - \frac{3y}{(y + h) \ln(y + h)} \right) \right].
\]

Since \( \ln(x + h) \leq \ln(y + h) \) and \( 0 \leq 3y \leq (y + h) \ln(y + h) \) (see proof of (2)),
\( \Delta g(x, y) \geq 0 \) follows from \( g(x) \geq 0 \). The case \( x > y \) follows from \( \Delta g(x, y) = \Delta g(y, x) \).

This proves (4). If \( x + h \geq (y + h)^{2/3} \), we get \( \ln(x + h) \geq (2/3) \ln(y + h) \) and
\[
\Delta g(x, y) \geq g(x) \frac{3y}{y + h} \frac{1}{\ln(y + h)} \geq g(x) \frac{3y}{y + h} \frac{2/3}{\ln(y + h)} \geq \frac{2}{3} \frac{3y}{y + h} \frac{g(x)}{\ln(x + h)}.
\]

If \( y \geq 3h \) then \( \Delta g(x, y) \geq \frac{3}{2} g(x)/\ln(x + h) \). If on the other hand \( x + h \leq (y + h)^{2/3} \),
we can bound \( \Delta g(x, y) \) by
\[
\Delta g(x, y) \geq g(x) \left[ 1 - \ln^3(x + h) \left( \frac{2}{3} \right) \right]
\]
\[
\geq 0.7 g(x)
\]
\[
\geq 1.4 \frac{g(x)}{\ln(x + h)} \quad \text{since} \quad \ln(x + h) \geq 2.
\]
Therefore we have shown (g7):

$$\Delta g(x, y) \geq 1.4 \frac{g(x)}{\ln (x + h)}$$

for all $0 \leq x \leq y$ and $y \geq 3h$.

For $n \in \mathbb{N}$, $n \geq 1$, define

$$f(n) = 1 + \sum_{m=2}^{n} \frac{1}{m \ln^2 m}.$$  

Then

$$f(n) \geq 1 + \sum_{m=2}^{\infty} \frac{1}{m \ln^2 m} = 1 + (\log_2 e)^2 \sum_{m=2}^{\infty} \frac{1}{m(m \log_2 m)^2}$$

$$= 1 + (\log_2 e)^2 \sum_{i=1}^{\infty} \sum_{m=2^{i+1}}^{2^i} \frac{1}{m \log_2 m}$$

$$\geq 1 + (\log_2 e)^2 \sum_{i=1}^{\infty} 2^i \frac{1}{2^i} = 1 + (\log_2 e)^2 \frac{\pi^2}{6} \leq 5.$$  

Thus (f1), $1 \leq f(n) \leq 5$, holds for all $n \geq 1$. Define $\delta f(n, m) = f(n) - f(m)$ for $1 \leq m \leq n$.  

For $16 \leq m \leq 2/3n$,

$$\delta f(n, m) = \sum_{j=m}^{n} \frac{1}{j \ln^2 j} \geq \sum_{j=m}^{\lfloor 3m/2 \rfloor} \frac{1}{j \ln^2 j}$$

$$\geq \frac{1}{\lfloor 3m/2 \rfloor \ln^2 \lfloor 3m/2 \rfloor} \geq \frac{1}{3 \ln^2 \lfloor 3m/2 \rfloor}.$$  

Since $m \geq 16$,

$$\lfloor 3m/2 \rfloor \geq \left(\frac{3}{2} + \frac{1}{20}\right) m \geq 1.6m \leq (16 + h)^{0.15} m \leq (m + h)^{1.15} \leq (m + h)^{4/73}.$$  

Hence

$$\ln \left[ \frac{3m}{2} \right] \geq \ln^2 (m + h)^{4/73} = \frac{4}{3} \ln^2 (m + h).$$  

This proves

$$\delta f(n, m) \geq \frac{1}{4} \ln^2 (m + h)$$

for all $16 \leq m \leq \frac{2}{3}n$.  

REFERENCES


