SIAM J. ALG. DISC. METH.
Vol. 6, No. 3, July 1985

011

CODING STRINGS BY PAIRS OF STRINGS*

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

Abstract. Let $X, Y \subset \{0, 1\}^*$. 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 $Y' \subset Y$ 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 $\# Y \geqq (k+1) \# X$.

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 $Y = \{\bigoplus_{x \in S} x \mid S \subset X\}$ where $\bigoplus$ denotes bitwise sum mod 2. Then $\# Y = 2^{\# X}$. Yet Y codes X robustly for $k = 2^{\# X - 1} - 1$. 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 $K(x|y) = O(\log |x|)$. We deliberately leave the implicit constant in the O-notation undefined. Let $X, Y \subset \{0, 1\}^*$. We say *Y 1-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 $Y' \subset Y$ with $\# Y' \leqq k$ the set of strings $Y \setminus Y'$ still 1-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 α, and that there are β times more strings in Y than in X. Then one would intuitively expect every $y \in Y$ to code at most α 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 $p \gg \alpha \log np$. Let $X = \{x_1, \cdots, x_n\} \subset \{0, 1\}^p$, $Y = \{y_1, \cdots, y_{\beta n}\} \subset \{0, 1\}^{\alpha p}$ and $K(x_1 \cdots x_n) \geqq np$ (i.e. $x_1 \cdots x_n$ is a random string). Then*

(a) *Each of $y \in Y$ codes at most α 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 $\{i_1, \cdots, i_s\} \subset \{1, \cdots, n\}$. Then

(1) $sp - O(s \log n) \leqq K(x_{i_1} \cdots x_{i_s})$ because $x_1 \cdots x_n$ is random [P, fact 5].

Suppose $y \in Y$ codes $x_{i_1}, \cdots, x_{i_s}$. Then

(2) $K(x_{i_1} \cdots x_{i_s}) \leqq \sum (K(x_{i_j}|y) + O(\log K(x_{i_j}|y)) + K(y) \leqq O(s \log p) + \alpha p$.

For $s = \alpha + 1$, (1) and (2) imply $(\alpha+1)p - O(\alpha \log n) \leqq \alpha p + O(\alpha \log p)$. Hence $p - O(\alpha \log np) \leqq 0$. This proves (a).

Suppose (b) is false. Then

$$\begin{aligned} \alpha\beta n &\geqq \sum_j \# \{x | y_j \text{ codes } x\} && \text{by (a)} \\ &= \sum_i \# \{y | y \text{ codes } x_i\} \\ &> (n/2) 2\alpha\beta = \alpha\beta n && \text{by assumption.} \end{aligned}$$ □

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 ith bit is the mod 2 sum of the ith bits of y and z for $1 \leqq i \leqq p$. For $1 \leqq i \leqq p$, let $e_i \in \{0, 1\}^p$ be the string which has 1 in the ith position and 0's in all other positions. Let $E_p = \{e_1, \cdots, e_p\}$. Let $\mathbf{0} \in \{0, 1\}^p$ be the string consisting of p 0's.

Let $X, Y \subset \{0, 1\}^p$. We say Y *simply* 2*-codes* X if for all $x \in X$ there are two strings $y, z \in Y$ such that $x = z \oplus y$. We say Y simply 2-codes X k-robustly if for all $Y' \subset Y$ with $\# Y' \leqq k$ 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 $i \leqq p$ and $Y_{p+1} = \mathbf{0}$.

Intuition suggests that in this example for $i \leqq p$, 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 $y_i = \bigoplus_{j \neq i} x_j$ for $i \leqq p$ and $y_{p+1} = \bigoplus_{j=1}^{p} x_i$.

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 $X, Y \subset \{0, 1\}^p$ the following edge-labelled graph $G(X, Y) = (V, E, L)$: $V = Y$ is the vertex set. For all $y, z \in Y$, there is an edge $\{y, z\} \in E$ 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.

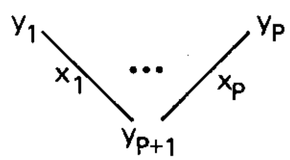

FIG. 1

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 $Y' = \{y | x \in L(y)\}$ has the property of a set of strings each of which carries specific information about x: $Y \setminus Y'$ 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 $i \leqq p$ and $L'(y_{p+1}) = \varnothing$.

3. Transformation of labellings for simple 2-coding. We define the labelled p-dimensional cube $C_p := G(E_p, \{0,1\}^p)$. If $Y \subset \{0,1\}^p$, then $G(E_p, Y)$ is a subgraph of C_p.

Any node labelling of C_p has to distribute $p2^{p-1}$ 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 $l(C_p) \geqq p/2$. This shows that the function $l(\cdot)$ is unbounded. As pointed out in the abstract, coding $X = \{x_1, \cdots, x_n\}$ by $Y = \{\bigoplus_{i \in I} x_i | I \subset \{1, \cdots, n\}\}$ works for arbitrary $X \subset \{0,1\}^p$. 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 $y \subset \{0,1\}^p$ and $\# y \leqq m$, then $l(G(E_p, Y)) \leqq \log m$.*

Proof. The proof is by induction on p. For $p = 1$, this is easily verified. Suppose the lemma holds for p. Let $Y \subset \{0,1\}^{p+1}$. For $i = 0, 1$, let $Y_i = \{y \in Y | y_{p+1} = i\}$ and $m_i = \# Y_i$. Then $l(G(E_{p+1}, Y_i)) \leqq \log m_i$ for $i = 0, 1$ by the induction hypothesis. Assume $m_0 \leqq m_1$. 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

$$l(G(E_{p+1}, y)) \leqq \max\{1 + l(G(E_{p+1}, Y_0)), l(G(E_{p+1}, Y_1))\}$$

$$\leqq \max\left\{1 + \log\frac{m}{2}, \log m\right\}. \qquad \square$$

COROLLARY 2. *Let $Y \subset \{0,1\}^p$, $\# Y = m$. For at least $p/2$ strings $e_i \in E_p$, there is a set $Y_i \subset Y$ such that $\# Y_i \leqq (2m \log m)/p$ 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

$$m \log m \geqq \sum_{y \in Y} \# L(y) = \sum_i \# \{y | e_i \in L(y)\}$$

$$> (p/2)(2m \log m)/p. \qquad \square$$

COROLLARY 3. *Let $Y \subset \{0,1\}^p$, $\# Y = m$, 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 $x, y, z \in \{0,1\}^*$. We say y and z 2-code x if $K(x|yz) = O(\log|x|)$. Let $X, Y \subset \{0,1\}^*$. 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 $Y' \subset Y$ with $\# Y' \leqq k$, the set of strings $Y \setminus Y'$ 2-codes X.

With $X, Y \subset \{0,1\}^*$, we associate again an edge-labelled graph $G(X, Y) = (Y, E, L)$: for each $y, z \in Y$ there is an edge $\{y, z\} \in E$ iff y and z 2-code some $x \in X$. For each edge $e = \{y, z\} \in E$, we set $L(e) = \{x \in X | y \text{ and } z \text{ 2-code } x\}$. Thus L is now a mapping from E into the power set of X. For $E' \subset E$, 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 $X = \{x_1, \cdots, x_n\} \subset \{0,1\}^p$, $Y = \{Y_1, \cdots, Y_{bn}\} \subset \{0,1\}^{ap}$ and $G(X, Y) = (Y, E, L)$. Let $K(x_1, \cdots, x_n) \geqq np$. Then*

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

Proof. Let $d = \# L(E)$ and let $L(E) = \{x_{i_1}, \cdots, x_{i_d}\}$. Then

(3) $dp - O(d \log n) \leqq K(x_{i_1} \cdots x_{i_d})$.

The string $x_{i_1} \cdots x_{i_d}$ 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 $K(x_{i_j}|y_k y_l) = O(\log p)$ and a program that produces x_{i_j} from $y_k y_l$.
- The bits of $y_1 \cdots y_{bn}$.

Thus

(4) $K(x_{i_1} \cdots x_{i_d}) \leqq O(d \log bn) + O(d \log p) + abnp.$

(3) and (4) imply the lemma. □

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 $\# L(E) \leqq \# Y$.

We now give an example of an edge-labelled graph G such that $\# L(E) \leqq \# Y$ 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_1 be a single edge with label x_1. For $i \geqq 1$, 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+1}. Call the resulting graph G_{i+1}. By induction on i, one easily verifies that $\# L(E) \leqq \# V - 1$ for any subgraph (V, E, L) of G_i.

Suppose G_8 is a subgraph of $G(X, Y)$. Consider any node y in G_8. Then $K(x_i|y) > 2p/3 - O(\log p)$ for some $i \in \{5, \cdots, 8\}$. Otherwise one gets the contradiction

$$\begin{aligned} 4p - O(\log p) &\leqq K(x_5 \cdots x_8) \leqq \sum_{i=5}^{8} (K(x_i|y) + O(\log p)) + K(y) \\ &\leqq \frac{11p}{3} + O(\log p). \end{aligned}$$

Consider in G_8 the subgraph drawn in Fig. 2. For all $j \in \{1, \cdots, 5\}$, we have

$$\begin{aligned} K(z_j|yx_i) &\leqq K(yx_iz_j) - K(yx_i) + O(\log p) \\ &\leqq K(yz_j) + K(x_i|yz_j) - K(yx_i) + O(\log p) \\ &\leqq K(y) + K(z_j|y) - K(y) - K(x_i|y) + O(\log p) \\ &\leqq p - \frac{2p}{3} + O(\log p). \end{aligned} \qquad \text{[ZL]}$$

This gives the contradiction

$$\begin{aligned} 4p - O(\log p) &\leqq K(x_1 \cdots x_4) \\ &\leqq K(yx_i) + \sum_{j=1}^{5} K(z_j|yx_i) + \sum_{j=1}^{4} K(x_j|z_jz_{j+1}) + O(\log p) \\ &\leqq \left(2 + \frac{5}{3}\right)p + O(\log p). \end{aligned}$$

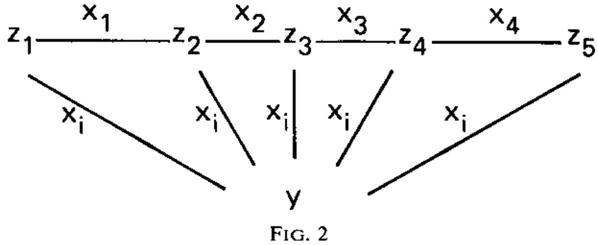

FIG. 2

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 $V'\subset V$, let $\# L((V'\otimes V')\cap E)\leqq \# V'$. Then $l(G)\leqq \alpha\sqrt{n}$ where $\alpha=2\sqrt{6}$.*

Proof. The proof is by induction on n. The theorem is true for $n\leqq\alpha$. Let $n>\alpha$. Find a node $u\in V$ such that $\# L((\{u\}\otimes V)\cap E)\geqq\alpha\sqrt{n}$ (if no such node exists, the nodes of G can be trivially labelled in the desired way). Let E_1 be a smallest set of edges adjacent to u such that $\# L(E_1)\geqq\alpha\sqrt{n}$. By hypothesis we have $\alpha\sqrt{n}-1\leqq\# E_1\leqq\alpha\sqrt{n}$. Let V_1 be the set of end points of edges in E_1 other than u.

Let $V_2=V\backslash(V_1\cup\{u\})$. 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

$$\alpha\sqrt{\# V_1}\leqq\alpha\sqrt{\alpha\sqrt{n}}\leqq\alpha\sqrt{n}-2$$

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

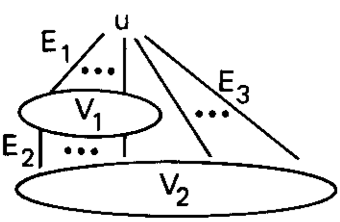

FIG. 3

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

$$\alpha\sqrt{\# V_2}\leqq\alpha\sqrt{n-\alpha\sqrt{n}+1}\leqq\alpha\left(\sqrt{n}-\frac{\alpha\sqrt{n}-1}{2\sqrt{n}}\right)$$

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

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 endpoint 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 evey node $w\in V_2$, we have*

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

Proof. Assume the lemma is false for node w. Let $V_3\subset V_1$ be a smallest set of nodes such that $\# L((w\otimes V_3)\cap E)\geqq 10$ (Fig. 4). We make three observations:

(i) $\# V_3\geqq 9$.

(ii) Let $V_4\subset V_3$ and $z\in V_3\backslash V_4$. Then

$$L(\{z, u\})\backslash L(\{V_4\otimes u\})\neq\varnothing,$$

$$L(\{z, w\})\backslash L(\{V_4\otimes w\})\neq\varnothing$$

by the minimality of V_1 and V_3.

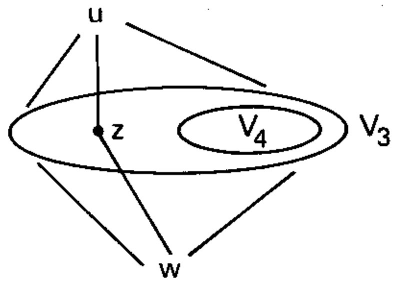

FIG. 4

(iii) Let $V_4 \subset V_3$, $\# V_4 \leqq 2$. Then $\# L(u \otimes V_4) \leqq 3$ and $\# L(w \otimes V_4) \leqq 3$. 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

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

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

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

COROLLARY 4. *Let* $X = \{x_1, \cdots, x_n\} \subset \{0, 1\}^p$, $Y = \{y_1, \cdots, y_{bn}\} \subset \{0, 1\}^p$, *let* $K(x_1 \cdots x_n) \geqq np$, *let* $(1 - O(\log p \# Y)/p) \leqq 1 + 1/\# Y$, *and suppose* Y *2-codes* X *k-robustly. Then* $4\sqrt{6\# Y} > k$.

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

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

For every node $v \in V$ and any edge label l that occurs on at least $n^\varepsilon + 1$ of the edges adjacent to v, put label l 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 $E_v^1, \cdots, E_v^\tau$ such that in every class E_v^i, every label occurs on at most one edge of E_v^i. Partition E into classes $E^{i,j}$, $1 \leqq i \leqq j \leqq n^\varepsilon$, by $\{u, v\} \in E^{i,j}$ if $[u, v] \in E_v^i \cap E_u^j$. For all i, j, let $G^{i,j} = (V, E^{i,j}, L^{i,j})$ where $L^{i,j}$ is L restricted to $E^{i,j}$. Then in $G^{i,j}$ for every node v, all edges adjacent to v have different labels. We will show $l(G^{i,j}) \leqq n^{1-3\varepsilon}$.

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 $G^{i,j}$. Continue this process as long as possible. If finally all of $G^{i,j}$ 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 μ that consists of ones only. Then $z(m, n, j, k) \leqq 1 + km + (j-1)^{1/k} m^{1-1/k} n$. In particular,

$$z(n, n, 9c^3, 2c) \leqq 1 + 2cn + (9c^3)^{1/(2c)} n^{2-1/(2c)} \leqq n^{2-6\varepsilon}$$

if $6\varepsilon < 1/(2c)$ and n is large enough.

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

If R is a set of at most $2c$ rows of μ, then at most $4c^2$ different labels occur in R. Each label occurs in at most $2c$ more rows of μ. 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 μ that contains $4c^2 + 2c$ different labels. The rows and columns of μ' correspond to a set V' of $4c+1$ vertices of μ. Thus

$$2c(2c+1) \leqq \# L((V' \otimes V') \cap E) \leqq c(4c+1). \qquad \square$$

COROLLARY 5. *Let* $X = \{x_1, \cdots, x_n\} \subset \{0,1\}^p$, $Y = \{Y_1, \cdots, Y_{bn}\} \subset \{0,1\}^{ap}$ *and* $p \gg \log bnp$. *Suppose* $K(x_1 \cdots x_n) \geqq pn$ *and* Y *2-codes* X *k-robustly. Then* $16a(bn)^{1-\varepsilon} > k$ *for* $\varepsilon < 1/(24a)$ *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 3. *For* $c \geqq 2$, *there exists* $G = (V, E, L)$, $\# V = n$, *having an edge labelling satisfying*

$$\# L((V' \otimes V') \cap E) \leqq c \# V' \quad \textit{for all } V' \subset V$$

and

$$l(G) \geqq c' n^a \quad \textit{where } a < \frac{1}{2} - \frac{1}{4c-2}.$$

Proof. Let $\delta = (\frac{1}{2} - 1/(4c-2) - a)/2$. 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,

$$(**) \qquad \#(V' \otimes V') \cap E \leqq c \# V' \quad \text{for all } V' \subset V \quad \text{with } \# V' \leqq n^\alpha.$$

The probability that for any set V' of cardinality $j \leqq n^\alpha$, $(**)$ does not hold, is at most

$$\begin{aligned} W_j &= \binom{n}{j} \binom{\binom{j}{2}}{cj} \binom{\binom{n}{2} - cj}{n^{1+\alpha} - cj} \Big/ \binom{\binom{n}{2}}{n^{1+\alpha}} \\ &\leqq \left(\frac{ne}{j}\right)^j \left(\frac{ej^2}{2cj}\right)^{cj} \frac{n^{1+\alpha} \cdots (n^{1+\alpha} - cj + 1)}{\binom{n}{2} \cdots \left(\binom{n}{2} - cj + 1\right)} \\ &\leqq \left(\frac{ne}{j}\right)^j \left(\frac{ej}{2c}\right)^{cj} (2.1 n^{-1+\alpha})^{cj} \\ &\leqq (c_2 j^{1-1/c} n^{-1+\alpha+1/c})^{cj}. \end{aligned}$$

For $2c+2 \leqq j \leqq \log n$, we estimate

$$W_j \leqq (\log^2 n \cdot n^{-1/2-1/(4c-2)-\delta+1/c})^{cj} \leqq (n^{-1/2-1/6+1/2})^{2.6} = n^{-2}.$$

For $\log n < j \leqq n^\alpha$, we have

$$\begin{aligned} W_j &\leqq (c_2 n^{\alpha(1-1/c)-1+\alpha+1/c})^{cj} \\ &= (c_2 n^{(\alpha(2c-1)-c+1)/c})^{cj} \\ &= (c_2 n^{-\delta(2c-1)/c})^{cj} \leqq (n^{-c_3})^{\log n} \leqq n^{-2}. \end{aligned}$$

Hence the probability that (**) does not hold is at most

$$\sum_{j=2c+2}^{n^\alpha} W_j \leqq n^\alpha n^{-2} \leqq n^{-1}.$$

Next we make use of the fact that with probability $1-o(n^{-1})$, the degree of every node in G is bounded by $3n^\alpha$ [ER]. Therefore there exists a graph G with n nodes, $n^{1+\alpha}$ edges, such that the degree of every node in G is bounded by $3n^\alpha$ and (**) holds for G.

Let L be any edge labelling of G, which labels every edge with exactly 1 label $l \in \{1, \cdots, n^\alpha\}$. Let $V' \subset V$. Then $\# L((V' \otimes V') \cap E) \leqq \min\{n^\alpha, \#((V' \otimes V') \cap E)\} \leqq c \# V'$.

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

$$\binom{d}{j}\left(\frac{1}{n^\alpha}\right)^j \leqq \left(\frac{de}{j}\right)^j\left(\frac{1}{n^\alpha}\right)^j \leqq \left(\frac{3n^\alpha e}{j}\right)^j\left(\frac{1}{n^\alpha}\right)^j = \left(\frac{3e}{j}\right)^j = O(n^{-3})$$

if $j \geqq \log n$. Therefore the probability that $\log n$ or more edges adjacent to the same node as G have the same label is at most $n \cdot n^\alpha \cdot O(n^{-3}) = O(n^{-1})$. 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 $\sum_v \# L'(v) \geqq n^{1+\alpha}/\log n$. 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 up to $m \log m$ 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 $X \subset \{0, 1\}^p$ and $1 \leqq i \leqq p$, define $r(X, i)$ as the number of edges in $G(E_p, X)$ with label i.

For $1 \leqq k \leqq p$ define

$$r_k(X) = \min_{\substack{D \subset \{1, \cdots, p\} \\ |D| \geqq k}} \sum_{i \in D} r(X, i)$$

and

$$\rho_k(X) = \min_{\substack{D \subset \{1, \cdots, p\} \\ |D| \geqq k}} \max_{i \in D} r(X, i),$$

and for $m \in IN$,

$$r_k(m) = \max_{|X| \leqq m} r_k(X)$$

and

$$\rho_k(m) = \max_{|X| \leqq m} \rho_k(X).$$

One checks easily that for $m \leqq 2^p$, $\rho_p(m) = \lfloor m/2 \rfloor$, whereas from Lemma 2 it follows that $r_p(m) = \theta(m \log m)$.

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

THEOREM 4. *There are constants $\alpha > 0$ and $C, h \geqq 1$ such that for all $1 \leqq k < p$ and $m/(p-k) \geqq C/\alpha$,*

$$\rho_k(m) \leqq \alpha \frac{m}{p-k} \ln^3\left(\alpha \frac{m}{p-k} + h\right).$$

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 $\beta > 0$, $h \geqq 1$ such that for any $X \subset \{0, 1\}^p$,*

$$\sum_{i=1}^{p} \frac{r(X, i)}{\ln^3(r(X, i) + h)} \leqq \beta |X|.$$

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 $i \in \{1, \cdots, p\}$. Define

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

Later it will be shown that for appropriate $h \geqq e^2$, $F(x)$ is monotonically increasing for $x \geqq 0$. Hence,

$$\sum_{i=1}^{p} F(r_i) \geqq \sum_i \sum_{r_i > r} F(r_i) > (p-k) F(r)$$

$$= \alpha m \frac{\ln^3(\alpha(m/(p-k)) + h)}{\ln^3(\alpha(m/(p-k)) \ln^3(\alpha(m/(p-k)) + h) + h)}.$$

For an appropriate $C \geqq 1$,

$$x + h \geqq \ln^3(x+h)$$

holds for all $x \geqq C$. Thus if $\alpha m/(p-k) \geqq C$, then

$$\ln^3\left(\frac{\alpha m}{p-k} \ln^3\left(\frac{\alpha m}{p-k} + h\right) + h\right) \leqq \ln^3\left(\frac{\alpha m}{p-k}\left(\frac{\alpha m}{p-k} + h\right) + h\right)$$

$$\leqq \ln^3\left(\left(\frac{\alpha m}{p-k} + h\right)^2\right) = 8 \ln^3\left(\frac{\alpha m}{p-k} + h\right).$$

Therefore, $\sum_{i=1}^{p} F(r_i) > \alpha m/8$. But this contradicts Lemma 5 if $\alpha/8 \geqq \beta$. $\square$

Proof of Lemma 5. Define $h = e^2 \approx 7.389$ and $\gamma = 0.16$.

$$g(x) = \gamma \frac{x}{\ln^3 (x+h)}, \quad \text{for } x \geqq 0$$

and

$$f(n) = 1 + \sum_{m=2}^{n} \frac{1}{m \ln^2 (m)}, \quad \text{for } n \in \mathbb{N}, \quad n \geqq 1.$$

We will show in the Appendix:

(g1) $\qquad 0 \leqq g(x) \leqq 0.16x \qquad$ for all $x \geqq 0$,

(g2) $\qquad g'(x) \geqq 0 \qquad$ for all $x \geqq 0$,

(g3) $\qquad g''(x) \leqq 0 \qquad$ for all $x \geqq 0$,

(f1) $\qquad 1 \leqq f(n) \leqq f(n+1) \leqq 5 \qquad$ for all $n \geqq 1$.

Now let $X \subset \{0, 1\}^p$. Lemma 5 follows from the following:

PROPOSITION. *If* $n = |\{i | r_i > 0\}|$, *then*

$$\sum_{i=1}^{p} g(r_i) \leqq f(n)|X|.$$

Proof. The proof is by induction on n. Define $r = \max_{1 \leqq i \leqq p} r_i$. For each i, the edges with label i are a matching. Hence, $|X| \geqq 2r$. For all $1 \leqq h \leqq n_0 = 98$, we get

$$\sum_{i=1}^{p} g(r_i) \leqq ng(r) \leqq 98\gamma \frac{r}{\ln^3 (r+h)} \leqq \frac{49\gamma|X|}{\ln^3 (r+h)} \leqq \frac{49\gamma}{2^3}|X| \leqq |X| \leqq f(n)|X|.$$

Thus, the claim holds for all $n \leqq n_0$. Now assume

(7.1) $$n+1 > n_0 = 98,$$

and the claim is true for all $n' \leqq n_0$. We may assume that $r_1 \geqq r_2 \geqq \cdots \geqq r_{n+1} > r_{n+2} = \cdots = r_p = 0$. Define for $l \in \{0, 1\}$,

$$X^l = \{x \in X | x_{n+1} = l\},$$

and for $1 \leqq i \leqq n$ r_i^l as the number of edges in $G(E_p, X^l)$ with label i, this means we cut X in dimension $n+1$. Obviously,

(7.2) $$X = X^0 \cup X^1, \qquad r_i = r_i^0 + r_i^1 \quad \text{for } 1 \leqq i \leqq n.$$

$D = D^0 \cap D^1$ and $d^l = |D^l|$. One can check easily

(7.3) $$|X^l| \geqq \max\{r_{n+1}, d^l + 1\} \quad \text{for } l = 0, 1.$$

Define $\Delta g(x, y) = g(x) + g(y) - g(x+y)$. Now

$$\sum_{i=1}^{p} g(r_i) = \sum_{i=1}^{n+1} g(r_i) = \sum_{i=1}^{n} g(r_1^0) + \sum_{i=1}^{n} g(r_i^1) - \sum_{i \in D} \Delta g(r_i^0, r_i^1) + g(r_{n+1}).$$

Applying the induction hypothesis to X^0 and X^1 gives

(7.4) $$\sum_{i=1}^{p} g(r_i) \leqq |X^0| f(d^0) + |X^1| f(d^1) - \sum_{i \in D} \Delta g(r_i^0, r_i^1) + g(r_{n+1}).$$

The idea of the proof is as follows: if D is large, then $\sum_{i \in D} \Delta g(r_0^0, r_i^1)$ is large enough to compensate the term $g(r_{n+1})$; otherwise one of the d^l must be relatively small, such that the difference between $|X^l|f(n+1)$ and $|X^l|f(d^l)$ 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) $\quad \Delta g(x, y) \geqq 0 \quad$ for all $x, y \geqq 0$,

(g5) $\quad \Delta g(x, y) \leqq \Delta g(x, z) \quad$ for all $0 \leqq x$ and $0 \leqq y \leqq z$,

(g6) $\quad \Delta g(1, 1) \geqq 0.0298\gamma$,

(g7) $\quad \Delta g(x, y) \geqq 1.4 \dfrac{g(x)}{\ln(x+h)} \quad$ for all $0 \leqq x \leqq y$ and $y \geqq 3h$.

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

(f2) $$\delta f(n, m) \geqq \frac{1}{4} \frac{1}{\ln^2(m+h)} \quad \text{for all } 16 \leqq m \leqq \frac{2}{3}n.$$

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

$$\sum_{i=1}^{p} g(r_i) \leqq |X^0| f(d^0) + |X^1| f(d^1) + g(r_{n+1})$$
$$\leqq (|X^0| + |X^1|) f(n+1) + g(r_{n+1}) - |X^1|\, \delta f(n+1, d^1).$$

If $d^1 \geqq 16$ and $d^1 \geqq r_{n+1}$, we get

$$\begin{aligned} g(r_{n+1}) - |X^1|\, \delta f(n+1, d^1) &\leqq g(r_{n+1}) - d^1 \frac{1}{4} \frac{1}{\ln^2(d^1+h)} && \text{by (7.3) and (f2)} \\ &\leqq g(r_{n+1}) - \gamma \frac{d^1}{\ln^3(d^1+h)} && \text{since } \frac{1}{4} \geqq \frac{\gamma}{\ln(d^1+h)} \\ &= g(r_{n+1}) - g(d^1) \leqq 0 && \text{by (g2).} \end{aligned}$$

If $15 \geqq d^1 \geqq r_{n+1}$, we get

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

If $16 \leqq d^1 \leqq r_{n+1}$, we have

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

If $1 \leqq d^1 \leqq 15$ and $d^1 \leqq r_{n+1}$, we have

$$\begin{aligned} g(r_{n+1}) - |X^1|\, \delta f(n+1, d^1) &\leqq \gamma \frac{r_{n+1}}{\ln^3(r_{n+1}+h)} - r_{n+1} \frac{1}{(d^1+1)\ln^2(d^2+1)} \\ &\leqq \frac{r_{n+1}}{\ln^2(d^1+1)} \left(\frac{\gamma}{\ln(d^1+1)} - \frac{1}{d^1+1} \right) < 0, \end{aligned}$$

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

$$\begin{aligned} g(r_{n+1}) - |X^1|\, \delta f(n+1, d^1) &\leqq \gamma \frac{r_{n+1}}{\ln^3(r_{n+1}+h)} - r_{n+1} \frac{1}{2 \ln^2 2} \\ &\leqq r_{n+1} \left(\frac{\gamma}{8} - \frac{1}{2 \ln^2 2} \right) < 0. \end{aligned}$$

Thus, $\sum_{i=1}^{p} g(r_i) \leqq |X| f(n+1)$. We now assume

$$d^l \geqq \tfrac{2}{3}n \quad \text{for } l=0,1. \tag{7.5}$$

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

$$c_1 n \ln^3(n+h) \geqq n. \tag{7.6}$$

This implies

$$g(r_{n+1}) \leqq g(c_1 n \ln^3(n+h)) = \gamma \frac{c_1 n \ln^3(n+h)}{\ln^3(c_1 n \ln^3(n+h)+h)} \leqq \gamma c_1 n.$$

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

$$\begin{aligned}
\sum_{i=1}^{p} g(r_i) &\leqq |X^0| f(d^0) + |X^1| f(d^1) - \sum_{i \in D} \Delta g(r_i^0, r_i^1) + g(r_{n+1}) \\
&\leqq |X| f(n+1) - \sum_{i \in D} \Delta g(1,1) + g(r_{n+1}) \quad \text{by (g5)} \\
&\leqq |X| f(n+1) - \frac{n}{3} 0.0298\gamma + \gamma c_1 n \qquad \text{by (g6)} \\
&\leqq |X| f(n+1) \qquad \text{since } \frac{0.0298}{3} \geqq c_1.
\end{aligned}$$

Let us now assume

$$r_{n+1} \geqq c_1 n \ln^3(n+h). \tag{7.7}$$

From (7.1) it follows that

$$r_{n+1} \geqq n \geqq n_0 \geqq 98 \geqq 6h. \tag{7.8}$$

For $1 \leqq i \leqq h$, define $z_i = \min\{r_i^0, r_i^1\}$ and $v_i = \max\{r_i^0, r_i^1\}$. We have

$$v_i \geqq \frac{r_i}{2} \geqq \frac{r_{n+1}}{2} \geqq 3h. \tag{7.9}$$

Case 3.

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

Then

$$\begin{aligned}
\sum_{i \in D} g(r_i^0, r_i^1) &= \sum \Delta g(z_i, v_i) \\
&\geqq \sum 1.4 \frac{g(z_i)}{\ln(z_i+h)} \qquad \text{by (7.9) and (g7)} \\
&\geqq 1.4 \frac{\sum g(z_i)}{\ln(\sum g(z_i)+h)} \qquad \geqq 1.4 \frac{(1/8)\ln^2(r_{n+1}/(r_{n+1}+h))}{\ln((1/8)(r_{n+1}/\ln^2(r_{n+1}+h))+h)} \\
&\geqq 0.175 \frac{r_{n+1}}{\ln^3(r_{n+1}+h)} \qquad \geqq g(r_{n+1}), \quad \text{since } 0.175 \geqq \gamma.
\end{aligned}$$

Hence in (7.4),

$$\sum_{i=1}^{p} g(r_i) \leqq (|X^0|+|X^1|)f(n+1)+g(r_{n+1})-\sum_{i\in 0}\Delta g(r_i^0, r_i^1) \leqq |X|f(n+1).$$

It remains the case that

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

Define for $l=0,1$, $B^l=\{i|r_i^l>r_i^{1-l}\}$ and $b^l=|B^l|$. Since $b^0+b^1\leqq n$, we may assume $b^1 \leqq n/2$.

If we remove from $G(E_p, X^1)$ edges with labels not in B^1, the remaining graph consists of some connected components $G(E_p, Y^1), \cdots, G(E_p, Y^u)$ where $\bigcup_{1\leqq j\leqq u} Y^j = X^1$. Let us denote by y_i^j the number of edges in $G(E_p, Y^j)$ with label i. Each such graph contains only labels from B^1. Hence by the induction hypothesis,

$$\sum_{i=1}^{p} g(y_i^j) \leqq |Y^j| f\left(\frac{n}{2}\right)$$

and

$$\sum_{i\in B^1} g(r_i^1) \leqq \sum_{i\in B^1}\sum_{j=1}^{u} g(y_i^j) \quad \text{since } r_i^1=\sum_{j=1}^{u} y_i^j$$

$$\leqq \sum_{j=1}^{u} |Y^j| f\left(\frac{n}{2}\right) = |X^1| f\left(\frac{n}{2}\right).$$

Thus we can conclude

$$\sum_{j=1}^{p} g(r_i) \leqq \sum_{i=1}^{n} g(r_i^0)+\sum_{i\in B^1} g(r_i^1)+\sum_{i\notin B^1} g(r_i^1)+g(r_{n+1})$$

$$\leqq |X^0|f(n+1)+|X^1|f(n+1)-|X^1|\,\delta f\left(n+1,\frac{n}{2}\right)$$

$$+\sum_{i=1}^{n} g(z_i)+g(r_{n+1}) \qquad \text{since } r_i^1=z_i \text{ for } i\notin B^1$$

$$\leqq |X|f(n+1)-\frac{1}{4}\frac{r_{n+1}}{\ln^2(n/2+h)}+\frac{1}{8}\frac{r_{n+1}}{\ln^2(r_{n+1}+h)}$$

$$+\gamma\frac{r_{n+1}}{\ln^3(r_{n+1}+h)} \qquad \text{by (f2)}$$

$$=|X|f(n+1)-\frac{r_{n+1}}{\ln^2(r_{n+1}+h)}\left[\frac{1}{4}-\frac{1}{8}-\frac{\gamma}{\ln(r_{n+1}-h)}\right]$$

$$\leqq |X|f(n+1).$$

This completes the proof of the Proposition and Theorem 4. □

For $Y\subset\{0,1\}^p$ and $Q\subset\{1,\cdots,p\}$, let $G^Q(E_p, Y)$ denote the subgraph of $G(E_p, Y)$ 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 $Y\subset\{0,1\}^p$ of size at most μp, one can find a set $Q\subset\{1,\cdots,p\}$ of size at least $(1-\varepsilon)p$ such that in $G^Q(E_p, Y)$ the occurrence of each label is bounded by $A(\varepsilon, \mu)$, and hence $G^Q(E_p, Y)$ 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

$$l_k(Y) = \min_{\substack{Q\subset\{1,\cdots,p\}\\ |Q|\geqq k}} \quad \min_{\substack{\text{transformation } L\\ \text{for } G^Q(E_p,Y)}} \quad \max_{v\in G^Q(E_p,Y)} \# L(v)$$

and

$$l_k(m) = \max_{|Y|\leqq m} l_k(Y).$$

Obviously, for $n-(\log p)/2\leqq k\leqq n$, it holds that $l_k(p)=\theta(\log p)$.

THEOREM 5. *For any $\varepsilon, \mu>0$ there exists a constant $R(\varepsilon,\mu)$ such that*

$$l_{(1-\varepsilon)p}(\mu p)\leqq R(\varepsilon,\mu), \quad \textit{for any } p.$$

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

Let $R=R(\varepsilon,\mu)>10A/\varepsilon g(1)$. If the theorem is false, then there exists $p\in\mathbb{N}$ and $Y\subset\{0,1\}^p$, $|Y|\leqq\mu p$ such that for any $Q\subset\{1,\cdots,p\}$ 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 $U\subset\{1,\cdots,p\}$ 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 $\tilde{L}$ of L to a graph $G^Q(Ep, Y)$ where Q is a subset of U of size $(1-\varepsilon)p$, $F(L)$ and $F(\tilde{L})$ 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 $Z\subset Y$ be the set of all nodes from which there is a path of length $\geqq 0$ 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 $z=z_0, z_1, \cdots, z_l$ be a path in $\bar{H}$ from z to a node z_l with indegree bigger than R. By definition of Z, such a path must exist.

Change L into $\bar{L}$ by assigning for $0\leqq i<l$ 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 $\#\bar{L}(z_0)\leqq R$, $\#\bar{L}(z_l)=\# L(z)-1\geqq R$ and $\#\bar{L}(z)=\# L(z)$ for all remaining $z\in Y$. Hence

$$F(L)>F(\bar{L}),$$

which contradicts the minimality of L. $\square$

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 $R|Z|\leqq Ap$. Hence

$$|Z|\leqq\frac{A}{R}p.$$

On the other hand, $G(E_p, Z)$ must have at least $\varepsilon p/2$ different labels; otherwise, deleting this set of labels from U would yield a subset Q of $\{1, \cdots, p\}$ of size at least $(1-\varepsilon)p$ 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

$$|Z| \geqq \frac{1}{f(n)} \sum_{i=1}^{p} g(r_i),$$

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

$$|Z| \geqq \frac{1}{5} \cdot \frac{\varepsilon}{2} p \cdot g(1) = \frac{\varepsilon}{10} g(1) p.$$

Combining the two inequalities for $|Z|$ gives

$$\frac{\varepsilon}{10} g(1) \leqq \frac{A}{R}.$$

Hence

$$R \leqq \frac{10A}{\varepsilon g(1)}.$$

This contradicts the definition of R. $\square$

COROLLARY 7. *If $Y \subset \{0, 1\}^p$, $\# Y = O(p)$ 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 $r \geqq 3$. 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 $l(G(x, y))$?

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

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

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

$$g'(x) = \gamma \frac{\ln^3 (x+h) - x3 \ln^2 (x+h)/(x+h)}{\ln^6 (x+h)} = \gamma \frac{1}{\ln^3 (x+h)} \left[1 - \frac{3x}{(x+h) \ln (x+h)} \right].$$

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

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

$$\begin{aligned} g''(x) &= \gamma \Bigg(\frac{-3}{\ln^4 (x+h)} \frac{1}{x+h} \left[1 - \frac{3x}{(x+h) \ln (x+h)} \right] \\ &\qquad - \frac{1}{\ln^3 (x+h)} \left[\frac{3(x+h) \ln (x+h) - 3x(\ln (x+h) + 1)}{(x+h)^2 \ln^2 (x+h)} \right] \Bigg) \\ &= -\gamma \frac{3}{\ln^5 (x+h)(x+h)^2} [(x+h) \ln (x+h) - 3x + h \ln (x+h) - x] \\ &= -\gamma \frac{3}{\ln^5 (x+h)(x+h)^2} [(x+2h) \ln (x+h) - 4x]. \end{aligned}$$

Let $\varphi(x) := (x+2h)\ln(x+h) - 4x$. Then for $x \geqq 0$, $g''(x) \leqq 0 \Leftrightarrow \varphi(x) \geqq 0$,

$$\varphi'(x) = \ln(x+h) + \frac{x+2h}{x+h} - 4,$$

$$\varphi''(x) = \frac{1}{x+h} + \frac{(x+h)-(x+2h)}{(x+h)^2} = \frac{x}{(x+h)^2}.$$

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

(g3) $$g''(x) \leqq 0 \quad \text{for all } x \geqq 0.$$

Define $\Delta g(x, y) = g(x) + g(y) - g(x+y)$. Calculation proves (g6):

$$\Delta g(1,1) = 2\gamma \left[\frac{1}{\ln^3(1+h)} - \frac{1}{\ln^3(2+h)}\right] \geqq 0.0298\gamma.$$

Assume $0 \leqq x$ and $0 \leqq y \leqq z$. Since for all $t \in [y, z]$, $g'(x+t) \leqq g'(t)$ by (g3), we can conclude that $g(x+z) - g(x+y) \leqq g(z) - g(y)$. This yields $g(x) + g(y) - g(x+y) \leqq g(x) + g(z) - g(x+z)$, or

(g5) $$\Delta g(x, y) \leqq \Delta g(x, z) \quad \text{for all } 0 \leqq x \text{ and } 0 \leqq y \leqq z.$$

For Δg we can show the bound for $0 \leqq x \leqq y$:

$$\Delta g(x, y) = g(x) + g(y) - g(x+y) \geqq g(x) - x \sup_{z \in [y, x+y]} g'(z) = g(x) - xg'(y).$$

This yields

$$\Delta g(x, y) \geqq g(x) - x\gamma \frac{1}{\ln^3(y+h)}\left(1 - \frac{3y}{(y+h)\ln(y+h)}\right)$$

$$= g(x)\left[1 - \frac{\ln^3(x+h)}{\ln^3(y+h)}\left(1 - \frac{3y}{(y+h)\ln(y+h)}\right)\right].$$

Since $\ln(x+h) \leqq \ln(y+h)$ and $0 \leqq 3y \leqq (y+h)\ln(y+h)$ (see proof of (g2)), $\Delta g(x, y) \geqq 0$ follows from $g(x) \geqq 0$. The case $x > y$ follows from $\Delta g(x, y) = \Delta g(y, x)$. This proves (g4). If $x + h \geqq (y+h)^{2/3}$, we get $\ln(x+h) \geqq (2/3)\ln(y+h)$ and

$$\Delta g(x, y) \geqq g(x)\frac{3y}{y+h}\frac{1}{\ln(y+h)} \geqq g(x)\frac{3y}{y+h}\frac{2/3}{\ln(x+h)} = \frac{2}{3}\frac{3y}{y+h}\frac{g(x)}{\ln(x+h)}.$$

If $y \geqq 3h$ then $\Delta g(x, y) \geqq \frac{3}{2} g(x)/\ln(x+h)$. If on the other hand $x + h \leqq (y+h)^{2/3}$, we can bound $\Delta g(x, y)$ by

$$\Delta g(x, y) \geqq g(x)\left[1 - \frac{\ln^3(x+h)}{\ln^3(y+h)}\right]$$

$$\geqq g(x)\left[1 - \left(\frac{2}{3}\right)^3\right]$$

$$\geqq 0.7 g(x)$$

$$\geqq 1.4 \frac{g(x)}{\ln(x+h)} \quad \text{since } \ln(x+h) \geqq 2.$$

Therefore we have shown (g7):

$$\Delta g(x, y) \geqq 1.4 \frac{g(x)}{\ln(x+h)}, \quad \text{for all } 0 \leqq x \leqq y \text{ and } y \geqq 3h.$$

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

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

Then

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

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

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

Thus (f1), $1 \leqq f(n) \leqq 5$, holds for all $n \geqq 1$. Define $\delta f(n, m) = f(n) - f(m)$ for $1 \leqq m \leqq n$. For $16 \leqq m \leqq 2/3n$,

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

$$\geqq \lceil m/2 \rceil \frac{1}{\lceil 3m/2 \rceil \ln^2 \lceil 3m/2 \rceil} \geqq \frac{1}{3} \frac{1}{\ln^2 \lceil 3m/2 \rceil}.$$

Since $m \geqq 16$,

$$\lceil 3m/2 \rceil \leqq \left(\frac{3}{2} + \frac{1}{20}\right) m \leqq 1.6m \leqq (16+h)^{0.15} m \leqq (m+h)^{1.15} \leqq (m+h)^{\sqrt{4/3}}.$$

Hence

$$\ln^2 \left\lceil \frac{3m}{2} \right\rceil \leqq \ln^2 (m+h)^{\sqrt{4/3}} = \frac{4}{3} \ln^2 (m+h).$$

This proves

(f2) $$\delta f(n, m) \geqq \frac{1}{4} \frac{1}{\ln^2 (m+h)} \quad \text{for all } 16 \leqq m \leqq \frac{2}{3} n.$$

REFERENCES

[H] C. HYLTEN-CAVALLIUS, *On a combinatorial problem*, Colloq. Math., 6 (1958), pp. 59–65.

[P] W. J. PAUL, *On heads versus tapes*, Proc. 22nd Symposium on Foundations of Computer Science, 1981.

[ZL] A. ZVONKIN AND L. LEVIN, *The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms*, Russian Math. Surveys, 6, 1970.

[ER] P. ERDÖS AND A. RENYI, *On the evolution of random graphs*, in P. Erdös, The Art of Counting, MIT Press, Cambridge, MA, 1973.