1 Problem Definition

Suppose that we are given a set of bit vectors. The only type of operation we can perform on these vectors is replacing an existing entry in some bit vector with the character `x`. We can use this to "anonymize" the data by crossing out any coordinate where the bit vectors differ:

Definition 1. Given a set S of bit vectors of length d, we say that the coordinate i is *anonymized* if all bit vectors in S have the same value at coordinate i. The *cost of anonymizing* the set S is equal to the product of $|S|$ and the number of non-anonymized coordinates.

The cost of anonymizing a set of bit vectors is easy to compute. However, the problem becomes much harder when we wish to divide our bit vectors into groups and then anonymize each group individually. This problem is known as the *k-distinct problem.* Formally, it is defined as follows:

Definition 2. Given a set S of bit vectors of length d, we wish to partition S into k sets $S_1, S_2, \ldots, S_k$ so that we minimize:

$$\sum_i \text{cost of anonymizing } S_i.$$

2 Proof of Hardness

Our NP-hardness proof relies on the following variant of the k-distinct problem:

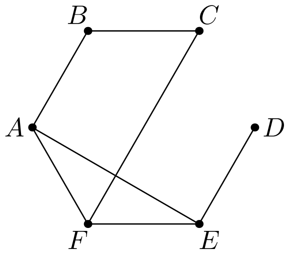
Definition 3. Given a set S of $2n$ bit vectors, the goal of the *balanced 2-distinct problem* is to find the following:

$$\min_{\substack{\text{partitions } (S_1, S_2) \\ |S_1|=|S_2|=n}} \left(\sum_i \text{cost of anonymizing } S_i \right).$$

In section 2.1, we show that it is NP-hard to solve the balanced 2-distinct problem. In section 2.2, we then use that hardness result to show that the original k-distinct problem is also NP-hard.

2.1 Hardness of the Balanced 2-Distinct Problem

We will show that the balanced 2-distinct problem is NP-hard by a reduction from the sparsest cut problem, which is known to be NP-hard:



(a) A graph G.

$$\begin{array}{ccl}
A & = & \langle 1, \; 0, \; 0, \; 0, \; 1, \; 0, \; 1 \rangle \\
B & = & \langle 1, \; 1, \; 0, \; 0, \; 0, \; 0, \; 0 \rangle \\
C & = & \langle 0, \; 1, \; 0, \; 0, \; 0, \; 1, \; 0 \rangle \\
D & = & \langle 0, \; 0, \; 1, \; 0, \; 0, \; 0, \; 0 \rangle \\
E & = & \langle 0, \; 0, \; 1, \; 1, \; 0, \; 0, \; 1 \rangle \\
F & = & \langle 0, \; 0, \; 0, \; 1, \; 1, \; 1, \; 0 \rangle
\end{array}$$

(b) The corresponding vectors.

Figure 1: An example reduction from the balanced 2-distinct problem to the sparsest cut problem.

Definition 4. Given a graph $G = (V, E)$, we say that a cut $(S, V - S)$ is *perfectly balanced* if $|S| = |V - S| = |V|/2$. The *sparsest cut* of G is the perfectly balanced cut[1] that minimizes the number of edges crossing the cut.

Our reduction is fairly straightforward. Suppose that we are given a graph G, and we wish to find the sparsest cut. Let $V = \{v_1, \ldots, v_{2n}\}$ be the list of vertices of G, and let $E = \{e_1, \ldots, e_m\}$ be the list of edges of G. We construct one bit vector $\mathbf{b}_i$ for each vertex v_i, and each bit vector has length $|E|$. The jth coordinate of the bit vector $\mathbf{b}_i$ is defined to be 1 if the vertex v_i is incident to the edge e_j, and 0 otherwise. See Figure 1 for an example.

The runtime of this reduction is clearly polynomial. To show that the reduction is correct, we must show a relationship between the sparsest cut and the optimal solution to this balanced 2-distinct problem.

Theorem 1. *A graph with $2n$ vertices and m edges has a sparsest cut of size t if and only if the cost of anonymizing the corresponding balanced 2-distinct problem is equal to $(m + t) \cdot n$*

Proof. Let $G = (V, E)$ be a graph, and let $(C, V - C)$ be a perfectly balanced cut of G with cost t. Let $S_1 = \{\mathbf{b}_i \mid v_i \in C\}$, and let $S_2 = \{\mathbf{b}_i \mid v_i \notin C\}$. Clearly, we have $|S_1| = |S_2| = n$. What is the cost of anonymizing these sets of vectors?

The number of coordinates of S_1 that are not anonymized is equal to the number of coordinates such that at least one vector in S_1 contains a 1

[1] Most definitions of sparsest cut actually require a weaker constraint: that both $|S|$ and $|V - S|$ must be at least $|V|/3$ (in other words, the cut is required to be *balanced*, instead of perfectly balanced). But by adding $|V|/3$ nodes that aren't connected to anything, it's easy to reduce the definition with balanced cuts to the definition with perfectly balanced cuts, which makes the problem NP-hard either way.

in that position. Our reduction ensures that the vector $\mathbf{b}_i$ contains a 1 in coordinate j if and only if v_i is one of the endpoints of e_j. Formally:

$$\begin{aligned}\text{cost of anonymizing } S_1 &= n \cdot |\{j \mid \exists \mathbf{b}_i \in S_1 : \mathbf{b}_i[j] = 1\}| \\ &= n \cdot |\{e_j = (v_k, v_\ell) \mid v_k \in C \vee v_\ell \in C\}|\end{aligned}$$

Similar analysis shows that:

$$\text{cost of anonymizing } S_2 = n \cdot |\{e_j = (v_k, v_\ell) \mid v_k \notin C \vee v_\ell \notin C\}|$$

The coordinate j will be anonymized in both sets if any only if one of $v_k, v_\ell \in C$ and the other is not — precisely when the edge e_j crosses the cut. Otherwise, the coordinate will be anonymized in only one set. Hence, the total cost of anonymizing both sets is $n \cdot (m + t)$.

Now suppose that we have a partition (S_1, S_2) of the bit vectors $\mathbf{b}_i$ such that the cost of anonymizing the partition is $n \cdot (m + t)$ for some value of t (not necessarily an integer). We construct a cut for the original graph $G = (V, E)$ by setting $C = \{v_i \mid \mathbf{b}_i \in S_1\}$. Because the partition (S_1, S_2) was required to be balanced, the cut $(C, V - C)$ must also be balanced. But what is the cost?

Consider the edge $e_j = (v_k, v_\ell)$. We constructed the cut based on which bit vectors were in S_1, so e_j will cross the cut if and only if $\mathbf{b}_k \in S_1$ and $\mathbf{b}_\ell \in S_2$, or vice versa. Therefore, if e_j crosses the cut, then both S_1 and S_2 contain a vector that is non-zero at the coordinate j, and so both S_1 and S_2 are not anonymous in coordinate j. If e_j doesn't cross the cut, then either S_1 or S_2 is not anonymous in coordinate j, but not both.

Hence, if z is the number of edges crossing the cut, then we know that the cost of anonymizing the partition (S_1, S_2) is equal to $2n \cdot z + n \cdot (m - z) = n \cdot (m + z)$. By assumption, this must also equal $n \cdot (m + t)$. The variables m and n are fixed, so the only way these equations can be equal are if $t = z$. $\square$

2.2 Hardness of the k-Distinct Problem

We show that the k-distinct problem is hard by reducing from the balanced 2-distinct problem to the 2-distinct problem. From there, it is easy to see that there is a reduction from the 2-distinct problem to the k-distinct problem.

Let $2n$ be the number of vectors, and let d be the length of all vectors. Let $\mathbf{b}_0, \ldots, \mathbf{b}_{2n-1}$ be the bit vectors. Let c be the value $2nd+1$. By construction, c is strictly greater than the maximum number of $\mathtt{x}$s necessary to anonymize the bit vectors $\mathbf{b}_i$.

$$\begin{aligned}
\mathbf{b}_1 &= \langle 0, \quad 0, \quad 0\rangle \\
\mathbf{b}_2 &= \langle 0, \quad 0, \quad 1\rangle \\
\mathbf{b}_3 &= \langle 0, \quad 1, \quad 0\rangle \\
\mathbf{b}_4 &= \langle 0, \quad 1, \quad 1\rangle \\
\mathbf{b}_5 &= \langle 1, \quad 0, \quad 0\rangle \\
\mathbf{b}_6 &= \langle 1, \quad 0, \quad 1\rangle
\end{aligned}$$

(a) The original vectors.

$$\begin{aligned}
\mathbf{b}'_1 &= \langle 0,0,0, \; \underbrace{1,\ldots,1}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}\rangle \\
\mathbf{b}'_2 &= \langle 0,0,1, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{1,\ldots,1}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}\rangle \\
\mathbf{b}'_3 &= \langle 0,1,0, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{1,\ldots,1}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}\rangle \\
\mathbf{b}'_4 &= \langle 0,1,1, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{1,\ldots,1}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}\rangle \\
\mathbf{b}'_5 &= \langle 1,0,0, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{1,\ldots,1}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}\rangle \\
\mathbf{b}'_6 &= \langle 1,0,1, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{0,\ldots,0}_{\text{19 times}}, \; \underbrace{1,\ldots,1}_{\text{19 times}}\rangle
\end{aligned}$$

(b) The transformed vectors.

Figure 2: An example showing how to reduce from the balanced 2-distinct problem to the 2-distinct problem.

We construct a new set of $2n$ vectors $\mathbf{b}'_i$ with length $d' = d + 2n \cdot c$ as follows:

$$\mathbf{b}'_i[j] = \begin{cases} \mathbf{b}_i[j] & \text{if } j < d \\ 1 & \text{if } i \cdot c \leq j - d < (i+1) \cdot c \\ 0 & \text{otherwise} \end{cases}$$

An example of this transformation is depicted in Figure 2.

Again, this reduction is clearly polynomial. To show that it works, we must show that:

Theorem 2. *A set of bit vectors $B = \{\mathbf{b}_1, \ldots, \mathbf{b}_{2n}\}$ has a balanced partition with anonymizing cost $t < c$ if and only if the bit vectors $B' = \{\mathbf{b}'_1, \ldots, \mathbf{b}'_{2n}\}$ constructed by this reduction has some partition with anonymizing cost $t' = 2n^2c + t$.*

Proof. Suppose that we have a balanced partition of B into sets S_1 and S_2 with anonymizing cost t. Consider the partition defined as follows: $S'_1 = \{\mathbf{b}'_i \mid \mathbf{b}_i \in S_1\}$ and $S'_2 = \{\mathbf{b}'_i \mid \mathbf{b}_i \in S_2\}$. What is the cost of anonymizing this partition? The cost of anonymizing the first d coordinates is the same as the cost of anonymizing S_1 and S_2. For the next $2nc$ coordinates, note that in both S'_1 and S'_2, exactly nc of those coordinates will be all zeroes, and exactly nc of those coordinates will contain exactly one 1. Hence, the total cost of anonymizing is $t + 2 \cdot (nc) \cdot (n) = 2n^2c + t$, just as we wanted.

Now suppose that we have a (not necessarily balanced) partition of B' into sets S'_1 and S'_2, and that the cost of anonymizing these sets is $t' = 2n^2c + t$, where $t < c$.

For the sake of contradiction, suppose that the partition (S'_1, S'_2) is not balanced. Let $n_1 = |S'_1|$ and let $n_2 = |S'_2|$. Without loss of generality, suppose $n_1 < n_2$. Then we must consider two cases:

1. $n_1 = 1$ and $n_2 = 2n - 1$. Because $n_1 = 1$, S'_1 will contribute nothing to the cost of anonymization. Of the $2nc$ additional columns added, S'_2 will match in only c of them, so the total cost of anonymization is at least:

$$\begin{aligned} n_2 \cdot (2nc - c) &= c \cdot (2n-1)^2 = c(4n^2 - 4n + 1) \\ &= 2n^2c + c + (2n^2c - 4nc) \\ &= 2n^2c + c + 2nc(n-2) \end{aligned}$$

   Because $n \geq 2$, we know that this cost must be strictly greater than $2n^2c + t$, so we have a contradiction.

2. $n_1 \neq 1$. Then both S'_1 and S'_2 will contribute to the cost of anonymization. Specifically, cn_1 columns of S'_1 will not be anonymous, and cn_2 columns of S'_2 will not be anonymous. For notational simplicity, let $n_1 = n - z$ and let $n_2 = n + z$. Then the total cost of anonymization will be at least:

$$\begin{aligned} cn_1 \cdot n_1 + cn_2 \cdot n_2 &= c((n-z)^2 + (n+z)^2) \\ &= c(n^2 - 2nz + z^2 + n^2 + 2nz + z^2) \\ &= c(2n^2 + 2z^2) = 2n^2c + c + c(2z^2 - 1) \end{aligned}$$

   Because $n_1 < n_2$, we know that $z \geq 1$, so the total cost of anonymization will be strictly greater than $2n^2c + t$, so we have a contradiction.

Hence, if $n_1 \neq n_2$, we have a contradiction: the cost of the split is too large. Therefore, if we do have a split with cost $2n^2c + t$ ($t < c$), then $n_1 = n_2$.

We construct S_1 and S_2 as follows: $S_1 = \{\mathbf{b}_i \mid \mathbf{b}'_i \in S'_1\}$ and $S_2 = \{\mathbf{b}_i \mid \mathbf{b}'_i \in S'_2\}$. We have just argued that this split will be balanced. The cost of anonymization contributed by the last $2nc$ coordinates will be equal to $2 \cdot (nc) \cdot (n) = 2n^2c$. Hence, the cost of anonymization contributed by the first d coordinates must be equal to t, just as we wanted. □