Journal of Combinatorial Optimization, 7, 339–352, 2003

Random Pooling Designs Under Various Structures*

F.K. HWANG
Y.C. LIU u8722518@math.nctu.edu.tw
Department of Applied Mathematics, National Chiao Tung University, Hsinchu 30050, Taiwan, ROC

Received February 5, 2003; Revised June 10, 2003; Accepted July 29, 2003

Abstract. Balding et al. (1995) showed that randomizing over the k-set space yields much better pooling designs than the random pooling design without the k-restriction. A natural question arises as to whether a smaller subspace, i.e., a space with more structure, will yield even better results. We take the random subset containment design recently proposed by Macula, which randomizes over a subspace of the k-set space, as our guinea pig to compare with the k-set space. Unfortunately the performance of the subset containment design is hard to analyze and only approximations are given. For a set of parameters, we are able to produce either an exact analysis or very good approximations. The comparisons under these parameters seem to favor the k-set space.

Keywords: random pooling designs, clone library screening, k-clique

1. Introduction

A d-disjunct matrix is a 0–1 matrix satisfying certain conditions. Let columns be indexed by $1, \ldots, n$ and rows by $1, \ldots, t$. Then a column can be viewed as a subset of $\{1, \ldots, t\}$ where the elements of the subset are the row indices corresponding to the 1-entries. The condition of d-disjunctness is that no column is contained in the union of any d other columns. d-disjunct matrices are used in pooling designs with a given probe where each column represents a *DNA* clone, either positive if it contains the probe as a subsequence or negative if not, and each row a pool. Namely, those columns having 1-entries in the pool are mixed together for a joint probing which gives a negative outcome if all clones in the pool are negative and a positive outcome otherwise. Such a pool is called negative or positive according to its outcome.

Suppose c of the n clones are positive. Kautz and Singleton (1964) showed that a d-disjunct matrix can identify all positives if $c \leq d$. Furthermore, the identification is easy by simply eliminating all clones which appear in rows with negative outcomes (those must be negative clones). Clones not eliminated are positive.

Construction of d-disjunct matrices have been studied (see Balding et al. (1995), Du and Hwang (2000) and Ngo and Du (2000) for a summary) but their existence is sparse. On the other hand, a random design in which each cell is 1 with probability p exists for all size $t \times n$. Of course, a random design has no d-disjunct property. So both negative and positive clones may remain unresolved at the end of probings. A negative clone is unresolved if it appears

*Research partially supported by the Republic of China NSC grant 90-2115-M-009-029.

in no negative pool. A positive clone is unresolved if all pools containing it either have unresolved negative clones or other positive clones. Let P_n^- (or P_n^+) denote the probability that a negative (or positive) clone is unresolved with n total clones randomly chosen from the sample space. (Delete the subscription if the probability is independent of the clone's number). Surprisingly, random designs can have small ratio $\frac{t}{n}$ and small P_n^-, P_n^+.

Let $[t]$ denote the set $\{1, \ldots, t\}$ and let the $2^{[t]}$-space denote the space consisting of 2^t (binary) t-vectors. Note that for $p = \frac{1}{2}$ in the random design, its n columns can be viewed as randomly chosen with replacement from the $2^{[t]}$-space. For general p, then the n columns are randomly chosen from a certain probability space. Balding et al. (1995) and Bruno et al. (1995) considered a subspace of the $2^{[t]}$-space consisting of all k-subsets of $[t]$. A random k-set design is then a $t \times n$ matrix whose n columns are randomly chosen from the k-subset space with replacement. They showed that the random k-set design improves over the random design significantly. The random k-set design has been extended to the random distinct k-set design (Balding et al., 1995; Bruno et al., 1995) in which the sampling is without replacement, thus a smaller sample space. The analysis in Hwang (2000) and Hwang and Liu (2001) showed that there is a further improvement, though slight.

It is natural to ask whether more structure can bring even bigger benefit. Thus we look into the random subset containment design (RSCD) proposed by Macula (1997) which randomizes over a subspace of the distinct k-set space. The subspace S, characterized by three parameters (m, k, d), $m > k > d \geq 1$, is obtained in the following way: Consider a $\binom{m}{d} \times \binom{m}{k}$ binary matrix where each column is labeled by a distinct k-subset of the set $[m]$ which satisfies $\binom{m}{k} \geq n$, and each row by a distinct d-subset. Cell (i, j) has a 1-entry if and only if the label of row i is a subset of the label of column j. S then consists of all columns in this matrix. Note that S is a subspace of the distinct $\binom{k}{d}$ -set space since the columns are distinct and have $\binom{k}{d}$ 1-entries. A RSCD is then a set of n random columns, $n \leq \binom{m}{k} \equiv N$, from S. Macula proved (1997) that a (m, k, d) RSCD is d-disjunct and he also proposed to use it for $c > d$ (Macula, 1999). As a RSCD is hard to analyze, Macula gave a decoding method which simplifies the analysis but also loses some power. We are concerned with two issues here.

(i) Does Macula's decoding give a fair evaluation of the RSCD?
(ii) How does the RSCD compare with the random distinct k-set design?

Empirical results show that the RSCD is most effective when k is small. Then the requirement $\binom{m}{k} \geq n$ forces m to be not so small. Hence the number of pools $\binom{m}{d}$ could be too large if $d \geq 3$. Thus $d = 2$ appears to be the practical choice in most situations and has become the focus of Macula's study. We will also assume $d = 2$ in the following sections.

We propose a different decoding method and prove that it always performs better than Macula's decoding. But the performance of the RSCD is hard to analyze under this decoding. So far, our analysis covers only the case $d = 2$ and $c = 3$. (c is derived from the number of cutting of the target *DNA* sequence and usually falls into the range $3 \leq c \leq 10$.) Nevertheless, since there is no reason to believe that this particular case favors any side in the comparisons we intend, it provides a platform to compare the two decoding methods and the two designs. Our numerical results will show that the gap between the two sides in each comparison is significant.

	123	125	134	135	145	234	235	245	345	outcome
12	1	1								1
13	1		1	1						1
14			1		1					1
15		1		1	1					1
23	1					1	1			1
24						1		1		1
25		1					1	1		0
34			1			1			1	1
35				1			1		1	0
45					1			1	1	1

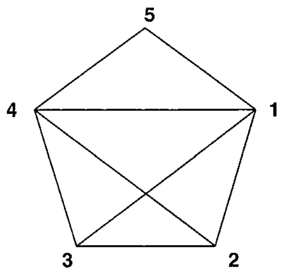

Figure 1. A $m = 5, k = 3, n = 9$ RSCD with three positive clones $\{1, 2, 3\}$, $\{2, 3, 4\}$ and $\{1, 4, 5\}$ and its outcome graph.

2. The representative decoding method

A given $t \times n$ RSCD yields a binary t-vector as outcomes of the t pools, namely, pool i has outcome 1 if it contains at least one positive clone, and 0 otherwise. For such an outcome vector, define its outcome graph as a graph G having $[m] = \{1, \ldots, m\}$ as its vertex-set and an edge (u, v) if the row labeled by $\{u, v\}$ has outcome 1. Figure 1 gives an example of a RSCD with $m = 5$, $k = 3$, $n = 9$ and its outcome graph when the positive clones are $\{1, 2, 3\}$, $\{2, 3, 4\}$ and $\{1, 4, 5\}$. Note that each column (clone) has a corresponding set of k vertices, which is just its column label, in the outcome graph. We denote the vertex-set of a clone C as $V(C)$.

A representative is a vertex with degree $k - 1$ in G. Macula proved that this vertex together with its $k - 1$ adjacent vertices, yield a k-set corresponding to a resolved positive clone. Hence the decoding method, we call it the representative decoding, is to identify clones with at least one representative as resolved positive clones. In figure 1, vertex 5 is a representative and $\{1, 4, 5\}$ is identified as a positive. But the two other positive clones $\{1, 2, 3\}$ and $\{2, 3, 4\}$ have no representatives and hence are unresolved. Note that whether a positive clone is unresolved is not affected by the sampling of unresolved negative clones in representative decoding. Hence we denote the probability of a positive being unresolved under representative decoding as P^+. Macula gave (1999)

$$P^+ = 1 - \sum_{i=1}^{k} (-1)^{i-1} \binom{k}{i} \binom{\binom{m-i}{k}}{c-1} \binom{\binom{m}{k} - 1}{c-1}^{-1},$$

where i is the number of vertices of the given positive clone which do not appear in the other $c - 1$ positive clones.

Let $P^v(\bar{P} = x)$ denote the probability that x positive clones are unresolved by representative decoding. Macula (1999) approximated $P^v(\bar{P} = 0)$ by $(1 - P^+)^c$. We now give an exact solution of $P^v(\bar{P} = x)$ for the case $c = 3$ for later comparisons.

Theorem 1.

$$\text{(i)}\quad P^v(\bar{P}=3)=\frac{1}{6}\binom{m}{k}\sum_{i\geq\lceil\frac{k}{2}\rceil}^{k-1}\binom{k}{i}\binom{m-k}{k-i}\binom{i}{2i-k}\binom{\binom{m}{k}}{3}^{-1},$$

$$\text{(ii)}\quad P^v(\bar{P}=2)=\frac{1}{2}\binom{m}{k}\sum_{i\geq\lceil\frac{k}{2}\rceil}^{k-1}\binom{k}{i}\binom{m-k}{k-i}\left[\binom{m-2(k-i)}{2i-k}-\binom{i}{2i-k}\right]\binom{\binom{m}{k}}{3}^{-1},$$

$$\text{(iii)}\quad P^v(\bar{P}=1)=\frac{1}{2}\binom{m}{k}\sum_{i=0}^{k-1}\binom{k}{i}\binom{m-k}{k-i}\sum_{j=1}^{k-i-1}\sum_{h=1}^{k-i-1}\binom{k-i}{j}\binom{k-i}{h}\binom{i}{k-j-h}\binom{\binom{m}{k}}{3}^{-1}.$$

Proof: Suppose $1^*, 2^*, 3^*$ are the three positive clones. Without loss of generality, assume $|V(1^*)\cap V(2^*)|=i$ for $i=0,1,\ldots,k-1$.

(i) Each vertex in G must be in at least two positive clones since any vertex in a single positive clone has degree $k-1$, hence would be a representative. Note that two vertices each in two positive clones must share a positive clone. Therefore G is a k'-clique with $k' > k$. There are $\binom{m}{k}$ ways of choosing $V(1^*)$, $\binom{k}{i}\binom{m-k}{k-i}$ ways of choosing $V(2^*)$ which intersects $V(1^*)$ in i elements, and $\binom{i}{2i-k}$ ways of choosing $2i-k$ elements from $V(1^*)\cap V(2^*)$, which, together with the $k-i$ elements from $V(1^*)\backslash V(2^*)$ and the $k-i$ elements from $V(2^*)\backslash V(1^*)$, yields $V(3^*)$. The lower bound $\lceil\frac{k}{2}\rceil$ of i guarantees $\binom{i}{2i-k}>0$, and the upper bound $k-1$ is due to the distinctness of $1^*, 2^*$. Finally, since $1^*, 2^*, 3^*$ are interchangeable, we divide the sum by six.

(ii) Let the two positive clones with no representative be 1^* and 2^*. Then $V(3^*)$ must contain all vertices in $V(1^*)\backslash V(2^*)$ and $V(2^*)\backslash V(1^*)$, as well as an vertex outside of $V(1^*)\cup V(2^*)$. Again, there are $\binom{m}{k}\binom{k}{i}\binom{m-k}{k-i}$ ways of choosing $V(1^*)$ and $V(2^*)$ with i intersections. $V(3^*)$ must have the $k-i$ vertices from each of $V(1^*)\backslash V(2^*)$ and $V(2^*)\backslash V(1^*)$, and $2i-k$ more vertices elsewhere, excluding the case that the $2i-k$ vertices are all in $V(1^*)\cup V(2^*)$ (case (i)). We divide the result by 2 since 1^* and 2^* are interchangeable.

(iii) Let the two clones with representatives be 1^* and 2^*. Then $V(3^*)$ must not contain all vertices in $V(1^*)\backslash V(2^*)$, or all vertices in $V(2^*)\backslash V(1^*)$. Further, $V(3^*)\subset V(1^*)\cup V(2^*)$.

□

Corollary 2. $P^v(\bar{P}=0)=1-P^v(\bar{P}=3)-P^v(\bar{P}=2)-P^v(\bar{P}=1)$ *when* $c=3$.

3. The unique-representative decoding

The unique-representative decoding was used in Balding et al. as follows:

1. A clone contained in a negative pool is a resolved negative clone.
2. Remove all resolved negative clones in the design. If a clone has a unique appearance in a positive pool, then it is a resolved positive clone.

For $d = 2$, each row is represented by an edge and each column (clone) a set of k vertices. The edges of a clone is simply all the 2-subsets of the k vertices, denoted as $E(C)$. Then the unique-representative decoding is equivalent to the edge-representative decoding defined as follows:

1. A clone C with at least one edge not in the outcome graph G is a resolved negative clone.
2. A clone C with all its edges in G is a resolved positive clone if at least one of its edge, called an edge-representative, is not in any other k-clique of G which represents a clone.

Figure 2 is an example demonstrating this relation. The 3 positive clones are $\{1, 2, 3\}$, $\{2, 3, 4\}$ and $\{1, 4, 5\}$. Corresponding to the negative outcomes of rows $\{2, 5\}$ and $\{3, 5\}$, edges $(2, 5)$ and $(3, 5)$ are not in the outcome graph. Hence clones $\{1, 2, 5\}$, $\{1, 3, 5\}$, $\{2, 3, 5\}$, $\{2, 4, 5\}$ and $\{3, 4, 5\}$ are resolved negatives. Since clone $\{1, 2, 3\}$ has edge-representative $(1, 2)$, which is not in any other k-clique of G (note that $\{1, 2, 4\}$ doesn't represent a clone), it is a resolved positive clone. With the same reason, $\{2, 3, 4\}$ and $\{1, 4, 5\}$ are both resolved positive clones.

Lemma 3. *For $d = 2$, unique-representative decoding is equivalent to edge-representative decoding.*

Proof: Suppose a clone C has an edge (u, v) not in G. Then C must be negative because there exists a row $\{u, v\}$ containing C but no positive clone. So C is resolved under the unique-representative decoding.

	1	1	1	1	1	2	2	2	3	
	2	2	3	3	4	3	3	4	4	outcome
	3	5	4	5	5	4	5	5	5	
12	1	1								1
13	1		1	1						1
14			1		1					1
15		1		1	1					1
23	1					1	1			1
24						1		1		1
25		1					1	1		0
34			1			1			1	1
35				1			1		1	0
45					1			1	1	1

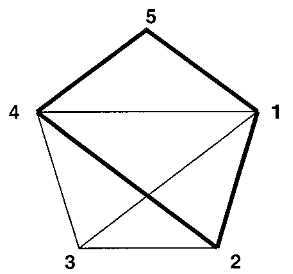

Figure 2. When $d = 2$, unique-representative decoding is equivalent to edge-representative decoding.

Next suppose C has all its edges in G. Further, C has an edge (u, v) not in any other k-clique of G which represents a clone. Then C must be positive since otherwise edge (u, v) would not be in G. The above argument also implies C is resolved.

Since our arguments are reversible, all clones not satisfying the conditions are unresolved under the unique-representative decoding. □

For easier distinction we will refer to a representative under the representative decoding as a vertex-representative.

Theorem 4. *A vertex-representative implies $k - 1$ edge-representatives.*

Proof: Let vertex v be a vertex-representative of clone C. Then v has $k - 1$ edges, while each is an edge-representative of C.

Suppose to the contrary that one of these $k - 1$ edges (v, u) is not an edge-representative, i.e., (v, u) is also an edge in $E(C^{'})$, $C^{'} \neq C$. Let w be a vertex in $V(C^{'})\backslash V(C)$. Then (v, w) is in G. Hence v has at least degree k, contradicting the fact that v is a vertex-representative. □

Theorem 4 says that whenever a positive clone is resolved under the representative decoding, it is resolved under the edge-representative decoding. (The representative decoding does not specify how a negative clone is resolved; we assume that it is the same as in the edge-representative decoding.) Further, the edge-representative decoding is at least as easy to use as the representative-decoding, if not easier. Therefore although P_n^- and P_n^+ are harder to compute under the edge-representative decoding, that shouldn't prevent us from using it as long as we know it guarantees better performance than its alternative.

4. P_n^- and P_n^+ for edge-representative decoding with $d = 2$ and $c = 3$

Define $V_C(A, B) = (V(A)\backslash V(B)) \cap C$, and $E_C(AB) = \{$edges between $V_C(A, B)\}$ and $V_C(B, A)\}$. Let $1^*, 2^*, 3^*$ denote the three positive clones.

Lemma 5. *Suppose C is a negative. Then $E(C) \subseteq E(1^*) \cup E(2^*) \cup E(3^*)$ if and only if* (i) $V(C) \subset V(1^*) \cup V(2^*)$, (ii) $E_C(1^*2^*) \subset E(3^*)$,(iii) $V(1^*) \cap V(2^*) \cap V(C) \neq \emptyset$.

Proof:

(i) Suppose there exists a vertex $v \in V(C)\backslash(V(1^*) \cup V(2^*))$. Note that $E(3^*)$ must contain all edges from v to $V(C)\backslash\{v\}$, which implies $V(C) = \{v\} \cup (V(C)\backslash\{v\}) \subseteq V(3^*)$, an absurdity since C and 3^* are distinct k-sets.

(ii) Obvious.

(iii) Suppose $V(1^*) \cap V(2^*) \cap V(C) = \emptyset$. Then $V_C(1^*, 2^*) \cup V_C(2^*, 1^*) = V(C)$. So $V(3^*)$ must contain $V(C)$, an absurdity.

□

Theorem 6.

$$P_N^- = \binom{N-1}{3}^{-1} \frac{1}{6} \sum_{i=1}^{k-1} \binom{k}{i} \binom{m-k}{k-i} \cdot \sum_{j=k-i+1}^{k-1} \binom{i}{i+j-k} \binom{m-k}{k-j} \left[\binom{m+i+j-2k}{i+j-k} - 1 \right].$$

Proof: To cover $E(C)$, $V(1^*)$ and $V(2^*)$ must cover all k vertices of $V(C)$. Suppose $V(1^*)$ covers i such vertices in $\binom{k}{i}$ ways. Then $V(2^*)$ must cover the remaining $k-i$ plus at least one from $V(C) \cap V(1^*)$ vertices. Let j denote the number of $V(C) \cap V(2^*)$. Then $V(1^*)$ must also take $k-i$ vertices not in $V(C)$ in $\binom{m-k}{k-i}$ ways, and $V(2^*)$ takes $k-j$ vertices in $\binom{m-k}{k-j}$ ways. Finally, $V(3^*)$ must take the $k-(i+j-k)$ vertices in $V_C(1^*, 2^*) \cup V_C(2^*, 1^*)$, hence it must take $i+j-k$ vertices from the remaining $(m-k)+(i+j-k)$ vertices not in $V_C(1^*, 2^*) \cup V_C(2^*, 1^*)$. But one of the choice is C, which must be subtracted.

Since $1^*, 2^*$ and 3^* can be interchanged, we divide the result by 6 to eliminate the repetitive counting. We also divide by the total number of ways of choosing three positive clones to obtain a probability. □

Corollary 7. *$P_n^- = P_N^-$ for all $1 \leq n < N$.*

Proof: P_n^- can be computed by counting the number of ways choosing 3 positive clones to cover C and choosing $n-4$ other negative clones, then divided by the total number of ways of choosing $n-1$ clones including 3 positive ones. Thus

$$P_n^- = \frac{(P_N^-)\binom{N-1}{3}\binom{N-4}{n-4}}{\binom{N-1}{n-1}\binom{n-1}{3}} = P_N^-$$

□

Since $P_n^- = P_N^-$ for all $1 \leq n < N$, we denote it as P^- in the following discussion.

It is more tricky to compute P_n^+. Suppose we compute P_n^+ for 1^*.

Lemma 8. *1^* is unresolved if and only if $V(1^*)$ is contained in a k'-clique with $k' > k$ in G and $E_{1^*}(2^*3^*)$ is covered by unresolved negative clones in the n-sample.*

Proof: For 1^* to be unresolved, necessarily $V(1^*) \subset V(2^*) \cup V(3^*)$. Suppose not. Then there exists a vertex $v \in V(1^*) \backslash (V(2^*) \cup V(3^*))$ which is a vertex-representative of 1^*, contradicting the assumption that 1^* is unresolved.

So we assume $V(1^*) \subset V(2^*) \cup V(3^*)$. Note that $E(1^*) \not\subseteq E(2^*) \cup E(3^*)$ since then, $E(3^*)$ must contain all edges between $V(1^*) \cap V(2^*)$ and $V(1^*) \backslash V(2^*)$, which implies $V(3^*)$ must contain $(V(1^*) \cap V(2^*)) \cup (V(1^*) \backslash V(2^*)) = V(1^*)$, an absurdity. Therefore for 1^* to be unresolved, there must exist some unresolved negative clones to cover $E_{1^*}(2^*3^*)$. Note that an unresolved negative clone C has $E(C) \subseteq E(1^*) \cup E(2^*) \cup E(3^*)$. For C to cover an edge in $E_{1^*}(2^*3^*)$, C must have a vertex u belonging to $V_{1^*}(2^*, 3^*)$, and a vertex v belonging to $V_{1^*}(3^*, 2^*)$. Further, since $V(C) \neq V(1^*)$, there exists a vertex w belonging to $V(C)\backslash V(1^*)$. Since the edge (u, w) can only be covered by $E(2^*)$, w belonging to $V(2^*)$. Similarly, since the edge (v, w) can only be covered by $E(3^*)$, w belonging to $V(3^*)$ too. Since $V(1^*)$ belonging to $V(2^*) \cup V(3^*)$, w has edges to every vertex in $V(1^*)$; thus $w \cup V(1^*)$ is a $(k+1)$-clique. □

Define $z = \binom{k+h}{k} - 1 - I_{h=j} - I_{h=k-i-j}$ where $I_q = 1$ if q holds and $I_q = 0$ otherwise.

Theorem 9.

$$
\begin{aligned}
P_n^+ \leq{} & \binom{N-1}{n-1}^{-1} \binom{n-1}{2}^{-1} \left(\frac{1}{2}\right) \cdot \sum_{i=0}^{k-2} \binom{k}{i} \sum_{j=1}^{k-i-1} \binom{k-i}{j} \binom{m-k}{k-i-j} \\
& \sum_{h=1}^{\min\{j,k-i-j\}} \binom{k-i-j}{h} \binom{m+i+j-2k}{j-h}. \\
& \left\{ \binom{N-3}{n-3} - \binom{N-3-z}{n-3}\binom{z}{0} - \binom{N-3-z}{n-4} \left[\binom{z}{1} - \binom{\binom{i+h}{i}-1}{1} \right] \right. \\
& \left. - \binom{N-3-z}{n-5} \cdot \left[\binom{z}{2} - F \right] \right\}.
\end{aligned}
$$

where

$$
\begin{aligned}
F ={} & \binom{\binom{i+h}{i}-1}{2} + \binom{\binom{i+h}{i}-1}{1} \binom{z-\binom{i+h}{i}+1}{1} \\
& + \frac{1}{2} \sum_{x=1}^{j-1} \sum_{y=0}^{x-1} \binom{j}{x} \binom{x}{y} \binom{i+h}{i+j-x} \binom{i+h}{i+x-y} \\
& + \frac{1}{2} \sum_{x=1}^{k-i-j-1} \sum_{y=0}^{x-1} \binom{k-i-j}{x} \binom{x}{y} \binom{i+h}{k-j-x} \binom{i+h}{i+x-y}.
\end{aligned}
$$

Proof: There are $\binom{k}{i}$ ways of choosing i vertices from $V(1^*)$ to appear in both $V(2^*)$ and $V(3^*)$. Among the remaining $k-i$ vertices of $V(1^*)$, $V(2^*)$ chooses j (and $V(3^*)$ the remaining $k-i-j$). $V(2^*)$ also chooses $k-i-j$ vertices outside of $V(1^*)$, and $V(3^*)$ chooses j among which h are from the $k-i-j$ vertices of $V(2^*)\backslash V(1^*)$, and $j-h$ from the remaining $(m-k)-(k-i-j)$ vertices. We divide the result by 2 since 2^* and 3^* can be interchanged.

There are $\binom{N-3}{n-3}$ ways of choosing the other $n-3$ clones. We subtract those cases for which 1^* is resolved. We count these cases according to the number of unresolved negatives contained in the $n-3$ samples. Suppose G contains a $(k+h)$-clique, where h is the number of vertices in $V(2^*) \cap V(3^*)$ outside of $V(1^*)$. Then any k-clique in the $(k+h)$-clique is an unresolved negative as long as it is not a positive clone or 1^* itself. We subtract 1 since 1^* is such a clique. We also subtract $I_{h=j}$ (implying 3^* is such a clique) and $I_{h=k-i-j}$ (implying 2^* is such a clique). Therefore

$$z = \binom{k+h}{k} - 1 - I_{h=j} - I_{h=k-i-j}$$

is the number of unresolved negatives. An unresolved negative is called universal if it contains $V_{1^*}(2^*, 3^*) \cup V_{1^*}(3^*, 2^*)$ (hence covering $E_{1^*}(2^*3^*)$. Since its other i vertices must be taken from the $i+h$ vertices of $V(2^*) \cap V(3^*)$ except the choice yielding 1^*, there are $\binom{i+h}{i} - 1$ universal unresolved negative clones).

If no unresolved negative is sampled, then 1^* is resolved. If one unresolved negative U is chosen, then 1^* is resolved unless U is universal.

Suppose two unresolved negative U is chosen, then 1^* is resolved unless one of the following subcases occurs:

(i) Both U_1 and U_2 are universal.
(ii) Exactly one of U_1 and U_2 is universal.
(iii) Both $V(U_1)$ and $V(U_2)$ contain $V_{1^*}(3^*, 2^*)$ and together, cover $V_{1^*}(2^*, 3^*)$ (but neither is universal). More specifically, U_1 consists of the $k-i-j$ vertices of $V_{1^*}(3^*, 2^*)$, x vertices of $V_{1^*}(2^*, 3^*)$, and $i+j-x$ vertices from $V(2^*) \cap V(3^*)$, while U_2 consists of the $k-i-j$ vertices of $V_{1^*}(3^*, 2^*)$, the $j-x+y$ vertices of $V_{1^*}(2^*, 3^*)$ including y of the x vertices chosen in U_1, and the $i+x-y$ vertices from $V(2^*) \cap V(3^*)$. We divide the result by 2 since U_1 and U_2 can be interchanged.
(iv) Both $V(U_1)$ and $V(U_2)$ contain $V_{1^*}(2^*, 3^*)$ and together, cover $V_{1^*}(3^*, 2^*)$.

If at least three unresolved negatives are sampled, we assume 1^* is unresolved even though there are counterexamples. Therefore, the formula is an upper bound of P_n^+. □

Corollary 10.

$$P_N^+ = \binom{N-1}{2}^{-1} \frac{1}{2} \sum_{i=0}^{k-2} \binom{k}{i} \sum_{j=1}^{k-i-1} \binom{k-i}{j} \binom{m-k}{k-i-j} \cdot \sum_{h=1}^{\min\{j,k-i-j\}} \binom{k-i-j}{k} \binom{m+i+j-2k}{j-h}.$$

Proof: When every member of $\binom{[m]}{k}$ is taken as a clone, then all the z unresolved negatives exist. Therefore 1^* is always unresolved as long as $h \geq 1$ (a k-clique which contains $V(1^*)$ exists in G with $k' > k$).

Note that for $n = N$, the upper bound in Theorem 9 is same as P_N^+ in Corollary 10 since z is at least 1 and it can be verified:

(i) $\binom{N-3-z}{n-3} = 0$ for $z \geq 1$,
(ii) $\binom{N-3-z}{n-4} = 0$ for $z \geq 2$, $\binom{z}{1} - \binom{\binom{i+h}{i}-1}{1} = 0$ for $z = 1$,
(iii) $\binom{N-3-z}{n-5} = 0$ for $z \geq 3$, the term in [ ] in Theorem 9 following $\binom{N-3-z}{n-5}$ is equal to 0 for $z = 1$ or 2.

□

Corollary 11. *By setting $h = 1$ in the upper bound in Theorem 9, we obtain a lower bound of P_n^+.*

Proof: For $h = 1$, each unresolved negative is obtained by replacing a vertex in $V(1^*)$ by the intersection vertex in $V(2^*)$ and $V(3^*)$ outside of $V(1^*)$. Suppose three unresolved negatives are sampled. Then either a universal type is sampled, or two of them satisfy the description in case (iii) (or (iv)) in the proof of Theorem 9. Hence 1^* is always unresolved. The reason of being a lower bound is that unresolvedness corresponding to large h is ignored.

□

Next we compute the probability that there's no unresolved positive clones by edge-representative decoding without sampling, and denote it as $P_N^e(\bar{P} = 0)$,

Theorem 12.

$$\text{(i)}\quad P_N^e(\bar{P} = 3) = \frac{1}{6}\binom{m}{k} \sum_{i=\lceil \frac{k}{2} \rceil}^{k-1} \binom{k}{i}\binom{m-k}{k-i}\binom{i}{2i-k}\binom{\binom{m}{k}}{3}^{-1},$$

$$\text{(ii)}\quad P_N^e(\bar{P} = 2) = \frac{1}{2}\binom{m}{k} \sum_{i=\lceil \frac{k}{2} \rceil}^{k-1} \binom{k}{i}\binom{m-k}{k-i}\left[\binom{m-2(k-i)}{2i-k} - \binom{i}{2i-k}\right]\binom{\binom{m}{k}}{3}^{-1},$$

$$\text{(iii)}\quad P_N^e(\bar{P} = 1) = \frac{1}{2}\binom{m}{k} \sum_{i=1}^{k-1} \binom{k}{i}\binom{m-k}{k-i} \sum_{j=0}^{i-1}\binom{i}{j} \cdot \sum_{h=1}^{\min\{k-i,i-j\}-1} \binom{k-i}{h}\binom{m+i-2k}{i-j-h}\binom{\binom{m}{k}}{3}^{-1}$$

Proof:

(i) and (ii) are same as their counterparts in Theorem 1 since the necessary and sufficient conditions for (i) and (ii) are same regardless of vertex-representative or edge-representative.

(iii) Suppose 3^* is the only unresolved positive. Then necessarily, $V(3^*) \subset V(1^*) \cup V(2^*)$ and $V(1^*) \cap V(2^*)$ in $h \geq 1$ vertices outside of $V(3^*)$. Further, $V(3^*)$ must not contain all vertices in $V(1^*) \backslash V(2^*)$ or all vertices in $V(2^*) \backslash V(1^*)$.

Suppose $V(1^*)$ contains i vertices from $V(3^*)$, and $V(2^*)$ contains $k - i + j$ vertices from $V(3^*)$, including the $k - i$ vertices not in $V(1^*)$. Then $V(1^*)$ also contains $k - i$ vertices outside of $V(3^*)$, while $V(2^*)$ contains $i - j$ vertices outside of $V(3^*)$, among which h also appears in $V(1^*)$. □

Corollary 13. $P_N^e(\bar{P} = 0) = 1 - P_N^e(\bar{P} = 3) - P_N^e(\bar{P} = 2) - P_N^e(\bar{P} = 1)$.

5. Some numerical comparisons

In this section, we give some numerical data about the two issues we are concerned with. Table 1 compares the probabilies of the non-existence of unresolved positive clones when $c = 3$ between the representative decoding and the edge-representative decoding, as well as their approximations $(1-P^+)^3$ and $(1-P_N^+)^3$. We find that representative decoding does lose significant power against the edge-representative decoding. For example, when $m = 12$ and $k = 4$, that is $t = 66$ and $N = 495$, $P^v(\bar{P} = 0) = 0.84095$ while $P_N^e(\bar{P} = 0) = 0.95753$.

Table 1. Comparisons between $P^v(\bar{P} = 0)$, $(1 - P^+)^3$, $P_N^e(\bar{P} = 0)$ and $(1 - P_N^+)^3$.

		$m = 6$	$m = 9$	$m = 12$	$m = 15$
		$t = 15$	$t = 36$	$t = 66$	$t = 105$
$k = 3$	$(1 - P^+)^3$	0.320309	0.702409	0.854944	0.920296
	$P^v(\bar{P} = 0)$	0.315789	0.717014	0.863139	0.924556
	$(1 - P_N^+)^3$	0.597172	0.90778	0.969771	0.98745
	$P_N^e(\bar{P} = 0)$	0.789474	0.95504	0.985296	0.993873
$k = 4$	$(1 - P^+)^3$	0.084929	0.589977	0.83255	0.923544
	$P^v(\bar{P} = 0)$	0.043956	0.60129	0.840947	0.927331
	$(1 - P_N^+)^3$	0.186589	0.785259	0.940043	0.979166
	$P_N^e(\bar{P} = 0)$	0.43956	0.852903	0.957527	0.984841
$k = 5$	$(1 - P^+)^3$	0	0.408412	0.778734	0.913776
	$P^v(\bar{P} = 0)$	0	0.402581	0.787051	0.91741
	$(1 - P_N^+)^3$	0	0.563164	0.881667	0.964157
	$P_N^e(\bar{P} = 0)$	0	0.654194	0.903663	0.969958
$k = 6$	$(1 - P^+)^3$		0.189026	0.678769	0.888164
	$P^v(\bar{P} = 0)$		0.150749	0.684508	0.891895
	$(1 - P_N^+)^3$		0.278761	0.773598	0.936329
	$P_N^e(\bar{P} = 0)$		0.388775	0.801976	0.943107
$k = 7$	$(1 - P^+)^3$		0.0357452	0.519217	0.839309
	$P^v(\bar{P} = 0)$		0.0117647	0.51417	0.843035
	$(1 - P_N^+)^3$		0.064	0.598251	0.885695
	$P_N^e(\bar{P} = 0)$		0.223529	0.630783	0.894235

Table 2. P_n^+.

	RSCD				
	Vertex representative	Upper bound	Lower bound	n	Random distinct K-set
$m = 6, t = 15,$		0.10062	0.10062	16	0.23349
$k = 3, K = 3$	0.31579	0.12771	0.12771	18	0.2705
		0.15789	0.15789	20	0.30739
$m = 9, t = 36,$		0.03185	0.02831	50	0.0031
$k = 4, K = 6$	0.16129	0.0637	0.05982	90	0.00705
		0.07742	0.07355	126	0.0119
$m = 12, t = 66,$		0.05483	0.03695	300	1.664×10^{-6}
$k = 6, K = 15$	0.12117	0.07876	0.06084	600	4.106×10^{-6}
		0.08201	0.06409	924	7.8796×10^{-6}
$m = 15, t = 105,$		0.01896	0.00872	1000	5.512×10^{-11}
$k = 7, K = 21$	0.05672	0.02903	0.01814	2000	1.0382×10^{-10}
		0.03559	0.0247	3000	1.5369×10^{-10}

$(1 - P^+)^3$ underestimates $P^v(\bar{P} = 0)$ in most cases. Only when the difference between m and k is small, for example $m = 9$ and $k = 5$, $(1 - P^+)^3$ overestimates $P^v(\bar{P} = 0)$. Furthermore, $(1 - P_N^+)^3$ underestimates $P_N^e(\bar{P} = 0)$ in all cases.

Table 2 compares P_n^+. Data in the "upper/lower bound" column are the upper/lower bounds for P_n^+ by edge-representative decoding. The difference between these two columns is very small, so we are very close to the real value. Furthermore, by Corollary 10, when $n = N$, the value of upper bound is just P_N^+. Again, representative decoding loses some power in computing P_n^+. When $d = 2$, a RSCD is a set of random columns from a subspace of distinct $\binom{k}{2}$-set space, hence we compare it with random distinct $\binom{k}{2}$-set designs. Denote $\binom{k}{2}$ as K, Table 2 also demonstrates the comparison of P_n^+ between RSCD and random distinct K-set designs. We find that RSCD are better only in very limited cases (when $m = 6$) in our data.

In Table 3, we present the comparison of P^- between RSCD and random distinct K-set designs. The probability of a negative clone being unresolved does not depend on n, and P^- is exact. As in the P_n^+ case, random distinct K-set designs are better except for limited cases.

6. Conclusions

(i) Since the edge-representative decoding is as simple as the representative decoding, but provably more powerful, it should always be used to decode the RSCD regardless of how difficult is the analysis.

Table 3. P^-.

		$m = 6$	$m = 9$	$m = 12$	$m = 15$
		$t = 15$	$t = 36$	$t = 66$	$t = 105$
$k = 3, K = 3$	Random distinct K-set	0.09044	0.008667	0.001532	0.0003949
	RSCD	0.027864	0.002351	0.00042	0.0001115
$k = 4, K = 6$	Random distinct K-set	0.19367	0.003212	0.0001155	8.10564×10^{-6}
	RSCD	0.153846	0.006294	0.0006409	0.0001072
$k = 5, K = 10$	Random distinct K-set	0.6673	0.004688	0.0000276	3.8274×10^{-7}
	RSCD	1	0.01989	0.0011896	0.0001263
$k = 6, K = 15$	Random distinct K-set		0.02274	0.000027	6.11551×10^{-8}
	RSCD		0.07493	0.002942	0.0002022
$k = 7, K = 21$	Random distinct K-set		0.173999	0.0001035	3.95241×10^{-8}
	RSCD		0.28342	0.00927	0.0004339

(ii) The representative decoding severely overestimates P_n^+.
(iii) $(1 - P_N^+)^3$ underestimates $P_N^e(\bar{P} = 0)$.
(iv) The random distinct k-set design is better than the RSCD in general when $c = 3$, thus crashing the hope that randomizing over a sensible subspace always brings improvement.

When $d = 2$, two different columns intersects at at most $\binom{k-1}{2}$ rows in a RSCD, while the corresponding number for the random distinct $\binom{k}{2}$-set design is $\binom{k}{2} - 1$. Intuitively, a smaller upper bound of the column intersection number should be good (Balding et al., 1995; Bruno et al., 1995) since then it would take the union of more columns to cover a given column. Surprisingly, this advantage is not enough to establish the RSCD as the favorable design against the random distinct k-set design.

References

D.J. Balding, W.J. Bruno, E. Knill, and D.C. Torney, "A comparative survey of non-adaptive pooling designs," *Genetic Mapping and DNA Sequencing*, IMA Volumes in *Mathematics and its Applications*, Springer Verlag, 1995, pp. 133–155.

W.J. Bruno, E. Knill, D.J. Balding, D.C. Bruce, N.A. Doggett, W.W. Sawhill, R.L. Stalling, C.C. Whittaker, and D.C. Torney, "Efficient pooling designs for library screening," *Genomics*, vol. 26, pp. 21–30, 1995.

D.Du and F.K. Hwang, *Combinatorial Group Testing and its Applications*, 2nd edition. World Scientific, 2000.

F.K. Hwang, "Random k-set pool designs with distinct columns," *Prob. Eng. Inform. Sci.*, vol. 14, pp. 49–56, 2000.

F.K. Hwang and Y.C. Liu, "The expected number of unresolved positive clones in various random pool designs," *Prob. Eng. Inform. Sci.*, vol. 15, pp. 57–68, 2001.

W.H. Kautz and R.R. Singleton, "Nonrandom binary superimposed codes," *IEEE Trans. Inform. Thy.*, vol. 10, pp. 363–377, 1964.

A.J. Macula, “A simple construction of d-disjunct matrices with certain weights,” *Disc. Math.*, vol. 80, pp. 311–312, 1997.

A.J. Macula, “Probabilistic nonadaptive group testing in the presence of errors and DNA library screening,” *Ann. Comb.*, vol. 1, pp. 61–69, 1999.

H.Q. Ngo and D.Z. Du, “A sruvey on combinatorial group testing algorithms with applications to DNA library screening,” *DIMACS Ser. Discrete Math. Theoret. Comput. Sci.*, 55, Amer. Math. Soc., 2000, pp. 171–182.