# Comparing File Copies with at Most Three Disagreeing Pages

F.K. Hwang and P.J. Wan

**Abstract**—Recently, Metzner gave a parallel algorithm to identify two disagreeing pages in comparing two replicated file copies. We show that this algorithm can also identify three disagreeing pages, if one additional comparison is allowed to resolve some possible ambiguity. We also discuss the relation between the file-copy problem and the group testing problem which has been widely studied.

**Index Terms**—Replicated files, disagreeing pages, corrupted pages, combined signatures, group testing.

## 1 Introduction

It is often desirable to keep replicated copies of large files at remote locations to prevent loss of information and to permit easy access. Active files usually require frequent updating. Metzner and Abidi [1] proposed a parity structure to provide a check that updating has been done correctly, and a mechanism for locating discrepancies when they do occur.

Let the file size be $m$ bits. Divide the file into $n$ units, referred to as *pages*, of $m / n$ (assuming divisibility) bits each where $m$ is much larger than $n$. A binary parity sequence $S_x$, or *signature*, is derived from each page $x$. Assume that a signature is a binary $s$-vector. The signature $S_X$ of a set $X$ of pages is simply the modulo two sum of the page-signatures, hence is also a binary $s$-vector. Let $F$ denote the set of pages in the file, and $F'$ a copy of $F$. Define $\Delta_X = S_X - S_{X'}$. If $X$ consists of a single page $x$, we simply write $\Delta_x$. It is usually assumed that $\Delta_k = \{\Delta_X : |X| \leq k\}$ is a set of *pseudorandom* $s$-vectors. Since $k$ is usually selected such that $|\Delta_k|$ is much smaller than $2^s$, it is reasonable to assume that elements in $\Delta_k$ are distinct. When the number of disagreeing pages is small, it is also reasonable to assume that $\Delta_X = 0$ implies $\Delta_x = 0$ for all $x \in X$. Violations of these assumptions will be referred to as *masking errors*.

A page $x$ is called *disagreeing* if $\Delta_x \neq 0$. Let $D$ denote the set of disagreeing pages. The problem is to identify $D$ by comparing sets of signatures, and the goal is to minimize such comparisons. An algorithm is *parallel* if all such comparisons are specified simultaneously, and *sequential* if the specification of one comparison can depend on the outcome of another. A sequential algorithm can further be classified into *k-stage* for $k = 2, 3, \ldots$, if all comparisons can be partitioned into $k$ stages, where comparisons in the same stage are parallel.

We call the problem a $d$-problem if $|D| = d$, and a $\overline{d}$-problem if $|D| \leq d$. Fuchs, Wu, and Abraham [2] proposed a parallel algorithm with $N \equiv \lceil \log_2 n \rceil$ comparisons for the $\overline{1}$-problem. Let page $k$ be indexed by the binary $N$-vector of the number $k$. Then the $i$th comparison consists of signatures of the set of pages with 1 in the $i$th bit. We call the set of comparisons the *binary representation matrix* (BRM). Metzner [3] proposed a sequential algorithm and later [4] improved it with two other sequential tree algorithms. In particular, for the $\overline{2}$-problem, the second algorithm in [4] is equivalent to a parallel algorithm consisting of BRM and a comparison on

- *F.K. Hwang is with AT&T Bell Laboratories, Murray Hill, NJ 07974-0636.*
- *P.J. Wan is with the Department of Computer Science, University of Minnesota, Minneapolis, MN 55455. E-mail: wan@cs.umn.edu.*

*Manuscript received 14 Apr. 1995; revised 10 Oct. 1996.*
*For information on obtaining reprints of this article, please send e-mail to: transcom@computer.org, and reference IEEECS Log Number 104111.0.*

the set of all pages. We will denote this set of $1 + N$ comparisons by BRM$^*$. Barbara, Garcia-Molina and Feijoo [5] also proposed an algorithm for the $\overline{2}$-problem, but their algorithm requires more comparisons.

What if the assumption of $d \le 2$ is found to be wrong after the BRM$^*$ is applied (say, the real problem has $d = 3$)? In this paper, we show that all we need to do is do one more comparison on one page. However, the additional comparison depends on the outcomes of the BRM$^*$, and hence the algorithm is a two-stage algorithm. We also show that the file-copy problem is related to the group testing problem which has been widely studied in the literature.

## 2 The $\overline{3}$-Problem

We first show that $BRM^*$ can differentiate $d \le 2$ from $d > 2$. Define $\Delta^* = \{\Delta_X : \Delta_X \ne 0, \Delta_X \ne \Delta_F\}$. We call two $\Delta$-values *complementary* if they add up to $\Delta_F$.

Lemma 1. *$BRM^*$ has the following properties:*

1) *$d = 0$ if and only if $\Delta^* = \varnothing$ and $\Delta_F = 0$.*
2) *$d = 1$ if and only if $\Delta^* = \varnothing$ and $\Delta_F \ne 0$.*
3) *$d = 2$ if and only if $\Delta^*$ either contains a single value or two complementary values.*

Proof.

1) This is obvious.
2) The only if part is trivial. Now we prove the if part by contradiction. Assume that $d > 1$. Then $D$ contains at least two members, say, $a$ and $b$. Then $BRM^*$ has one row which contains either $a$ or $b$ but not both. The feedback of this row is neither 0 nor $\Delta_F$. This contradicts that $\Delta^* = \varnothing$. Therefore $d \le 1$. From condition 1), $d$ must be 1.
3) We first prove the only if part. Suppose that $D = \{a, b\}$. Then there are only two candidates $\Delta_a$ and $\Delta_b$ as members of $\Delta^*$ . Since $\Delta_a + \Delta_b = \Delta_F$, $\Delta^*$ either contains a single value or two complementary values.

   Now we prove the if part. From conditions 1) and 2), we only need to prove that $d$ cannot be greater than two. Suppose to the contrary that $D$ contains at least three members, say, $a$, $b$, and $c$. Since $a$ and $b$ are different, there exists a row in $BRM^*$ which contains either $a$ or $b$ but not both. Let $X(a, b)$ denote the $\Delta$-value of this row and let $Y(a, b)$ denote the sum of $\Delta$-values in this row other than $\Delta_a$ and $\Delta_b$. Then $X(a, b)$ is either $\Delta_a + Y(a, b)$ or $\Delta_b + Y(a, b)$. Similarly, we can define $X(b, c)$, $Y(b, c)$, and $X(c, a)$, $Y(c, a)$ such that $X(b, c) = \Delta_b + Y(b, c)$ or $\Delta_c + Y(b, c)$ and $X(c, a) = \Delta_c + Y(c, a)$ or $\Delta_a + Y(c, a)$. We show that any of the eight combinations of $X(a, b)$, $X(b, c)$, and $X(c, a)$ will yield two noncomplementary $\Delta$-values or three distinct values. The eight combinations form the following two general patterns.

   *Pattern 1.* Each of $\Delta_a$, $\Delta_b$, and $\Delta_c$ is chosen once. For example, $X(a, b) = \Delta_a + Y(a, b)$, $X(b, c) = \Delta_b + Y(b, c)$, and $X(c, a) = \Delta_c + Y(c, a)$. Clearly, these three $\Delta$-values are distinct.

   *Pattern 2.* One of $\Delta_a$, $\Delta_b$, and $\Delta_c$ is chosen twice, another chosen once. For example, $X(a, b) = \Delta_a + Y(a, b)$, $X(b, c) = \Delta_b + Y(b, c)$, and $X(c, a) = \Delta_a + Y(c, a)$. Clearly, $X(a, b)$ and $X(c, a)$ are not complementary as both contain $\Delta_a$. If they are not equal, then we are done. If they are, then $Y(a, b) = Y(c, a)$. Hence, $X(a, b)$ and $X(b, c)$, which are clearly distinct, are not complementary as both miss $\Delta_c$.

   Therefore, in either case, $\Delta^*$ contains at least two noncomplementary values, which contradicts the assumption. Hence $d \le 2$. □

The above lemma immediately leads to the following result.

Corollary 1. *$d > 2$ if and only if $\Delta^*$ contains at least two distinct noncomplementary values.*

The next theorem states the main result of this paper.

Theorem 1. *The $\overline{3}$-problem can be solved by using $BRM^*$ plus one more comparison on a single page after observing the feedbacks of $BRM^*$.*

Proof. By Corollary 1, we know whether $d \le 2$ after observing $BRM^*$. If it is the $\overline{2}$-problem, the disagreeing pages can be identified by $BRM^*$ from Lemma 1. Therefore, it suffices to consider $d = 3$. Let $D = \{a, b, c\}$ and let $A$, $B$, $C$ denote $\Delta_a$, $\Delta_b$, $\Delta_c$, respectively.

Without loss of generality, assume that $n = 2^N - 1$ since we can always add imaginary pages assumed to be agreeing, whose inclusion in a comparison can be omitted without affecting the feedback(a comparison involving only imaginary pages can be skipped with 0 recorded as its feedback). By Corollary 1, $\Delta^*$ contains at least two noncomplementary values $u$ and $v$ (distinct). We will consider two cases depending on whether $\Delta^*$ contains a third value noncomplementary to either $u$ or $v$. For $\Delta' \subseteq \Delta^*$, let $I(\Delta')$ denote the $N$-vector whose bit $i$ is 1 if and only if row $i$ of $BRM$ yields a delta-value in $\Delta'$, and is 0 otherwise. $I(\Delta')$ could be interpreted as the page which appears in all the rows of $BRM$ that yield delta-values in $\Delta'$. Also, let $\overline{z}$ denote the value $\Delta_F + z$.

**Case 1.** $\Delta^*$ contains a third value $w$ noncomplementary to either $u$ or $v$. There are eight possibilities for the set $\{u, v, w\}$, which can be classified into two groups.

**Subcase 1(i)**. $u + v + w \equiv \Delta_F$. There are four possibilities which can be differentiated by one more comparison on $I(\Delta_F, u, v, w)$. Let $\Delta_I$ denote the feedback of this comparison. Then $\Delta_I$ for each case is given below.

| $u$ | $v$ | $w$ | $\Delta_I$ |
|---|---|---|---|
| $A$ | $B$ | $C$ | $0$ |
| $A$ | $A+B$ | $A+C$ | $u$ |
| $A+B$ | $A$ | $A+C$ | $v$ |
| $A+B$ | $A+C$ | $A$ | $w$ |

To see this, note that in the first case $I(\Delta_F, u, v, w) \ne a$ for vector $a$ cannot have a 1-bit in those comparisons not involving $a$, but $I(\Delta_F, u, v, w)$ has a 1-bit in those comparisons with feedback $v(w)$, which involves $b(c)$ but not $a$. Similarly, $I(\Delta_F, u, v, w) \ne b$ or $c$. Hence, $\Delta_I = 0$. In the second case, $I(\Delta_F, u, v, w)$ has a 1-bit only in those comparisons involving $a$, hence $I(\Delta_F, u, v, w) = a$. Similarly, $I(\Delta_F, u, v, w) = a$ in the third and fourth cases.

**Subcase 1(ii)**. $u + v + w \equiv 0$. Again, there are four possibilities which can be differentiated by one more comparison on $I(\Delta_F, \overline{u}, \overline{v}, \overline{w})$.

| $u$ | $v$ | $w$ | $\Delta_I$ |
|---|---|---|---|
| $A+B$ | $A+C$ | $B+C$ | $0$ |
| $A$ | $B$ | $A+B$ | $\overline{w}$ |
| $A$ | $A+B$ | $B$ | $\overline{v}$ |
| $A+B$ | $A$ | $B$ | $\overline{u}$ |

**Case 2**. $\Delta^*$ does not contain a third noncomplementary value. There are four possibilities for $\{u, v\}$ which can be differentiated by one more comparison on $I(\Delta_F, u, v)$.

$$\begin{array}{ccc} u & v & \Delta_I \\ A & B & 0 \\ A & A+B & u \\ A+B & A & v \\ A+B & A+C & \Delta_F+u+v \end{array}$$

Once $A$, $B$, $C$ are determined, we can express every comparison feedback as one of the eight terms $O$, $A$, $B$, $C$, $A + B$, $A + C$, $B + C$, $A + B + C$. Correspondingly, we can label each comparison by a subset of $\{a, b, c\}$. Then $a(b, c)$ is just the binary number which has a 1-bit in those comparisons whose labels contain $A(B, C)$. □

EXAMPLE. $n = 16$

$$\begin{array}{cc} D:\{6,\ 10,\ 13\} & \{9,\ 10,\ 13\} \\ u & \Delta_F \\ v & u \\ w & v \\ \overline{w} & \overline{v} \end{array}$$

Suppose $D = \{6, 10, 13\}$. Then the sequence of $\Delta$-values in BRM is $(u, v, w, \overline{w})$ with $u + v + w \equiv 0$. This is Subcase 1(ii). The additional comparison is $I(\Delta_F, \overline{u}, \overline{v}, \overline{w},) = (0, 0, 0, 1) = 1$ with $\Delta_I = 0$. Therefore, $u = A + B$, $v = A + C$, $w = B + C$, $\overline{w} = A$. Consequently, $a = (1, 1, 0, 1) = 13$, $b = (1, 0, 1, 0) = 10$, $c = (0, 1, 1, 0) = 6$.

Suppose $D = \{9, 10, 13\}$. Then the sequence of $\Delta$-values in BRM is $(\Delta_F, u, v, \overline{v})$. This is Case 2. The additional comparison is $I(\Delta_F, u, v) = (1, 1, 1, 0) = 14$ with $\Delta_I = 0$. Therefore, $u = A$, $v = B$, and $\overline{v} = A + C$. Consequently, $a = (1, 1, 0, 1) = 13$, $b = (1, 0, 1, 0) = 10$, $c = (1, 0, 0, 1) = 9$.

Note that the conditions given for identifying $|D| = 0, 1, 2, 3$ disagreeing pages are all necessary and sufficient. Therefore, when none of these conditions is met, we know $|D| \geq 4$ and our underlying assumption $|D| \leq 3$ is wrong.

## 3 GROUP TESTING

In the group testing problem, we have a set $I$ of items each of which is either good or defective. The problem is to identify all the defective items with a minimum number of group tests. Assume that item $i$ has a parameter $\theta_i$, where $\theta_i = 0$ if and only if item i is good. A group test can be conducted on any arbitrary subset $S \subseteq I$ with the feedback $\sum_{i \in S} \theta_i$. In different group testing models, the sum means different things, and various restrictions are placed on the $\theta_i$. In the *residual model* ([6, Section 6.2]), $\theta_i$ is nonnegative and $\Sigma$ is the ordinary sum. If $S' \subset S$ and $\sum_{i \in S'} \theta_i < \sum_{i \in S} \theta_i$, then we can deduce that $S \setminus S'$ contains a defective item. We can extend the residual model to allow negative $\theta_i$. Then $\sum_{i \in S'} \theta_i \neq \sum_{i \in S} \theta_i$ implies $S \setminus S'$ contains a defective item. However, some information may be lost. For example, in case 2 of Section 2, $u + v < \Delta_F$ for the first pattern and $u + v > \Delta_F$ for the last pattern if $\theta_i$ is known to be nonnegative.

By interpreting the pages as items, the disagreeing pages as the defective items, and the comparisons as the tests, the file-copy problem has a similar flavor as the extended residual group testing problem except that $\Sigma$ is a modulo-2 sum, and the additional assumption that the masking errors are negligible. Note that the information provided by the modulo-2 sum is strictly less than the ordinary sum (but incomparable to the Boolean sum). Thus, any algorithm for the file-copy problem is an algorithm for the extended residual group testing problem if the masking errors can be ignored.

## 4 SOME CONCLUDING REMARKS

We could be more specific about the set $\Delta_k$. For the $\overline{2}$-problem, we need to assume that elements in $\Delta_1$ are distinct. For the $\overline{3}$-problem, we need to assume the same for $\Delta_2$.

A referee commented that our result also "follows (but not obviously) from the concept of null combinatorics related to Theorem 1 in [7]" which he or she noted "has not been published to my knowledge; so the result in the Hwang and Wan paper is new as to published result."

We thank two referees for excellent comments.

## REFERENCES

[1] J.J. Metzner and M.A. Abidi, "Remote Comparison and Correction of Duplicate Data Files," *Proc. Nat'l Telecomm. Conf.*, pp. 59.4.1-59.4.4, Nov. 1979.

[2] W. Fuchs, K.L. Wu, and J.A. Abraham, "Low-Cost Comparison and Diagnosis of Large Remotely Located Files," *Proc. Symp. Reliability Distributed Software and Database Systems*, pp. 67-73, Los Angeles, 1986.

[3] J.J. Metzner, "A Parity Structure for Large Remotely Located Replicated Data Files," *IEEE Trans. Computers*, vol. 32, no. 8, pp. 727-730, Aug. 1983.

[4] J.M. Metzner, "Efficient Replicated Remote File Comparison," *IEEE Trans. Computers*, vol. 40, no. 5, pp. 651-660, May 1991.

[5] D. Barbar, H. Garcia-Molina, and B. Feijoo, "Exploiting Symmetries for Low-Cost Comparison of File Copies," *Proc. Eighth Int'l Conf. Distributed Computing Systems*, June 1988.

[6] D.Z. Du and F.K. Hwang, *Combinatorial Group Testing and Its Applications.* World Scientific, 1993.

[7] J.J. Metzner and E.J. Kapturowski, "A General Decoding Technique Applicable to Replicated File Disagreement Location and Concatenated Code Decoding," *IEEE Trans. Information Theory*, vol. 36, pp. 911-917, July 1990.