# A Bound on the Overlap of Same-Sized Subsets

Olof Sivertsson, Pierre Flener, and Justin Pearson
Department of Information Technology
Uppsala University, Box 337, SE – 751 05 Uppsala, Sweden
Contact: Pierre.Flener@it.uu.se

**Abstract**

We prove a new lower bound on the number of shared elements of any pair of same-sized subsets drawn from a given set.

## 1 Introduction

Our objective is to determine a lower bound on the number of shared elements of any two of a number of subsets of the same size drawn from a given set. We first derive a bound from an existing result in extremal set theory and then establish a tighter bound, using a double-counting argument. This tighter bound has yielded crucial improvements in the construction of financial portfolio designs [2] in the credit derivatives market.

A known application of double counting leads to an optimal lower bound on the size of a set $B$ from which we can draw $v$ subsets of size $r$ such that any two of them share at most $\lambda$ elements. Write these $v$ subsets as $V_1, \ldots, V_v$ with $\lambda = \max_{i \neq j} |V_i \cap V_j|$. Then:

**Theorem 1** (Corrádi 1969 [1, 4])**.** *Let $V_1, \ldots, V_v$ be $r$-element sets and let $B$ be their union. If $|V_i \cap V_j| \leq \lambda$ for all $i \neq j$, then*

$$|B| \geq \frac{r^2 v}{r + (v-1)\lambda}. \tag{1}$$

However, we here actually know the set $B$ and are instead interested in the number $\lambda$ of shared elements of any two of $v$ subsets of size $r$ drawn from $B$. Rearranging (1) and taking $|B| = b$ gives the following lower bound:

$$\lambda \geq \frac{r(rv - b)}{b(v-1)}. \tag{2}$$

This lower bound is not always exact. For example, with $v = 9$, $b = 8$, and $r = 3$, we obtain $\lambda \geq \frac{57}{64}$, hence $\lambda \geq 1$. However, it is not difficult to show [2] that there are no 9 subsets of size 3 in an 8-element set such that any two of

them share at most $\lambda = 1$ element. In fact, this problem instance is at best solved with $\lambda = 2$; some of the sets of such optimal solutions share only one element.

Worse, the right-hand-side expression of (2) is negative when $b > rv$, that is when more elements are available than needed. This suggests that a tighter lower-bound expression ought to exist. For example, with $v = 2$, $b = 8$, and $r = 3$, we obtain $\lambda \geq -\frac{3}{4}$, hence $\lambda \geq 0$. Obviously, it *is* possible to construct 2 subsets of size 3 from an 8-element set that such that any two of them share at most $\lambda = 0$ elements, but a good lower-bound expression should evaluate to zero for such trivial problem instances.

When $x = qy + s$ denotes a division, with $x, y, q, s \in \mathbb{Z}$ and $0 \leq s < y$, then we define $\mathrm{mod}(x, y) = s$.

The rest of this paper is organised as follows. In Section 2, we state and prove a new lower bound on $\lambda$. Finally, in Section 3, we prove that the new lower bound is both tighter than the one of (2) and that its expression is never negative.

# 2 A Tighter Bound

Let $V_1, \ldots, V_v$ be $r$-element sets and $B$ be their union, with $b = |B|$. If $|V_i \cap V_j| \leq \lambda$ for all $i \neq j$, then we shall prove (see Theorem 2 below) that

$$\lambda \geq \frac{\left\lceil \frac{rv}{b} \right\rceil^2 \mathrm{mod}\,(rv, b) + \left\lfloor \frac{rv}{b} \right\rfloor^2 (b - \mathrm{mod}(rv, b)) - rv}{v(v-1)}. \tag{3}$$

The proof is structured as follows. We first prove a simple number-theoretic result in Lemma 1 and then use it to prove Lemma 2, which gives a lower bound on the sum of the squares of a non-empty sequence of $n$ natural numbers in terms of their sum and $n$. Finally, using Lemma 2 and well-known results (Lemma 3) on replication numbers (Definition 1), we prove (3) in Theorem 2.

**Lemma 1.** *Let $x \in \mathbb{Z}$ and $y \in \mathbb{Z}^+$. Then*

$$x = \left\lceil \frac{x}{y} \right\rceil \mathrm{mod}\,(x, y) + \left\lfloor \frac{x}{y} \right\rfloor (y - \mathrm{mod}(x, y)).$$

*Proof.* If $y \mid x$ then trivially the right-hand side is $x$. Else, letting $q$ and $s$ be the quotient and remainder of $\frac{x}{y}$, the right-hand side is

$$(q+1)s + q(y - s) = qy + s = x$$

thereby establishing the stated equality. □

**Lemma 2.** *Let $x_1, x_2, \ldots, x_n$ be a non-empty sequence of natural numbers with*

$$\sum_{i=1}^{n} x_i = a. \tag{4}$$

*Then*

$$\sum_{i=1}^{n} x_i^2 \geq \left\lceil \frac{a}{n} \right\rceil^2 \bmod (a,n) + \left\lfloor \frac{a}{n} \right\rfloor^2 (n - \bmod(a,n)).$$

*Proof.* Intuitively, if there are $x_j$ and $x_k$ among the $x_i$ with $x_j < x_k - 1$, then replacing them with $x_j + 1$ and $x_k - 1$ keeps the sum of the $x_i$ constant, but decreases the sum of their squares, namely by $2(x_k - x_j - 1)$. By Lemma 1, we can write

$$\sum_{i=1}^{n} x_i = a = \left\lceil \frac{a}{n} \right\rceil \bmod (a,n) + \left\lfloor \frac{a}{n} \right\rfloor (n - \bmod(a,n)).$$

Since $\bmod(a,n) + (n - \bmod(a,n)) = n$ we can think of $\bmod(a,n)$ occurrences of $\lceil \frac{a}{n} \rceil$ and $n - \bmod(a,n)$ occurrences of $\lfloor \frac{a}{n} \rfloor$ as choices for the $x_i$:

$$\sum_{i=1}^{n} x_i = \underbrace{\left\lceil \frac{a}{n} \right\rceil + \cdots + \left\lceil \frac{a}{n} \right\rceil}_{\bmod(a,n) \text{ times}} + \underbrace{\left\lfloor \frac{a}{n} \right\rfloor + \cdots + \left\lfloor \frac{a}{n} \right\rfloor}_{n - \bmod(a,n) \text{ times}} .$$

With this choice of the variables $x_i$ we get

$$\sum_{i=1}^{n} x_i^2 = \left\lceil \frac{a}{n} \right\rceil^2 \bmod (a,n) + \left\lfloor \frac{a}{n} \right\rfloor^2 (n - \bmod(a,n))$$

and we have proved that the stated lower bound is indeed attainable. □

**Definition 1.** *The* replication number *or* degree *of a point $x$ in a family $\mathcal{F}$, denoted by $d(x)$, is the number of members of $\mathcal{F}$ containing $x$.*

**Lemma 3** ([4], page 16)**.** *Let $\mathcal{F}$ be a family of subsets of some set $X$. Then*

$$\sum_{x \in Y} d(x) = \sum_{A \in \mathcal{F}} |Y \cap A| \quad \text{for any } Y \subseteq X \tag{5}$$

$$\sum_{x \in X} d(x)^2 = \sum_{A \in \mathcal{F}} \sum_{x \in A} d(x) \tag{6}$$

**Theorem 2.** *Let $V_1, \ldots, V_v$ be $r$-element sets and $B$ be their union, with $b = |B|$. If $|V_i \cap V_j| \leq \lambda$ for all $i \neq j$, then*

$$\lambda \geq \frac{\left\lceil \frac{rv}{b} \right\rceil^2 \bmod (rv,b) + \left\lfloor \frac{rv}{b} \right\rfloor^2 (b - \bmod(rv,b)) - rv}{v(v-1)}$$

*Proof.* The sets $V_1, \ldots, V_v$ can be thought of as a boolean matrix with $v$ rows and $b$ columns with a 1 (respectively 0) in row $i$ and column $j$ meaning that $j \in V_i$ (respectively $j \notin V_i$). Using this point of view, $d(x)$ is the sum of the $x^{\text{th}}$ column.

By (5) of Lemma 3, we have for each $i \in \{1, \ldots, v\}$

$$\begin{aligned}\sum_{x \in V_i} d(x) &= \sum_{j=1}^{v} |V_i \cap V_j| \\ &= |V_i| + \sum_{j \neq i} |V_i \cap V_j| \\ &\leq r + (v-1)\lambda \end{aligned} \tag{7}$$

where the final inequality comes from the fact that no two subsets share more than $\lambda$ elements. Next sum over all the sets $V_i$ to get

$$\sum_{i=1}^{v} (r + (v-1)\lambda) \geq \sum_{i=1}^{v} \sum_{x \in V_i} d(x)$$

and use (6) of Lemma 3 to arrive at

$$v(r + (v-1)\lambda) \geq \sum_{x \in B} d(x)^2$$

which we reorder to get an expression for $\lambda$

$$\lambda \geq \frac{\sum_{x \in B} d(x)^2 - rv}{v(v-1)}. \tag{8}$$

By counting the number of ones in the matrix both column-wise and row-wise, we get

$$\sum_{x \in B} d(x) = rv$$

and we can then replace the sum in (8) with the lower bound from Lemma 2, yielding the stated lower bound. □

Essentially, the improvement thus comes from the fact that the lower bound on the sum of the squares of the degrees coming from convexity can be increased slightly by using the fact that the degrees are all integers.

# 3 Concluding Remarks

Note that when $b \mid rv$ the lower bound of Theorem 2 degenerates into the lower bound of (2).

Even this new lower bound is not always exact. Consider for example $v = 9$, $b = 8$, and $r = 3$: using (2) we get $\lambda \geq 0.890625$ while Theorem 2 gives $\lambda \geq 0.91\bar{6}$. However, as stated in the introduction, this problem instance can only be solved with $\lambda \geq 2$. The lower bound of Theorem 2 is thus closer to the optimum than the one of (2), but there are still cases when it is too low.

Since inequality (7) degenerates into an equality when $v = 2$, the lower bound of Theorem 2 is always sharp in this case, while the lower bound of (2) is not sharp then. Neither bound is sharp for $v = 3$.

We now prove that the new lower bound of Theorem 2 satisfies our two requirements: it is tighter than the old lower bound of (2) and its expression is never negative. It is an open question whether there is a tighter lower bound that is easy to compute.

**Proposition 1.** *The lower bound of Theorem 2 is tighter than the lower bound of (2).*

*Proof.* We can apply Jensen's inequality

$$\sum_{i=1}^{b} x_i^2 \geq \frac{1}{b}\left(\sum_{i=1}^{b} x_i\right)^2$$

to the sum of squares in Theorem 2 to get

$$\begin{aligned}\lambda &\geq \frac{\left\lceil \frac{rv}{b} \right\rceil^2 \bmod (rv,b) + \left\lfloor \frac{rv}{b} \right\rfloor^2 (b - \bmod(rv,b)) - rv}{v(v-1)} \\ &\geq \frac{\frac{1}{b}\left(\left\lceil \frac{rv}{b} \right\rceil \bmod (rv,b) + \left\lfloor \frac{rv}{b} \right\rfloor (b - \bmod(rv,b))\right)^2 - rv}{v(v-1)}\end{aligned}$$

Now use Lemma 1 to rewrite this as

$$\lambda \geq \frac{\frac{(rv)^2}{b} - rv}{v(v-1)} = \frac{r(rv-b)}{b(v-1)}.$$

So the right hand side of Theorem 2 is at least as large as the one of (2).

Now consider for example $v = 10$, $b = 350$, and $r = 100$: Theorem 2 gives $\lambda \geq 21.\bar{1}$ (and a solution does exist with $\lambda = 22$, see [2]) whereas the bound of (2) only gives $\lambda \geq 20.63$. Hence the new bound is sometimes strictly tighter and we can now be sure no solution with $\lambda = 21$ exists, which is a claim that required a separate proof previously [3]. □

We next establish that the new lower bound satisfies our second requirement, namely that its expression is never negative.

**Proposition 2.** *The expression of the lower bound of Theorem 2 is never negative.*

*Proof.* By Lemma 1 we have

$$\left\lceil \frac{rv}{b} \right\rceil \bmod (rv,b) + \left\lfloor \frac{rv}{b} \right\rfloor (b - \bmod(rv,b)) = rv.$$

Consider the left-hand side above as a summation of $b$ terms, with $\bmod(rv,b)$ ceilings and $b - \bmod(rv,b)$ floors. Since these terms and coefficients are all non-negative integers, squaring the terms results in a larger or equal sum:

$$\left\lceil \frac{rv}{b} \right\rceil^2 \bmod (rv,b) + \left\lfloor \frac{rv}{b} \right\rfloor^2 (b - \bmod(rv,b)) \geq rv.$$

Hence the numerator of the bound of Theorem 2 is non-negative, and since its denominator $v(v-1)$ is positive, the fraction is never negative. □

The difference between the right-hand-side expressions of (2) and (3) can get arbitrarily large, especially since the one of (2) keeps decreasing into negative values when $b > rv$.

By choosing all subsets of size $\lambda + 1$ from each row (keeping the previous matrix analogy), which contains $r$ elements, and comparing this with choosing all subsets, also of size $\lambda + 1$, from the whole set $B$ we arrive at

$$v\binom{r}{\lambda+1} \leq \binom{b}{\lambda+1}$$

which holds because no two chosen subsets are equal since by definition no two subsets overlap by more than $\lambda$ elements. (So we could have used subsets of any size $\lambda + i \leq r$ but that would result in a weaker bound below.) We can rewrite this as

$$\begin{aligned} v &\leq \frac{b(b-1)\cdots(b-\lambda)}{r(r-1)\cdots(r-\lambda)} \\ &\leq \left(\frac{b-\lambda}{r-\lambda}\right)^{\lambda+1} \\ &= e^{(\lambda+1)\ln\frac{b-\lambda}{r-\lambda}} \end{aligned}$$

Since $\ln(1+x) \leq x$ we can use this to get

$$v \leq e^{(\lambda+1)\frac{b-\lambda}{r-\lambda}}$$

which results in

$$\lambda \geq \frac{r\ln v - b + r}{\ln v + b - r}$$

It turns out that this bound seems better than the one of Theorem 2 when $v$ becomes closer to its maximum value $\binom{b}{r}$. For example, when $b = 350$ and $r = 100$, this bound is better than the one of Theorem 2 when $v$ is larger than roughly $\sqrt{\binom{350}{100}}$, but more work and comparisons should be made. It should also be investigated if the large approximations in the steps to retrieve this bound could be limited to smaller approximations to get a better bound.


# Acknowledgements

We thank Ian P. Gent, Luis G. Reyna, and Nic Wilson for fruitful discussions. We also thank the anonymous referees for their useful comments, which significantly improved the presentation of this work.

# References


[1] K. Corrádi. Problem at Schweitzer competition. *Mat. Lapok*, 20:159–162, 1969.

[2] P. Flener, J. Pearson, L. G. Reyna, and O. Sivertsson. Design of financial CDO squared transactions using constraint programming. *Constraints*, 12(2):179–205, April 2007.

[3] I. P. Gent and N. Wilson. Minimising pairwise intersections problem. Personal communications to Justin Pearson, October 2004.

[4] S. Jukna. *Extremal Combinatorics*. Springer-Verlag, 2001.