Theorems on redundancies in distributed computing

Sahasrajit Sarmasarkar, Harish Pillai
Department of Electrical Engineering
Indian Institute of Technology, Bombay

Abstract

We consider the problem of redundancies in distributed computing where a master server wishes to compute some tasks and is provided a few child servers to compute. We consider a noisy environment where some child servers may fail to communicate their results to the master. We attempt to distribute tasks to the servers so that master is able to get the results for most of the tasks even if a few servers fail to communicate. We formulate some conditions on the distribution such that the number of tasks returned is the maximum and also show that constructions using "Balanced Incomplete Block Design" [1] attains this optimality.

I. Preliminaries

Given a set of n jobs(tasks) and c servers, we would like to distribute k jobs to each server with no two same jobs in any server along with each job appearing in exactly r distinct servers. This would imply that $n \times r = k \times c$. We analyse the number of distinct jobs returned when any subset of x servers return with equal probability.

A. Notations and symbols used

Let us denote the jobs as $\{a_1, a_2, ..., a_n\}$ and the servers as $\{s_1, s_2, ..., s_c\}$. Let us try to analyse the number of distinct jobs (d) when any subset of servers (S) return i.e. are able to communicate their results to the master.

Consider all jobs in the subset of servers S. There would be some jobs appearing multiple number of times in the subset of servers S. Thus, we say job a_i occurred $n_{i,S}$ times in subset S for any $i \leq n$. Note that $n_{i,S}$ could possibly be zero for some values of i as well. Note that $\mathbb{1}_E$ denotes the indicator random variable corresponding to event E. Let us denote the distribution of jobs in servers by $\mathcal{D}$.

We are interested in those jobs which have $n_{i,S} > 1$ in subset S and thus claim that number of distinct jobs (d) returned from subset of workers S is $(k \times x - \sum_{i=1}^{n}(n_{i,S} - 1)\mathbb{1}_{n_{i,S}>1})$ if the cardinality of S is x. For a given distribution of jobs to servers $\mathcal{D}$, we denote the expectation in the number of distinct jobs returned when any set of x servers return uniformly at random by $\mathbb{E}_{\mathcal{D},x}[d]$ ad the corresponding variance by $\sigma_{\mathcal{D},x}[d]$.

We first state our main theorem below on attaining the least variance on the number of distinct jobs when any set of x servers (chosen uniformly at random) are able to communicate their results to the master.

Theorem 1. *Consider a distribution D for a given n, k, r, c satisfying the conditions in I such that every pair of jobs are present together in exactly l or $l+1$ servers for some integer l. This distribution D would have the least variance in the number of distinct jobs d returned when any set of x servers return uniformly at random amongst all the distributions satisfying I.*

Also under a special constraint $n = c$, the above theorem can also be stated as follows.

Theorem 2. *Consider a distribution for a given n, k, k, n satisfying the conditions in I such that every pair of servers have l or $l+1$ common jobs for some positive integer l. Then the distribution has the least variance on the number of distinct jobs when any set of x servers return uniformly at random amongst all distributions satisfying property I.*

II. THEOREMS AND PROOFS

Theorem 3. *For a given n, k,c and r, any distribution which satisfies the condition mentioned in I has the same expectation of distinct jobs returned for every $x \leq c$ assuming S can be any subset of the servers of cardinality x with equal probability. Also the expectation can be computed as*

$$\mathbb{E}_{\mathcal{D},x}[d] = k \times x - \frac{n\sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{x}} = k \times x - n.\frac{r \times \binom{c-1}{x-1} + \binom{c-r}{x} - \binom{c}{x}}{\binom{c}{x}} \tag{1}$$

where d denotes the number of distinct jobs returned when any set of x servers return and the job distribution is denoted by $\mathcal{D}$. Additionally, the variance of the number of distinct jobs can be written as:

$$\sigma_{\mathcal{D},x}(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\|S|=x}}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}(n_{j,S}-1)\mathbb{1}_{n_{j,S}>1}}{\binom{c}{x}} - \left(\frac{n\sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{x}}\right)^2 \tag{2}$$

Also under the constraint $n = c$, we obtain $\mathbb{E}_{\mathcal{D},x}[d] = n.\left(1 - \frac{\binom{n-x}{r}}{\binom{n}{r}}\right)$

Proof. Let us attempt to compute the expectation of distinct jobs(d) returned.

$$\begin{aligned}\mathbb{E}_{\mathcal{D},x}[d] =& \frac{\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\|S|=x}}(k\times x - \sum_{i=1}^{n}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1})}{\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\|S|=x}} 1}\\ \overset{(a)}{=} & k\times x - \frac{\sum_{i=1}^{n}\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\|S|=x}}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}}{\binom{c}{x}}\end{aligned}$$

(a) follows on interchanging summations.

Now we show $\sum_{S\subset\{s_1,s_2,...,s_c\},|S|=x}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}$ is same for every distribution $\mathcal{D}$ for every job a_i under a given x.

$$\begin{aligned}\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\|S|=x}}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1} &\overset{(a)}{=} \sum_{t=1}^{r-1}\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\|S|=x,n_{i,S}=t+1}} t\\ &\overset{(b)}{=} \sum_{t=1}^{r-1} t\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\|S|=x,n_{i,S}=t+1}} 1\\ &\overset{(c)}{=} \sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}\end{aligned}$$

(a) follows since $n_{i,S}$ can take value only from $\{0,1,2,...r\}$. (b) follows since t is a constant for the second summation whereas (c) follows from the argument below.

Note that $\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\ |S|=x, n_{i,S}=t+1}} 1$ denotes the number of subsets of servers of cardinality x which have job a_i occurring $(t+1)$ times.

This would be unique irrespective of distribution since job a_i would occur in exactly r servers. This the subsets should have $(t+1)$ servers from these r servers and all the other $(x-t-1)$ servers from the remaining $(c-r)$ servers.

Hence, there are $\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}$ such subsets.

Thus, we have $\mathbb{E}_{\mathcal{D},x}[d] = k \times x - \frac{n\sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{x}}$

This would imply that the mean number of distinct elements returned for a given x remains the same irrespective of job distribution chosen as long as it follows the rules in I.

Note that using the idea of summation of series using coefficients of binomial expressions we can show that $\frac{n\sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{x}} = r \times \binom{c-1}{x-1} + \binom{c-r}{x} - \binom{c}{x}$.

This can be argued using the following binomial expressions

$r.(1+y)^{r-1} + 1/y - \frac{(1+y)^r}{y} = \sum_{t=0}^{r-1} t.\binom{r}{t+1}$ and $(1+y)^{c-r} = \sum_{u=0}^{c-r}\binom{c-r}{u}y^u$. Thus, $\frac{n\sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{x}}$ becomes the coefficient of y^{x-1} in $r(1+y)^{c-1} + \frac{(1+y)^{c-r}}{y} - \frac{(1+y)^c}{y}$ which equals $r \times \binom{c-1}{x-1} + \binom{c-r}{x} - \binom{c}{x}$

$$
\begin{aligned}
\sigma_{\mathcal{D},x}(d) &= \sigma_{\mathcal{D},x}(k \times x - \sum_i (n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}) \\
&\overset{(d)}{=} \sigma_{\mathcal{D},x}(\sum_i (n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}) \\
&= \frac{\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\ |S|=x}} (\sum_{i=1}^{n}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1})^2}{\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\ |S|=x}} 1} - \left(\frac{\sum_{S\subset\{s_1,s_2,...,s_c\};|S|=x} (\sum_{i=1}^{n}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1})}{\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\ |S|=x}} 1}\right)^2 \\
&\overset{(e)}{=} \frac{\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\ |S|=x}} (\sum_{i=1}^{n}\sum_{j=1}^{n}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}(n_{j,S}-1)\mathbb{1}_{n_{j,S}>1})}{\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\ |S|=x}} 1} - \left(\frac{\sum_{i=1}^{n}\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\ |S|=x}} ((n_{i,S}-1)\mathbb{1}_{n_{i,S}>1})}{\binom{c}{x}}\right)^2 \\
&\overset{(f)}{=} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{S\subset\{s_1,s_2,...,s_c\};\\ |S|=x}} (n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}(n_{j,S}-1)\mathbb{1}_{n_{j,S}>1}}{\binom{c}{x}} - \left(\frac{n\sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{x}}\right)^2
\end{aligned}
$$

Let us attempt to compute variance of the number of distinct jobs.

(d) follows since $\sigma(c-X) = \sigma(X)$ where c is a constant and X is the random variable. The first term in (e) follows since $(\sum_i b_i)^2 = \sum_i \sum_j b_i b_j$. The first term in (f) follows from interchange of summations whereas the second term follows from (c) in previous equation.

□

Claim 4. *Consider a pair (i_m, j_m) such that jobs a_{i_m} and a_{j_m} are present together in exactly m servers for every positive integer m. Now we define $f(i,j)$ as.*

$$f(i,j) = \sum_{\substack{S\subset\{s_1,s_2,\ldots,s_c\};\\ |S|=x}} (n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}(n_{j,S}-1)\mathbb{1}_{n_{j,S}>1} \tag{3}$$

.

We can say that

$$f(i_m,j_m) - f(i_{m-1},j_{m-1}) = \left[\binom{(c-2)}{(x-1)} - 2\binom{(c-r-1)}{(x-1)} + \binom{(c-2r+m-1)}{(x-1)}\right] \tag{4}$$

Proof. We can show that the function $f(i_m, j_m)$ is just a function of m. This is because which m servers have jobs a_{i_m} and a_{j_m} does not make any difference in the summation in $f(i_m, j_m)$.

Now consider $f(i_{m-1}, j_{m-1}) = \sum_{\substack{S\subset\{s_1,s_2,\ldots,s_c\};\\ |S|=x}} (n_{i_{m-1},S}-1)\mathbb{1}_{n_{i_{m-1},S}>1}(n_{j_{m-1},S}-1)\mathbb{1}_{n_{j_{m-1},S}>1}$.

Let us try to compute the difference between the functions and for this sake let us assume that servers $s_1, \ldots s_{m-1}$ are shared by both the pairs. s_m is shared by only the pair (a_{i_m}, a_{j_m}), however s_m only contains $a_{i_{m-1}}$, and server s_{m+1} contains only $a_{j_{m-1}}$ but none of the elements in the other pair.

Let $s_{k_1}, s_{k_2}, \ldots s_{k_{r-m}}$ denote the servers containing only one element from each pair say a_{i_m} and $a_{i_{m-1}}$, however $s_{l_1}, s_{l_2}, \ldots s_{l_{r-m}}$ denote the servers containing only one element from each pair say a_{j_m} and $a_{j_{m-1}}$. This is because job a_{i_m} occurs without a_{j_m} in $(r-m)$ such servers and similarly job $a_{i_{m-1}}$ occurs without $a_{j_{m-1}}$ in $(r-m+1)$ such servers which we have satisfied above.

Let us denote the remaining servers by R.

Now let us compare the difference between two functions for each subset of cardinality x.

Let us look at

$$(n_{i_m,S}-1)\mathbb{1}_{n_{i_m,S}>1}(n_{j_m,S}-1)\mathbb{1}_{n_{j_m,S}>1} - (n_{i_{m-1},S}-1)\mathbb{1}_{n_{i_{m-1},S}>1}(n_{j_{m-1},S}-1)\mathbb{1}_{n_{j_{m-1},S}>1} \tag{5}$$

for all subsets S of cardinality x.

1) If the set S neither contains s_m or s_{m+1}, then we can say that (5) goes to zero.

2) If the set S contains both s_m and s_{m+1}, then also we can say that (5) goes to 0.

Now consider a subset s which does not contain either s_m or s_{m+1} but this set has say $(\alpha+1)$ occurrences of $a_{i_{m-1}}$ and a_{i_m} and $(\beta+1)$ occurrences of $a_{j_{m-1}}$ and a_{j_m} for some $\alpha, \beta \geq 0$.

Now let us look at the set $s \cup \{s_m\}$. For this set (5) goes to $(\alpha+1)(\beta+1) - (\alpha+1)\beta$. Similarly, for the set $s \cup \{s_{m+1}\}$, (5) goes to $\alpha\beta - \alpha(\beta+1)$.

Thus the sum of (5) for these two subsets goes to 1.

We can also show that if set s contains zero occurrences of $a_{i_{m-1}}$ and a_{i_m} or zero occurrences of $a_{j_{m-1}}$ and a_{j_m}, the sum of (5) for sets $s \cup \{s_m\}$ and $s \cup \{s_{m+1}\}$ goes to 0, however there are $2\binom{(c-r-1)}{(x-1)} - \binom{(c-2r+m-1)}{(x-1)}$ such subsets. This follows from the fact that such a subset can only exist if it is a subset of $\{s_{k_1}, s_{k_2}, \ldots, s_{k_{r-m}}\} \cup R$ or $\{s_{l_1}, s_{l_2}, \ldots, s_{l_{r-m}}\} \cup R$.

Hence, the function $f(i_m, j_m) - f(i_{m-1}, j_{m-1})$ goes to $\left[\binom{(c-2)}{(x-1)} - 2\binom{(c-r-1)}{(x-1)} + \binom{(c-2r+m-1)}{(x-1)}\right]$ □

Claim 5. *Consider any distribution of jobs in servers denoted by $\mathcal{D}$ satisfying the conditions in I. Let us define $n_{p,\mathcal{D}}$ as the number of pairs of jobs which appear together in exactly p servers. We can show that the variance can be written as*

$$\sigma_{\mathcal{D},x} = \frac{2.\sum_{p=0} n_{p,\mathcal{D}} f(i_p,j_p) + n\sum_{t=1}^{r-1} t^2 \binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{x}} - \left(\frac{n\sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{x}}\right)^2 \tag{6}$$

Proof. We know from equation (2) in Theorem 3 that

$$\sigma_{\mathcal{D},x}(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{S\subset\{s_1,s_2,\ldots,s_c\};\\|S|=x}}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}(n_{j,S}-1)\mathbb{1}_{n_{j,S}>1}}{\binom{c}{x}} - \left(\frac{n\sum_{t=1}^{r-1}t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{x}}\right)^2$$

Now consider the numerator in the first term of $\sigma_{\mathcal{D},x}(d)$.

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\substack{S\subset\{s_1,s_2,\ldots,s_c\};\\|S|=x}}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}(n_{j,S}-1)\mathbb{1}_{n_{j,S}>1}\\
=&2.\sum_{1\le i<j\le n}\sum_{\substack{S\subset\{s_1,s_2,\ldots,s_c\};\\|S|=x}}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}(n_{j,S}-1)\mathbb{1}_{n_{j,S}>1}\\
&+\sum_{1\le i=j\le n}\sum_{\substack{S\subset\{s_1,s_2,\ldots,s_c\};\\|S|=x}}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}(n_{j,S}-1)\mathbb{1}_{n_{j,S}>1}\\
\overset{(a)}{=}&2.\sum_{p=0}n_{p,\mathcal{D}}f(i_p,j_p)+\sum_{1\le i\le n}\sum_{\substack{S\subset\{s_1,s_2,\ldots,s_c\};\\|S|=x}}((n_{i,S}-1)\mathbb{1}_{n_{i,S}>1})^2\\
\overset{(b)}{=}&2.\sum_{p=0}n_{p,\mathcal{D}}f(i_p,j_p)+n\sum_{t=1}^{r-1}t^2\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}
\end{aligned}$$

Note the first term in (a) follows since $f(i,j)=\sum_{\substack{S\subset\{s_1,s_2,\ldots,s_c\};\\|S|=x}}(n_{i,S}-1)\mathbb{1}_{n_{i,S}>1}(n_{j,S}-1)\mathbb{1}_{n_{j,S}>1}$. The second term in (b) is computed using a very similar technique as used in theorem 3 for computation of $\sum_{\substack{S\subset\{s_1,s_2,\ldots,s_c\};\\|S|=x}}((n_{i,S}-1)\mathbb{1}_{n_{i,S}>1})^2$

□

Claim 6. *Recall the definition of $f(i_p,j_p)$ as defined in Claim 3. Then, the following can be said:*

- $\frac{f(i_{p+k_1},j_{p+k_1})-f(i_p,j_p)}{k_1} > \frac{f(i_p,j_p)-f(i_{p-k_2},j_{p-k_2})}{k_2}\ \forall\ k_1,k_2\in\mathbb{N}$.
- $\frac{f(i_{p+k_1},j_{p+k_1})-f(i_p,j_p)}{k_1} > \frac{f(i_{p+1},j_{p+1})-f(i_{p-k_2+1},j_{p-k_2+1})}{k_2}\ \forall\ k_1,k_2\in\mathbb{N}$.
- $\frac{f(i_{p+k_1+1},j_{p+k_1+1})-f(i_{p+1},j_{p+1})}{k_1} > \frac{f(i_p,j_p)-f(i_{p-k_2},j_{p-k_2})}{k_2}\ \forall\ k_1,k_2\in\mathbb{N}$.
- $\frac{f(i_{p+k_1},j_{p+k_1})-f(i_p,j_p)}{k_1} > \frac{f(i_{p+k_2},j_{p+k_2})-f(i_p,j_p)}{k_2}\ \forall\ k_1,k_2\in\mathbb{N}, k_1>k_2$.

Proof. Let us define $g(p)=f(i_p,j_p)$. Now consider

$$\begin{aligned}
\frac{g(p+k_1)-g(p)}{k_1} = \frac{\sum_{i=0}^{i=k_1-1}(g(p+i+1)-g(p+i))}{k_1} &\overset{(a)}{\ge} \frac{k_1(g(p+1)-g(p)}{k_1}\\
&\ge g(p+1)-g(p)
\end{aligned}$$

Now consider,

$$\begin{aligned}
\frac{g(p+k_1+1)-g(p+1)}{k_1} = \frac{\sum_{i=1}^{i=k_1}(g(p+i+1)-g(p+i))}{k_1} &\overset{(d)}{\ge} \frac{k_1(g(p+2)-g(p+1)}{k_1}\\
&\ge g(p+2)-g(p+1)
\end{aligned}$$

Note (a) and (d) follow since since $g(p+i+1)-g(p+i)\ge g(p+1)-g(p)\ \forall i\in\mathbb{N}$

Similarly,

$$\frac{g(p)-g(p-k_2)}{k_2}=\frac{\sum_{i=0}^{i=k_2-1}(g(p-i)-g(p-i-1))}{k_2}\overset{(b)}{\leq}\frac{k_2(g(p)-g(p-1)}{k_2}$$
$$\leq g(p)-g(p-1)$$

Similarly,

$$\frac{g(p+1)-g(p-k_2+1)}{k_2}=\frac{\sum_{i=0}^{i=k_2-1}(g(p-i+1)-g(p-i))}{k_2}\overset{(c)}{\leq}\frac{k_2(g(p+1)-g(p)}{k_2}$$
$$\leq g(p+1)-g(p)$$

Note (b) and (c) follow since $g(p-i)-g(p-i-1)\leq g(p)-g(p-1)\ \forall i\in\mathbb{N}$.
Thus we can say

$$\frac{g(p+k_1)-g(p)}{k_1}\geq\frac{g(p)-g(p-k_2)}{k_2}\forall k_1,k_2\in\mathbb{N}$$

and

$$\frac{g(p+k_1)-g(p)}{k_1}\geq\frac{g(p+1)-g(p-k_2+1)}{k_2}\forall k_1,k_2\in\mathbb{N}$$

$$\frac{g(p+k_1+1)-g(p+1)}{k_1}\geq\frac{g(p+1)-g(p-k_2+1)}{k_2}\forall k_1,k_2\in\mathbb{N}$$

Thus the first, second and third inequalities follow.
Now consider the following:

$$\frac{g(p+k_1)-g(p)}{k_1}\geq\frac{g(p+k_2)-g(p)}{k_2}\Leftrightarrow\frac{g(p+k_1)-g(p+k_2)}{k_1-k_2}\geq\frac{g(p+k_2)-g(p)}{k_2}$$

Note that since $k_1>k_2$, we have proven the inequality on R.H.S. □

Let us now state and prove Theorem 1.

Theorem. *Consider a distribution D for a given n,k,r,c satisfying the conditions in I such that any pair of jobs are present together in exactly l or $l+1$ servers for some integer l. This distribution D would have the least variance in the number of distinct jobs amongst all the distributions satisfying I.*

Proof. Now we know that $n_{p,D}=0$ only for $p\neq l,l+1$.
Let us consider another distribution D_1 which satisfies the constraints in I.
Recall the definition of $f(i_p,j_p)$ as defined in Claim 3. we

- Case 1: $n_{p,D_1}=0$ for $p\neq l,l+1$. Since we have the constraints $\sum_p n_{p,D_1}=\sum_p n_{p,D}=\binom{n}{2}$ and $\sum_p p.n_{p,D_1}=\sum_p p.n_{p,D}=c\binom{k}{2}$, we can say that $n_{p,D_1}=n_{p,D}\ \forall\ p$ which would imply distribution D_1 has same variance of distinct jobs as that of distribution D.
- Case 2: $n_{l,D_1}<n_{l,D}$ but $n_{l+1,D_1}\geq n_{l+1,D_1}$
  Let us denote $x_p=n_{p,D_1}-n_{p,D}\ \forall p$.
  Now we know that $\sum_p p.n_{p,D}=\sum_p p.n_{p,D_1}=n\binom{k}{2}$ implying $\sum_p p.x_p=0$. Similarly, we can also argue $\sum_p x_p=0$.
  Now we know that $x_p<0$ only for $p=l$. Let us denote $\sum\limits_{p<l}x_p=x$ and $\sum\limits_{p>l}=y$. Since $\sum\limits_p x_p=0$, we can say that $x_l=-(x+y)$.
  Thus $\sum\limits_p p.x_p=0\Leftrightarrow(x+y)l=\sum\limits_{p>l}p.x_p+\sum\limits_{p<l}p.x_p\Leftrightarrow\sum\limits_{p>l}x_p(p-l)=\sum\limits_{q<l}x_q(l-q)$.
  However, we know from Claim 6 that $\frac{f(i_p,j_p)-f(i_l,j_l)}{p-l}>\frac{f(i_l,j_l)-f(i_q,j_q)}{l-q}\forall p>l>q$.

These two results above would imply:

$$\sum_{p>l} x_p(f(i_p,j_p)-f(i_l,j_l)) > \sum_{q<l} x_q(f(i_l,j_l)-f(i_q,j_q)) \Leftrightarrow \sum_{p\neq l} x_p.f(i_p,j_p) - \sum_{p\neq l} x_p f(i_l,j_l) > 0$$
$$\Leftrightarrow \sum_{p\neq l} x_p.f(i_p,j_p) + x_l.f(i_l,j_l) > 0$$
$$\overset{(c)}{\Leftrightarrow} \sum_{p} x_p f(i_p,j_p) > 0$$
$$\Leftrightarrow \sum_{p} n_{p,D_1} f(i_p,j_p) > \sum_{p} n_{p,D} f(i_p,j_p)$$

Note (c) follows since $x_l = -\sum_{p\neq l} x_p$.

Now let us consider the numerator of the first term in $\sigma_{\mathcal{D},x}(d)$ as in theorem 6 which can be written as $2.\sum_p n_{p,\mathcal{D}} f(i_p,j_p) + n.\Big(\sum_{t=1}^{r-1} t^2 \binom{r}{(t+1)} \binom{(c-r)}{(x-t-1)}\Big)$.

Thus the inequality proven in the previous result would imply that distribution D_1 has higher variance of number of distinct jobs returned than that of distribution D.

- Case 3: $n_{l+1,D_1} < n_{l+1,D}$ but $n_{l,D_1} \geq n_{l,D}$
  Note this can be proven in a very similar way as that of Case 2. The entire proof could be done for $l+1$ instead of l
- Case 4: $n_{l+1,D_1} < n_{l+1,D}$ and $n_{l,D_1} < n_{l,D}$
  Let us denote $x_p = n_{p,D_1} - n_{p,D}\ \forall p$.
  Now we know that $\sum_p p.n_{p,D} = \sum_p p.n_{p,D_1} = c\binom{k}{2}$ implying $\sum_p p.x_p = 0$. Similarly, we can also argue $\sum_p x_p = 0$.
  Now we know that $x_p < 0$ only for $p = l, l+1$. Let us denote $\sum_{p<l} x_p = x_1 + x_2$ and $\sum_{p>l+1} x_p = y_1 + y_2$.
  Since $\sum_p x_p = 0$, we can say that $x_l = -(x_1+y_1)$ and $x_{l+1} = -(x_2+y_2)$ for some $x_1, y_1, x_2, y_2 \in \mathbb{N}$.
  Choose $y_p = x_p.\frac{x_1}{x_1+x_2}$ and $z_p = x_p.\frac{x_2}{x_1+x_2}$ for $p < l$. Choose $y_p = x_p.\frac{y_1}{y_1+y_2}$ and $z_p = x_p.\frac{y_2}{y_1+y_2}$ for $p > l+1$
  Thus $\sum_{p>l+1} y_p = -y_1$; $\sum_{p>l+1} z_p = -y_2$; $\sum_{p<l} y_p = -x_1$; $\sum_{p<l} z_p = -x_2$;
  Thus

$$\sum_p p.x_p = 0 \Leftrightarrow (x_1+y_1)l + (x_2+y_2)(l+1) = \sum_{p>l+1} p.x_p + \sum_{p<l} p.x_p$$
$$\overset{(d)}{\Leftrightarrow} \sum_{p>l+1} y_p(p-l) + \sum_{p>l+1} z_p(p-l-1) = \sum_{q<l} y_q(l-q) + \sum_{q<l} z_q(l+1-q)$$

Note (d) follows since $y_p + z_p = x_p \forall p$; $\sum_{p>l+1} y_p + \sum_{q<l} y_q = -(x_1+y_1)$ and $\sum_{p>l+1} z_p + \sum_{q<l} z_q = -(x_2+y_2)$

Now we can say from Claim 6 that
$\frac{f(i_p,j_p)-f(i_t,j_t)}{p-t} > \frac{f(i_u,j_u)-f(i_q,j_q)}{u-q}\ \forall p > l+1, q < l$ and $t, u \in \{l, l+1\}$.
Thus we can say the following:

$$\sum_{p>l+1} y_p(f(i_p,j_p)-f(i_l,j_l)) + \sum_{p>l+1} z_p(f(i_p,j_p)-f(i_{l+1},j_{l+1}))$$
$$> \sum_{q<l} y_q(f(i_l,j_l)-f(i_q,j_q)) + \sum_{q<l} z_q(f(i_{l+1},j_{l+1})-f(i_q,j_q))$$
$$\overset{(e)}{\Leftrightarrow} \sum_{p\neq l,l+1} (y_p+z_p)f(i_p,j_p) + (x_1+y_1)f(i_l,j_l) + (x_2+y_2)f(i_{l+1},j_{l+1}) > 0$$
$$\Leftrightarrow \sum_{p\neq l,l+1} x_p.f(i_p,j_p) + x_l.f(i_l,j_l) + x_{l+1}.f(i_{l+1},j_{l+1}) > 0$$
$$\Leftrightarrow \sum_{p} x_p.f(i_p,j_p) > 0$$

Note (e) follows since $\sum_{p>l+1} y_p + \sum_{q<l} y_q = -(x_1+y_1)$ and $\sum_{p>l+1} z_p + \sum_{q<l} z_q = -(x_2+y_2)$
Now let us consider the numerator of the first term in $\sigma_{\mathcal{D},x}(d)$ as in theorem 6 which can be written as $2.\sum_p n_{p,\mathcal{D}} f(i_p,j_p) + n.\Big(\sum_{t=1}^{r-1} t^2 \binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}\Big)$.
Thus the inequality proven in the previous result would imply that distribution D_1 has higher variance of number of distinct jobs returned than that of distribution D.

□

Claim 7. *Consider any distribution D satisfying the conditions in I. Consider any pair of jobs chosen uniformly at random and let this pair occur together in y servers. Then variance of y is linearly proportional to the variance of distinct jobs returned when any 2 servers return. We can also state it as follows.*

$$\sigma_{D,2}(d) = \frac{\binom{n}{2}.var(y) + \left(\frac{c\binom{k}{2}}{\binom{n}{2}}\right)^2 + n.\Big(\binom{r}{2}\Big) - c.\Big(\binom{k}{2}\Big)}{\binom{c}{2}} - \left(\frac{n\binom{r}{2}}{\binom{c}{2}}\right)^2 \tag{7}$$

Proof. Consider the numerator of first term in $\sigma_{D,x}(d)$ which had been shown to be $2.\sum_{p=0} n_{p,\mathcal{D}} f(i_p,j_p) + n.\Big(\sum_{t=1}^{r-1} t^2 \binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}\Big)$.

Note that here we consider $x=2$ as we consider the case when any 2 servers return.

Thus the first term in numerator of $\sigma_{D,2}(d)$ becomes $2.\sum_{p=0} n_{p,\mathcal{D}} f(i_p,j_p) + n.\Big(\sum_{t=1}^{r-1} t^2 \binom{r}{(t+1)}\binom{(c-r)}{(1-t)}\Big)$.

Now we can show that $f(i_0,j_0) = \sum_{i=2}^{k+1}\sum_{j=2}^{k+1}(i-1).(j-1).\binom{(n-2*k)}{(x-i-j)}.\binom{k}{i}.\binom{k}{j}$. For $x=2$, we can show that it goes to 0.

Now consider $f(i_p,j_p) - f(i_{p-1},j_{p-1}) = \Big[\binom{(c-2)}{(x-1)} - 2\binom{(c-r-1)}{(x-1)} + \binom{(c-2r+m-1)}{(x-1)}\Big]$.

However for $x=2$, $f(i_p,j_p) - f(i_{p-1},j_{p-1})$ reduces to $\Big[(n-2) - 2(n-k-1) + (n-2k+p-1)\Big] = p-1$. So we can say that $f(i_p,j_p) = \frac{p.(p-1)}{2}$.

Consider $n.\Big(\sum_{t=1}^{k-1} t^2 \binom{r}{(t+1)}\binom{(c-r)}{(1-t)}\Big)$ = $n.\Big(\binom{r}{2}\Big)$.

Thus for $x=2$,

$$
\begin{aligned}
&2.\sum_{p=0} n_{p,\mathcal{D}} f(i_p, j_p) + n.\Big(\sum_{t=1}^{k-1} t^2 \binom{k}{(t+1)} \binom{(n-k)}{(x-t-1)}\Big) \\
=&\sum_{p=0} n_{p,D} p(p-1) + n.\Big(\binom{r}{2}\Big) \\
\overset{(a)}{=}&\sum_{p=0} p^2 n_{p,D} + n.\Big(\binom{r}{2}\Big) - c.\Big(\binom{k}{2}\Big)
\end{aligned}
$$

(a) follows since $\sum_{p=0} p.n_{p,D} = n.\binom{k}{2}$.
Now consider

$$
\begin{aligned}
&\sigma_{D,2}(d) \\
=&\frac{\sum_{p=0} p^2.n_{p,D}}{\binom{c}{2}} - \left(\frac{n\sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(x-t-1)}}{\binom{c}{2}}\right)^2 \\
=&\frac{\sum_{p=0} p^2.n_{p,D} + n.\Big(\binom{r}{2}\Big) - c.\Big(\binom{k}{2}\Big)}{\binom{c}{2}} - \left(\frac{n\binom{r}{2}}{\binom{c}{2}}\right)^2
\end{aligned}
$$

Hence, proved.

Now we know that $\text{var}(y) = \frac{\sum_{p=0} p^2.n_{p,D}}{\binom{n}{2}} - \left(\frac{c\binom{k}{2}}{\binom{n}{2}}\right)^2$

We also know from theorem 3 that $\mathbb{E}_{D,2}[d] = 2.k - \frac{n.\sum_{t=1}^{r-1} t\binom{r}{(t+1)}\binom{(c-r)}{(2-t-1)}}{\binom{c}{2}} = 2k - \frac{n\binom{r}{2}}{\binom{c}{2}}$ whereas $\mathbb{E}[y] = \frac{\sum_{p=0} p.n_{p,D}}{\binom{n}{2}} = \frac{c.\binom{k}{2}}{\binom{n}{2}}$.

□

III. A SPECIAL CASE UNDER $n = c$

The theorem in 1 can be strengthened to Theorem 2 if the number of jobs equals the number of servers. It is restated and proven below.

Theorem. *Consider a distribution for a given n, k satisfying the conditions in I such that every pair of servers have l or $l+1$ common jobs for some positive integer l. Then the distribution has the least variance on the number of distinct jobs amongst all distributions satisfying property I.*

Proof. Under $n = c$, theorem 7 would imply that $\text{var}(y) = \sigma_{D,2}(d)$. Also we can show that $\mathbb{E}[y] = 2k - \mathbb{E}_{D,2}[d]$ from theorem 3.

Thus random variables $2k - d$ and y have the same mean and variance when $x = 2$. Recall that x denotes the number of servers we are able to communicate to.

Suppose every pair of servers of servers have l or $l+1$ jobs common. Under this criterion, d (under $x = 2$) would be able to take at most 2 consecutive values implying its variance is the least. Since, we know that $\text{var}(y) = \sigma_{D,2}(d)$ and $\mathbb{E}[y] = 2k - \mathbb{E}_{D,2}[d]$, we can say that y would have the least variance too. This would imply that y can take exactly at most 2 values, thus every pair of jobs occur together in l or $l+1$ servers.

Thus, using theorem 1, we can say that under this condition, the variance of the number of distinct jobs would be the least for any x.

□

IV. Constructions:

In general, these types of constructions can be done using balanced incomplete block designs as explained in [1]. In these type of construction, treatments are divided into blocks so that each block has the same number of treatment and each treatment occurs exactly the same number of times. However in balanced block constructions we also ensure that every pair of treatments occur together in the same number of blocks. Using theorem 1 we can say that it would have the least variance amongst distributions satisfying property I for a given n, k, r and c.

Various methodologies of balanced block constructions have been proposed in [1], [2]. The famous Bruck-Ryser-Chowla theorem gives some necessary condition on n, k, c, r so that it might be possible to have a balanced symmetric design.

We discuss one construction from [1] using vector spaces.

A. Construction using vector sub-spaces

Let us discuss such a construction. Choose a finite field vector space of dimension N, Let us denote each job as the subspaces of this vector space of dimensionality $K = 1$. Also we denote each server as a subspace of dimensionality C ($C > 1$). Now only those jobs are present in a server such that the vector-space corresponding to a job is the subspace of the vector-space denoting the corresponding server.

Note that the number of jobs and server is same since number of subspaces of dimensionality $N - K$ is same as the number of subspaces of dimensionality K. Also by geometry we can show that each sub-space of dimensionality K has a fixed number of sub-spaces of dimensionality $N - K$. Thus both the conditions in I are satisfied.

In this construction, every pair of jobs must occur together in exaclty the same number of servers. Thus, according to Claim 1, we can say that this construction has the least variance.

However such constructions are only possible from $n = \frac{q^a-1}{q-1}$, $k = \frac{q^b-1}{q-1}$, $c = \begin{Bmatrix} a \\ b \end{Bmatrix}$ where $\begin{Bmatrix} a \\ b \end{Bmatrix}$ denotes the number of b-dimensional subspace of a-dimensional space on a finite field of order q. This holds true for some positive integer a and q is a power of some prime number since cardinality of finite fields can also be a power of some prime number.

B. Construction using 2-D spread of points

Note this construction is somewhat similar to the planar construction you had described for $n = 9, k = 3$ case. We can generalise it for any $n = a.b$ (using a field for $\mathbb{F}_a$), $k = a$ and $a < b$ such that a is a power of some prime number say p and b is a multiple of $\frac{a}{p}$.

We could also do a similar construction for $n = a.b$, $k = b$, $a > b$ such that a is a power of some prime number say p and b is a multiple of $\frac{a}{p}$.

Recall that these constructions were done by treating the jobs as points and servers were treated as lines. Since these lines are constructed on a finite field, no two lines would have more than one job common.

Thus every pair of servers intersect in at-most one point, thus theorem 1 would imply that it has the least variance for every x.

References

[1] R. C. BOSE, "On the construction of balanced incomplete block designs," *Annals of Eugenics*, vol. 9, no. 4, pp. 353–399, 1939. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1939.tb02219.x

[2] D. A. Sprott, "Balanced incomplete block designs and tactical configurations," *Ann. Math. Statist.*, vol. 26, no. 4, pp. 752–758, 12 1955. [Online]. Available: https://doi.org/10.1214/aoms/1177728433