1. Introduction

The whole story has started with the Hilton. Rényi's car. It was a kind of a number of his family. That time, in the early nineteen, say five people had a trip in Hungary, the question was simple, there was no packing problem. Once he gave me a ride from the Institute to the car park, less than 100 metres. One day, however, the Hilton stopped in front of the service. Obviously, there was some electrical problem with it. The electrician, however, was unable to find the source of the trouble. Rényi tried to find it, himself. He has found it, and in the same time he has developed a general mathematical model for the situation.

The car can be considered as a finite set of its parts. The car does not work properly, when it has exactly one, or we suppose that all of its parts does not work properly. When trying to find it, some are performed. One time to check which is in defect. If it does not work then the defective part in the car, otherwise it is not noticed in it. After performing several such tests we have to determine the defective part.

Let us formulate it in some more mathematically. A finite set $X$ of objects is given. A distinguished element $x$ of $X$ is given that it is known by us. Furthermore, a family of subsets of $X$ is given. We can ask the questions $x \in A$ or $x \notin A$ for elements $A$ of $X$. We have to identify $x$ on the basis of the answers for the above questions. We call the members of $A$ positive sets. They are the positive questions.

A family $A_1, A_2, \ldots, A_n$ of actual questions (sufficiency of $A_i$) is fixed in advance then we say that this is a linear search. The obvious mathematical way to estimate the number of questions. On the other hand, the choice of the next question may depend on the answers to the previous questions. The first question set in

\[ \text{The work of the author is supported by the Hungarian National Foundation for Scientific Research Grant No. 3565.} \]
If the answer is \( \mathbf{A} \), then the second question set is \( \mathbf{A, A} \) while in the case of the answer \( \mathbf{C} \), the second question set is \( \mathbf{C, C} \). The third question set is \( \mathbf{A, A, C} \), \( \mathbf{A, C, A} \) and \( \mathbf{C, A, A} \). For \( \mathbf{A} \), the two previous answers are the same as the correct answer. In the case of the maximum number of questions (the length of the longest path of the tree), it is recommended not to count the number of answers. The tree search seems to be more practical, however it is important to be able to organize, so in the era of fast computers it might be equally important.

With the model based [1], instead of an area, the combinational search problems [2] and [3].

He and his collaborators have written many papers. However, his work was not the only source of these investigations. Few new ideas are often easier to accept.

Let \( \mathbf{A} \) be a set of soldiers in World War II. A sample of blood is drawn from every soldier. The case containing negligible variance should be found using the Waterspan test. It was an original idea that the blood sample could be passed and tested together. In this way it can be detected if a certain select of soldiers contain an infected person or not. This model is called the Waterspan test. The blood samples should be tested with the same number of elements to be identified in a known, in advance. See [4] and [5]. This kind of models are called group testing and continued to belong to Mathematical Statistics.

An even older question was posed by Steward [6]. A set of n table tennis players is given. Suppose that their abilities are constant. It can be described with a real number, then in the second round the better one always defeats the weaker one. The aim is to determine their total order by pairwise comparisons, that is, table tennis matches. Although it is not clear at the first sight, this problem is also covered by the above model. Let \( \mathbf{A} \) be the set of all permutations of the players. One of these permutations is the one that is the correct order. One unknown permutation belongs to the set of all permutations where player \( \mathbf{a} \) is better than player \( \mathbf{b} \).

So, one can say that the area has a winner if the correct permutation wins a survey paper [7] containing 86 references, the book by Alspach and Wagner [8] has 106 references and finally the most recent summarization of the area, the book of Algern [9] quotes 509 papers. These numbers show that the area because quite large, a small paper cannot cover it. Therefore the area of the present paper is to survey those papers which are written (mostly by Hungarian authors) under a direct or an indirect influence of Milg2.

[1] Milg2, 1944
[2] Steward, 1945
2. Qualitatively independent sets and partitions

Let $A, B, C$ be two question sets. Suppose that they are disjoint. First ask if the answer $a$ is in $A$ or not. If the answer is no, we have to ask $B$, as well. However, if the answer is yes then there is no need to ask $B$, we know that $a$ is not in $B$. If one of

$$A \cap B, A \cap C, A \cap \overline{B}, A \cap \overline{C},$$

is empty, the situation is similar, that is, one of the possible answers to the first question makes the second question superfluous. We say that $A$ and $B$ are qualitatively independent if none of the sets in (1) is empty.

Bécsi [30] raised the question what is the maximum of $|A|$ on an $n$-element set if any two different numbers of $A$ are qualitatively independent.
The answer is easy for even \( n \). It is easy to see that the set in (1) is an \( n \)-set containing all \( n \)-sets. A family \( F \) of sets is called a \( \Sigma \)-family if there is no inclusion in \( F \), that is, \( \mathcal{C} \subseteq \mathcal{D} \) holds for any two distinct members of \( F \). Using this notion, one can state that any two members of \( F = \{A_1, A_2, \ldots, A_n\} \) are qualitatively independent if and only if \( F' = \{A_1, A_2, \ldots, A_n, A_{n+1}, \ldots, A_{2n}\} \) is a \( \Sigma \)-family. However, the maximum size of a \( \Sigma \)-family is known:

\[
\binom{n}{\frac{n}{2}}
\]

Therefore, if \( F \) is a family of pairwise qualitatively independent sets then \( F' \) is a \( \Sigma \)-family and \( n \) is less than equal to the above binomial coefficient, so

\[
n \leq \frac{n}{2}
\]

This inequality is true for any \( n \) but it is also sharp for even \( n \), due to the following construction. Take all \( n/2 \)-element subsets containing a fixed element \( f \). The odd case is non-trivial, but easy. It was independently discovered by many authors [3, 6, 7, 10].

**Theorem 1.** The maximum size of a family of pairwise qualitatively independent sets on \( n \) elements is

\[
\binom{\frac{n}{2} - 1}{\frac{n}{2} - 1}
\]

The construction coincides with the even case

[2] also contains good estimates on a more general problem. We say that \( r \geq 2 \) sets are qualitatively independent, if they divide \( X \) into \( r \) non-empty sets. The maximum size of a family in which any \( r \) sets are qualitatively independent is estimated.

One may consider a more general condition. If all the sets in (1) are of size \( n \), then we say that \( A \) and \( B \) are qualitatively independent of depth \( d \).

**Problem 1.** What is the maximum size of a family of pairwise qualitatively independent sets of depth \( d \) on \( n \) elements?
It might be true that the above generalization of Theorem 1 holds for fixed $d$ and large $n$. The case when $d = n$ seems to be hard.

In what follows, we will consider another generalization. Before that a further motivation will be presented. It can be considered as the fourth score of the area of constructual search problems. Given a tree, one of them is constructable. It is known that the constructual tree is of higher than the good cases and it should be found by the minimum number of avoidings using an optimal path behavior. (No additional weights can be used.) The necessity of this result is that we should get an optimal number of paths through the constructed tree. In order to do this, we divide the tree into three parts: the set of nodes in the left ends, the set of nodes in the right ends, and the rest. In our earlier model the problem was divided into two main questions and its consequences. The example of the equal sets balance suggests to introduce the notion of the question partition $\mathcal{P} = (A_1, ..., A_k)$, where $A_1, ..., A_k$ is a partition of the grounded $X$. The answer to this question determines the unique $i$ satisfying $i \in A_k$. In this case a family $\mathcal{F}$ of partitions is given and the partition for a linear search or tree search we choose from the $\mathcal{F}$. If the number of parts in a partition does not exceed $n$ we call it an $n$-partition.

The notion of the constructual independence of partitions if they divide $X$ into $n$ nonempty subsets. Our result shows that the exact constructual number $\mathcal{N}(A, \mathcal{C})$ of the particular question can be determined, only its exponent. (Majzik and Tóth [36] proved that

$$\log_2 \mathcal{N}(A, \mathcal{C}) \leq \frac{n}{2}.$$  

A recent great achievement in

Theorem 2 ([130], [131], [132], [133]).

$$\log_2 \mathcal{N}(A, \mathcal{C}) = \frac{n}{2}.$$ 

One should mention the preliminary work of Rabin and Rabin [36].

The following result does not belong to this section, but it is a generating function used in the construction of certain points to mention it. Suppose that in $m$ cases are given, in $n$ of them are of weight $1$ (constructual tree) the rest of them are of weight $0$ (good tree). Find the shortest tree search determining all the constructual cases. As the number of possibilities is $2^n$ and one case has three different answers,
Observe that any question can be replaced by its complement. Thus, \( R \subseteq N_1 \) can be replaced. Furthermore, if after some question in a tree search it is known that the unknown is in the subset of size \( n \leq 20 \) then a can be found by (66). Further questions, using Theorem 3, then to get an optimal tree search. Note that in the form \( n = g + b \) where \( a < c < 24 \) and take a partition \( b_1, b_2, \ldots, b_k, b \) where \( b_i = n - b \). (66) \( \begin{align*}
\int_{0}^{1} \frac{dx}{\ln x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\frac{\ln x}{x} &= \int_{0}^{1} \frac{dx}{\ln x} \\
\end{align*}
\)

Theorem 5. Suppose that the question asks for the subset of size at most \( k \). Then the shortest true search needs
\[
\left\lceil \log_k (n-k+1) \right\rceil.
\]

When, in our previous, the reader asked what the minimum number of questions for a tree search is, many readers [B. Bollobas, J. Cokerhow, T. Hirst, and D. S.sad] read all the questions for the same answer for the case \( k \leq 5 \). This is the case for previous answers for the case \( k = 5 \). Then the present answer for the case \( k = 5 \) is as follows. (See also [20]). The binary search for \( k \), the number of elements that need to be searched, is, using questions 1 to \( k \), has a time complexity of \( \log_2 k \). Since the question asks for the subset of size at most \( k \), then the shortest true search needs
\[
\left\lceil \log_k (n-k+1) \right\rceil.
\]

Theorem 6. Denote by \( m(k) \) the minimum length of a linear search finding an unknown element in an \( n \)-element set using question sets of size at most \( k \).

\[
\begin{align*}
\log_2 m(k) &\leq \log_2 \frac{n}{k} \\
\log_2 m(k) &\leq \log_2 \frac{n}{k}
\end{align*}
\]

(See also [20] and [20]). As it is pointed out by B. Bollobas, the lower bound is asymptotically tight when \( k = n \).

Theorem 7. (B. Bollobas [20]) Suppose that \( k \) divides \( n \). Then the set of all \( k \)-element subsets of the \( n \)-element set \( X \) can be partitioned into such classes that each class forms a partition of \( X \).
Although this famous result does not belong to the Combinatorial Search Theory, it was created under an indirect influence of Bailey. It is obvious to ask what can be said if $k$ does not divide $n$. To formulate a conjecture concerning this general case, a new definition is needed. Suppose that the elements of the ground set $X$ are ordered: $X = (x_1, x_2, \ldots, x_n)$. Define $A = \{x_1, x_2, \ldots, x_m\}$, where the indices are considered modulo $n$. The family $A, A_2, \ldots, A_k$ where $\pi = \sigma(x, y, z)$ is called a search.

(Hoogland's theorem states that the number of such families is of the form $n/\pi$.)

**Theorem** (Raynaud and Sazyman). There are permutations of the ground set such that these permutations of the wreath give all possible states exactly once.

It seems to be hard to settle this conjecture. Sazyman's conjecture was further studied by other methods and an algorithmic way of solving Sazyman's brilliant idea was to use matrices and flows in networks. This conjecture, however, seems to be too difficult. Our paper does not expect to solve it without algorithms. ([Despite it is not true.)

Let us turn back to the search problem with restrictions on $k$. We will use the problem of Fredholm to obtain solutions. The problem actually becomes an important problem of combinatorial science (with numbers rather than tables being played). It is the simplest case of the so-called counting problems (see [3]). It is obvious by Thoma's rule that the search must be self-contained. The search is a process of finding a path from the initial state to the final state. The search can be regarded as a path in a graph. The search is determined, but not the second one. Let us see the problem of why the lower estimate $[\log k]$ cannot be reached by a new search. To check this, we have to look at each stage of the set of possible cases, therefore the question was divided the ground set of partitions into 4 equal parts. This is, however, not always possible. Consider the case $k = 1$ and $n = 2$. Denote by $A$ and $B$ the set of partitions giving positive answer to the first and the second question, respectively. Then two of the sets in [2] have size $2/3$ and the other two have size $1/6$. Thus, in the problem of the following investigations.

**Theorem 2** ([18]). The minimum length of a linear search using permutations satisfying...
Let us mention that this strange formula is equal to the smallest \( w \) such that

\[
\frac{Gm^2 - 1}{2}
\]

also gives the exact minimum up to an additional constant 1 for the case \( (n^a < 2) \). For \( n^a \) and \( m \) the same results are given for the case \( (n^a > 2) \) where \( k \leq n^a \).

For the case of two search let us start with an observation of Szemédi [25].

If the first question set is \( S \) and the answer is \( x \) and the second question set is \( T \), then it can be replaced by \( S \setminus T \). On the other hand, when \( x \) is a 0 then \( T \) can be replaced by \( S \). Continuing in this way, we obtain a modified tree search where the question sets on different branches of the tree are disjoint while the case along the same branch are in isolation. Of course the length of the branches are unchanged. Thus, when looking for the shortest tree search, this strong property may be impossible. E.g. \( x = 2 \) and \( x = 1 \) then the original condition becomes simply the condition that all question sets, with exception of the very first one, are of size 1.

Furthermore, the minimum length of a tree search under the condition \( (n^a > 2) \) is bounded, however, the formula is rather complicated and we give only the case \( x = 1 \) here.

**Theorem 8** (Szemédi [25]). The minimum length of a tree search using question sets satisfying

\[
(n^a > 2)
\]

is

\[
\frac{Gm^2 - 1}{2}
\]

Compare it with Theorem 9. The best linear search is not longer than the best tree search, in this case. For \( k = 1 \), however, the former case is about \( \sqrt{2}m \)-times larger than the latter case.

For another good estimate see for the case when the intersection of any set of question sets is bounded.

The combination of the previous two types of constraints has not been studied yet.
PROBLEM 2. Determine the length of the shortest linear and tree search, resp., under the conditions

$$|A| \leq k, |A| \leq L$$

for all $A, B \in D$.

Let us see the situation at the search of a permutation by binary comparison. We observe that the comparison $x < y$ and $y < x$ sharply the center of two questions are dividing the set of permutations into four parts containing one third, one third, one sixth, one sixth of the whole set, instead of the "proportional" one fourth, one fourth, one fourth, one fourth. However, this is not true for all parts of questions! What we can see is that among any $\frac{3}{4}$ questions there is one each pair. One way of expressing the fact that two questions are not intersecting each other is a "proportional" one of the entropy function of the Information Theory. The entropy of the partition $D : D = \frac{A}{\bar{A}}$ of $D$ is

$$H = \sum_{[X]} \log \frac{|A|}{|\bar{A}|}$$

This expression is 2 for the case when $A = \frac{1}{2} D$. It is known from Information Theory that it is smaller in all other cases. This suggests the following problem.

PROBLEM 3. Determine the length [D, D] of the shortest linear search under the condition that among any $r$ questions with them is a pair such that the $r$-th question implied by them has an entropy at most $E$.

We have only estimates, to formulate them some more definitions are needed. Put $H(R) = -\sum_{\frac{1}{2} \log_2 |A| - 1 - \log_2 |\bar{A}| - 1}$. The increase of $A$ is defined using the interval $[D, D] \leq J_0$ where $J_0$ is measured.

**Theorem 10.**

$$H(\frac{1}{2} \log_2 \frac{|A|}{|\bar{A}|}) - \frac{1}{2} \log_2 \frac{|A|}{|\bar{A}|} \leq [D, D] \geq \frac{1}{2} \log_2 \frac{|A|}{|\bar{A}|} + O(\log_2 n).$$

**Proof.** Start with the base estimate. Let $g$ be a random variable taking on values from $D$. Define the probabilities to be equal: $P(R) = \frac{1}{n}$. Denote the question sets by $A_1, \ldots, A_n$. They define some further random variable

$$D = \bigcap_{i} \bigcap_{j} A_i.$$
Now define the entropy of a random variable $y$ taking on its different values with probabilities $p_i$ as:

$$H(y) = -\sum_{i=1}^{m} p_i \log p_i.$$ 

Observe that $I$ determines the random vector $(x_1, \ldots, x_m)$. On the other hand, so do the answers to the question "is $u \in A^j$" determine $y$, therefore the covariance is also true, $(x_1, \ldots, x_m)$ determine $I$. Consequently, the distribution of the two random variables are identical and

$$I(y) = H(x_1, \ldots, x_m).$$

Now we use an elementary lemma from Information Theory (see any textbook on Information Theory, e.g. [14]):

(1) $H(y) = H(x_1, \ldots, x_m) = \log n$.

(2) and (3) imply

(4) $\log n \geq H(x_1, x_2, \ldots, x_m) + H(x_1) + H(x_2) + \ldots$

for any partitioning of the set $(x_1, \ldots, x_m)$ into two and more-element subsets. If they are all constant sets then (4) leads to $\log n \geq m$, hence the entropy of one $x_i$ is bounded by one. However, if we find $m$ such disjoint parts that $H(x_1, x_2, \ldots, x_m) \geq H(x_1)$ then (4) needs $m \leq n$.

To find the best $m$, the problem will be reformulated for graphs in an obvious way. Define a graph whose vertices are $G$ and two vertices are connected if $H(x_i, x_j) \geq H(x_i)$. The following graph-theoretic lemma is needed:

**Lemma:** Given a simple graph $G$ on $m$ vertices, there is at least one edge among any $t$ vertices. Then it contains at least

$$\frac{m(m-1)}{2}$$

edges, if each edge $G$. This result is sharp.
Prior. Take the largest set $S$ of vertex-disjoint edges. Let $|S| = k$. If $m - 2k > 0$ then there are $m$ vertices not contained in any member of $S$. By the conditions of the lemma there is an edge among these $m$ edges which is vertex-disjoint to the members of $S$. This contradicts the maximality of $S$.

The contradiction proves $m - 2k = 0$ and the first part of the lemma.

Now consider the graph $G$ of $m$ vertices consisting of a complete graph on $m - 1$ vertices and isolated vertices. This graph obviously satisfies the conditions of the lemma and cannot contain more vertex-disjoint edges than $m - 1$.

The lemma and (1) imply

$$\log_2 n = \frac{m - 2k - 1}{2} \leq m$$

and the lower estimate in Theorem 10.

The upper estimate will be proved by a simple construction. Define $A = (m-1)(E_2)$. Question some set of size $k$ will be used. Then, by the maximality of $S$, we have $|S| \leq m - 1$. Thus, (2) implies $\log_2 n \leq m$, as needed.

Use the construction mentioned after Theorem 6 or $k/n$ is a constant.

Then the lower estimate is sharp in Theorem 10. It gives the upper estimate

$$\log_2 n = O(\log m)$$

which coincides with the one given in the Theorem.

Let us note some remarks concerning the Theorem.

1. It gives an approximate solution to the problem of Theorem 6 in a new way.

2. Problems 9 and Theorem 10 are not intended to help finding the shortest lower estimate for a permutation by pairwise comparison. It is a trivial problem, one has to compare all $\binom{n}{2}$ pairs. However the solution of the edge-graph problem for the tree search might give a better lower estimate on the permutation problem.

4. Miscellaneous

As it was mentioned earlier, a tree search needs $\log_2 n = O(n)$ steps to find the proper permutation of $n$ objects (already) by pairwise comparison. Modern computers have special hardware able to execute many operations simultaneously. This is called parallel computation. As one object can be
used only in one comparison at each moment, not more than \( w/2 \) parallel operations are possible. Therefore at least \( \Omega(n^x) \) steps are needed even if parallel steps are allowed. Ajtai, Komlós and Szemerédi [2] proved that this can really be done in this many steps.

A variant has been proved: an interesting result of sorting type. Let us denote by \( a_1, a_2, \ldots, a_n \) the distinct integers in \( X \). Let \( b_1, b_2, \ldots, b_n \) be their neighbors in the natural order of all three numbers. It is easy to see that \( X - \{ b_1, b_2, \ldots, b_n \} \) is enough. However, in one case it could be proved that there is no shorter than this algorithm. It is proved in [6].

If there are more unknown elements in \( X \), then one question will give different answers. One of the normal models is that there are two possible answers, either \( a_1 \) or \( a_n \). If there are more unknown elements, then in some cases, Bangu and Vesic [11] gave good estimates for the minimum length of a linear search when the number of unknown elements is known in advance.

There are two connected search theories in the theory of integer lists and finite automata. The usual idea of linear search (in a fixed order) to determine the weights of different order factors for the connected quantity. Sometimes it is safe to assume that there are many possible factors, very few of them have a real influence on the prime one has no influence (or is negligible). However, it is not known which ones are the non-negligible factors. A weaker version of this is that the number of non-negligible factors is finite, and the number of non-negligible factors is infinite. This is called the connected search theory. Two investigations led to the problem of finding the largest factor of the operation minimizing the difference and such that \( f(n) \) is a finite factor. Generalization can be found in [4].

REFERENCES


[34] Lebesgue, H., Sur la mesure des ensembles de points d'accumulation, Math. Ann. 68 (1915), 615–618.


[37] WEGENER, I., On separating systems whose elements are sets of at most k elements, Discrete Math. 28 (1979), 219–222. MR 84c:05019

(Received August 8, 1991)