Graph theoretical tools for the study of the cumulative build-up of empirical sociology

Dénes, Tamás mathematician

email: tdenest@freemail.hu

Introduction

The empirical apprehension of sociological phenomenons /sociological investigation/ takes place in four main steps:

I. Assessment of variables describing the phenomenon /hypothetical model/.

II. Selecting the social subsystem that is medium of the phenomenon /sampling/.

III. Measurement of the variables assumed on the sample.

IV. Evaluation of measuring data /empirical model/.

Considering sociological investigations as a phase each of the process of cognition serving for describing the given phenomenon, rather than as independent units, so the following problems arise:

a. The set of variables adopted does not coincide in general in two different investigations.

b. Representativity of the selected samples can not be secured to an identical degree in the different investigations, all the less as for the reason of the problem mentioned under point a.

c. There often appear deviations in the measurement even of identical variables in different investigations.

d. Most varied methods are adopted for the evaluation of measuring data.

Consequently there presents itself as a fundamental question, what may warrant the compare of different investigations concerning a given phenomenon, or how can their superposition /cumulativity/ be investigated ?

We shall desist here from a detailed tracing of the contents of the problems mentioned above and rather concentrate on the description of a model which may prove to be an efficient means for their solution.

1. Mathematical model

The basic principle of the model as follows:

a. To the given sociological phenomenon /sign: J/ /like to a system/ can be corresponded the structure of adopted variables to be determined from measured data, which structure can be described upon the media as an uniform reference system. /See: [2] , [3] /

b. The function of the individual variables within the given phenomenon is exactly determined by the structure-environment.

Mathematical description of the model:

Let $T_1, T_2, \ldots, T_r$ denote the investigations performed with a view to describing the phenomenon J, where the indexes can be denoted a time sequence.
Taking into consideration the steps I., II., III. as described in the introduction, planning each T_i /i=1,2,...,r/ investigation can be characterized by a set-triad (V_i, H_i, M_i) , where

V_i : the set of variables adopted in the investigation T_i

H_i : The set of sample elements adopted in the investigation T_i

M_i : The set of measuring procedures belonging to the variables in the set V_i.

The sets V_i and M_i the following relation realizes:

$$(1.1) \qquad v_{ij} \in V_i \Longrightarrow \exists\, m_{ij} \in M_i$$

where m_{ij} is the measuring procedure belonging to the variable v_{ij}, namely a measuring prodedure within M_i belongs to any variable within V_i, which means that

$$(1.2) \qquad |V_i| \leq |M_i|$$

On the <u>measurement of any variables</u> in V_i $/\forall m_{ij} \in M_i/$ we understand as follows :

$$(1.3) \qquad m_{ij} : H_i \longrightarrow K_{ij},$$
$$h \in H_i \Rightarrow m_{ij}(h) = k \in K_{ij}$$

where K_{ij} is the set of codes of the variable $v_{ij} \in V_i$.

The set of codes terminology is adopted for the sake of the possibility of a uniform treatment, as in this way it is not necessary to separate the socalled quantitative and qualitative variables. This does not cause any /theoretical/ difficulties, since it is easy to see that it is always possible to give a set of codes consisting of discrete codevalue for the range of interpretation of any quantitative variable /forming adequate classes/.
Thus, according to formula (1.3) the elements of the set M_i are such transformations, as each element of H_i correspondent to exactly one code value /belonging to a respective variable/.
Then introducing the following notations

$$D_{ij} = \bigcup_{h \in H_i} \{m_{ij}(h)\} \tag{1.4}$$

$$D_i = \bigcup_{m_{ij} \in M_i} D_{ij} \tag{1.5}$$

or, D_i is the set of data arising from the investigation T_i, so the performance of the investigation T_i can be described by the transformation :

$$V_i \times H_i \longrightarrow D_i \tag{1.6}$$

where "x" indicates the Descartes multiplication.

1.1. DEFINITION

On <u>structure of variables</u> in the set V_i we understand the structure of the relation R interpreted on the set V_i, namely

$$(1.7) \qquad R \subseteq V_i \times V_i$$

Based on the definition 1.1. and the known connection among relations and graphs any (V_i, R) system of variables can be assigned to a $\Gamma_i = (P_i, E_i)$ graph. Each p_{ij} vertices in P_i exactly represents the respective v_{ij} variable in V_i, and on the edges in E_i it realizes that

$$(1.8) \qquad (p_{il}p_{iz}) \in E_i \Leftrightarrow v_{il}Rv_{iz} \quad /p_{il}, p_{iz} \in P_i/$$

Thus, according to the above, there exists a one-one correspondence between the $T_1, T_2, \ldots, T_r$ investigations and $\Gamma_1, \Gamma_2, \ldots\ldots, \Gamma_r$ graphs.

1.2. DEFINITION

Let $\Gamma_i = (P_i, E_i)$ and $\Gamma_j = (P_j, E_j)$ denote the graphs representing the structures of variables belonging to

investigations T_i and T_j. Investigation T_j is called <u>cumulative</u> reference to investigation T_i, if

(1.9) $V_i \subset V_j$ and $\Gamma_i \subset \Gamma_j$

(1.10) Γ_j is connected

Thus, cumulativity is the relation interpreted on the set of investigations. Its symbol is R_c.

The formalized writeing of relation "T_j cumulative reference to T_i" is: $T_j R_c T_i$.

Hereinafter we shall deal with the description of the structure of relation R_c, which permits analysing of the empirical cognitive process.

2. <u>Analysis of the structure of empirical cognitive process</u>

Cumulativity /relation R_c/ permits a generalization of the comparative problematicalties of two arbitrary investigations, namely the analysis of the complete empirical cognitive process of a phenomenon J performed up to a given point of time.
Let see the graph of relation R_c, i.e. the graph $\Gamma_{R_c} = (P_{R_c}, E_{R_c})$ the vertices of which represent

$T_1, T_2, \ldots, T_r$ investigations in a way that P_i vertex is represents investigation T_i, while the edges of the graph are defined as follows:

$$(2.1) \qquad (P_i P_j) \in E_{R_c} \Longleftrightarrow T_i R_c T_j$$

Based on the known properties of relation $\subset$/irreflexive, antisymmetrical, transitive/ it follows from definition 1.2. that relation R_c possesses the irreflexive, antisymmetrical, trancitive properties. This implies that <u>cumulativity is an ordering relation</u> interpreted on the set of investigations. Consequently, the Γ_{R_c} graph corresponding to the relation R_c is a so-called transitive graph, i.e. directed graph with no loop and cycle. An essential property of transitive graphs is that any partial graphs of their also possesses these properties, namely any partial graphs of their is a transitive graph.

The four figures below we show those elementary types of graphs, which lend themselves for building up any Γ_{R_c} graph of the property mentioned above.

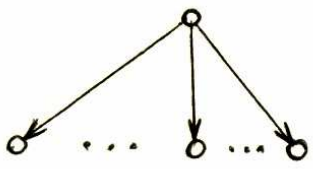

Figure 1.

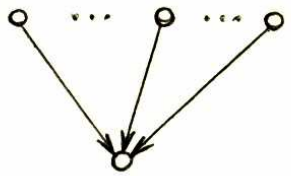

Figure 2.

Figure 3.

Figure 4.

Figure 3. actually demonstrates that extreme case, which is the limitation case of the types represented in Figures 1., 2., 3.
The structure of cognition belonging to the elementary graphs as in Figure 1. may be mentioned as a summing up type, that belonging to the elementary graph as in Figure 2. as a searching type.
The structure of cognition as in Figure 3. is designated as an elementary ideal structure.
The cognition structure which is a sequence of the elementary ideal structure, we are called ideal.
Graph of an ideal structure of cognition is a unique directed path /see Figure 5./.
The term directed path means such a directed series of edges $(u_1, u_2, \ldots, u_i, \ldots, u_q)$ where the terminal endpoint of edge u_i is the initial andpoint of the edge u_{i+1} for all $i < q$.

Figure 5.

It is known that any transitive graph /thus Γ_R too/ can be decomposed into directed paths. Consequently any structure of cognition can be produced as union of ideal structures of cognition.

Therefore it is just to denote as ideal such structure of cognition as consists of exactly one such /ideal/ member, since this structure represents in fact the cumulative cognition without roundabouts /ramifications/.

Let $\Gamma_{R_c}/t_i/$ denote the graph of R_c relation on the set $T_1, T_2, \ldots, T_r$ investigation in point of time t_i /$i = 0,1,2,\ldots$/, where

$$\Gamma_{R_c}(t_i) = \left(P(t_i),\ E(t_i)\right) \tag{2.2}$$

Define a $k/t_i, t_j/$ integer function as follows /where t_i, t_j two arbitrary moment, and $j > i \geqslant o$/:

$$k(t_i, t_j) \geqslant \begin{cases} 2 & \text{if } i = 0 \\ 1 & \text{if } i \geqslant 1 \end{cases} \tag{2.3}$$

On the basis of /2.3/ we can give the number of investigation in any t_i moment by the following recursive mode:

$$\left|P(t_o)\right| = 0 \tag{2.4}$$

$$\forall j > i \geqslant 1 \Longrightarrow \left|P(t_j)\right| = \left|P(t_i)\right| + k(t_i, t_j) \tag{2.5}$$

On the basis of /2.5/ we give the maximal number of edges in $\Gamma_{R_c}(t_j)$ graph at the moment $t_j > t_i$:

$$\frac{n_j(n_j-1)}{2} = \frac{\left(n_i+k(t_i,t_j)\right)\left(n_i+k(t_i,t_j)-1\right)}{2} \tag{2.6}$$

(2.6) is the edge number of a complete graph with n_j vertex.

Let denote $|P(t_i)| = n_i$ /i = 0,1,2,...../, then we define the efficiency index of the cognitive process of phenomenon J between t_i, t_j moments by the following way:

$$(2.7) \quad H(t_i,t_j) = \frac{|E(t_j)| - |E(t_i)|}{\frac{(n_i+k(t_i,t_j))(n_i+k(t_i,t_j)-1)}{2} - \frac{n_i(n_i-1)}{2}}$$

What is true on the lower and upper bound of $H(t_i,t_j)$?

On the basis of /2.6./ we give:

$$(2.8) \quad 0 \leq |E(t_i)| \leq \frac{n_i(n_i-1)}{2}$$

$$(2.9) \quad 0 \leq |E(t_j)| \leq \frac{n_j(n_j-1)}{2}$$

$$(2.10) \quad 0 \leq |E(t_j)| - |E(t_i)| \leq \frac{n_j(n_j-1)}{2} - \frac{n_i(n_i-1)}{2}$$

From the /2.3/, /2.4/, /2.5/ connections it is clear, that

$$(2.11) \quad \forall i \geq 1 \Longrightarrow k(t_i,t_j) \geq 1 \quad \text{and} \quad n_i \geq 1$$

Thus from the above and /2.6/ connections, we have that

$$(2.12) \qquad \frac{n_j(n_j-1)}{2} - \frac{n_i(n_i-1)}{2} \geqslant 1$$

Furthermore

$$(2.13) \qquad i=0 \Longrightarrow k(t_i,t_j) \geqslant 2 \quad \text{and} \quad n_i = 0$$

in this case /2.12/ is realized too, and we find the next values of the bounds of $H(t_i,t_j)$:

$$(2.14) \qquad 0 \leqslant H(t_i,t_j) \leqslant 1$$

On the $H(t_o,t_j)$ index we find a very simple connection:

$$(2.15) \qquad H(t_o,t_j) = \frac{2\cdot|E(t_j)|}{(k(t_o,t_j)-1)\cdot k(t_o,t_j)} = \frac{2\cdot|E(t_j)|}{(n_j-1)n_j}$$

Let's have to demonstrate the connection between the structure of cumulativity and efficiency of the cognitive process. Therefore we show the efficiency index of the elementary types of graphs /see figure 1-4./.

Let see for the first the summing up type between the t_i, t_j moments /see figure 6./

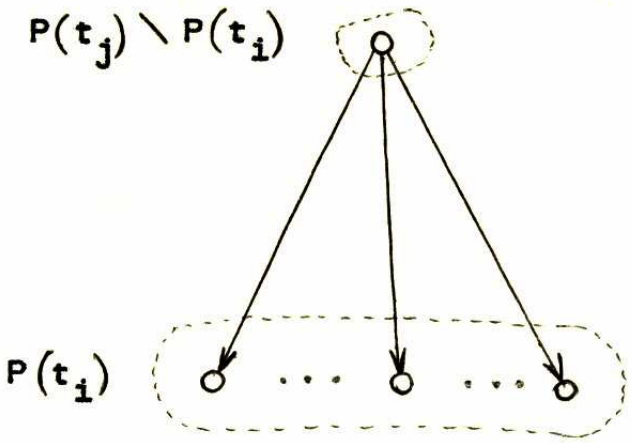

Figure 6.

In this case the $k(t_i,t_j) = 1$ and $|E(t_j)| = |E(t_i)| + n_i$ connections are realize, from which

$$(2.16) \quad H(t_i,t_j) = \frac{2\left(|E(t_i)| + n_i - |E(t_i)|\right)}{(2\cdot n_i + 1 - 1)\cdot 1} = 1$$

Let see for the second the searching type /see figure 7./.

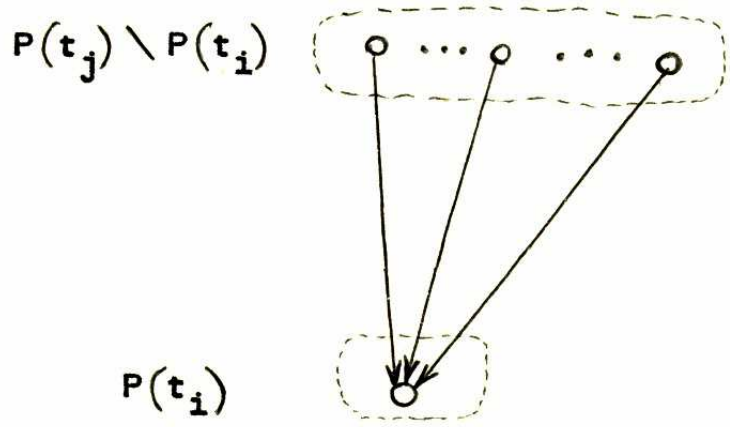

Figure 7.

In this case from the $n_i = 1$ and $|E(t_j)| = |E(t_i)| + k(t_i,t_j)$ connections we have

$$(2.17) \qquad H(t_i,t_j) = \frac{2\left(|E(t_i)| + k(t_i,t_j) - |E(t_i)|\right)}{\left(2n_i + k(t_i,t_j) - 1\right)\cdot k(t_i,t_j)} =$$

$$= \frac{2}{2n_i + k(t_i,t_j) - 1} = \frac{2}{k(t_i,t_j) + 1}$$

i.e., if we take the opportunity of abstraction of infinite number investigations , then

$$(2.18) \qquad \lim_{k(t_i,t_j)\to\infty} H(t_i,t_j) = 0$$

From the /2.16/ and /2.18/ connections we can see, that the efficiency of cognitive process determined by proportion of the searching and summing type subprocess.

Now, to demonstrate it, we show a $\Gamma_{R_c}(t_j)$ graph of a fictive cognitive process /j = 5, without the transitive edges, see figure 8./ and its $H(t_o, t_j)$ index.

The adjacency matrix of the original $\Gamma_{R_c}(t_j)$ graph /with the traneitive edges/ is on figure 9.

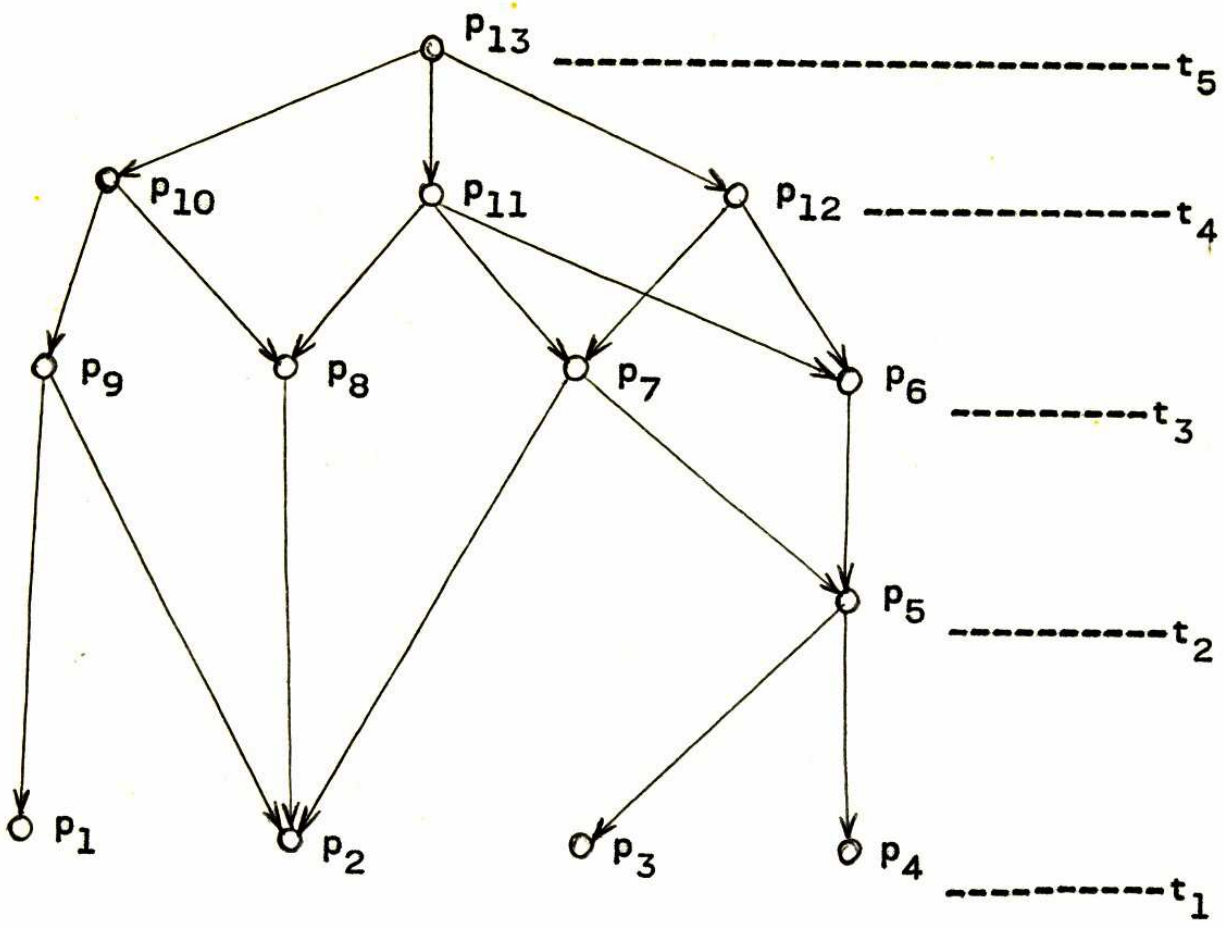

Figure 8.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1													
2													
3													
4													
5			1	1									
6			1	1	1								
7		1	1	1	1								
8		1											
9	1	1											
10	1	1						1	1				
11		1	1	1	1	1	1	1					
12		1	1	1	1	1	1						
13	1	1	1	1	1	1	1	1	1	1	1	1	

Figure 9.

From the adjacency matrix it is clear that

$$(2.19) \qquad \left|P(t_j)\right| = n_j = 13 \quad \text{and} \quad \left|E(t_j)\right| = 41$$

thus in this example:

$$(2.20) \qquad H(t_o,t_5) = \frac{2\left|E(t_j)\right|}{n_j(n_j-1)} = \frac{82}{13 \cdot 12} \approx 0{,}52$$

Finaly we show the efficiency indexes of Γ_{R_c} at disjunct time-intervals /see figure 8./.

$$\left.\begin{array}{ll} n_o & = 0 \\ k(t_o,t_1) & = 4 \\ |E(t_o)| & = 0 \\ |E(t_1)| & = 0 \end{array}\right\} \Longrightarrow H(t_o,t_1) = \frac{2(0-0)}{(2\cdot 0+4-1)\cdot 4} = 0$$

$$\left.\begin{array}{ll} n_1 & = 4 \\ k(t_1,t_2) & = 1 \\ |E(t_1)| & = 0 \\ |E(t_2)| & = 2 \end{array}\right\} \Longrightarrow H(t_1,t_2) = \frac{2(2-0)}{(2\cdot 4+1-1)\cdot 1} = 0{,}5$$

$$\left.\begin{array}{ll} n_2 & = 5 \\ k(t_2,t_3) & = 4 \\ |E(t_2)| & = 2 \\ |E(t_3)| & = 12 \end{array}\right\} \Longrightarrow H(t_2,t_3) = \frac{2(12-2)}{(2\cdot 5+4-1)\cdot 4} = 0{,}38$$

$$\left.\begin{array}{ll} n_3 & = 9 \\ k(t_3,t_4) & = 3 \\ |E(t_3)| & = 12 \\ |E(t_4)| & = 29 \end{array}\right\} \Longrightarrow H(t_3,t_4) = \frac{2(29-12)}{(2\cdot 9+3-1)\cdot 3} = 0{,}56$$

$$\left.\begin{aligned} n_4 &= 12 \\ k(t_4,t_5) &= 1 \\ |E(t_4)| &= 29 \\ |E(t_5)| &= 41 \end{aligned}\right\} \Longrightarrow H(t_4,t_5) = \frac{2(41-29)}{(2\cdot 12+1-1)\cdot 1} = 1$$

The graph Γ_{R_c} defined in the foregoing /being a transitive graph/ possesses the property as well that its vertices can be ordered into levels, that means that from any vertex of a given level just can be connected only vertices of lower level. /see [1]/ Consequently we obtain another characteristic feature, which can be expressed numerically, of the structure of recognition, when investigating the homology of the ordering in time of the investigations adequate to the vertices of the graph Γ_{R_c} correspond with the ordering created in the set of investigations by the relation /R_c/ of cumulativity. Let us assume for the investigation of this problem that the indexes of investigations $T_1, T_2, \ldots\ldots, T_r$ indicate the sequence in time of the investigations.

2.1. DEFINITION

The relation R_I interpreted on the set of $T_1,\ldots,T_r$ investigations we are called <u>inversion relation</u>, if

$$(2.21) \qquad T_i R_I T_j \iff T_i R_c T_j \qquad i < j$$
$$/1 \leq i \leq r,\ 1 \leq j \leq r/$$

Similarly to the relation R_c, an $\Gamma_{R_I} = (P_{R_I}, E_{R_I})$ directed graph belongs to the relation R_I as well. The set of edges of this latter graph contains exactly the edges connecting vertices of those investigations in the case of which investigation-pairs the sequence in time does not coincide with the cumulativity sequence.
Let us then form the index δ_I as follows:

$$(2.22) \qquad \delta_I = \frac{|E_{R_I}|}{|E_{R_c}|}$$

It is obvious that if the two kinds of ordering are isomorphic then the number of edges of Γ_{R_I} is minimal, i.e. $|E_{R_I}| = 0$ and exactly then the value of δ_I is minimal, so $\min \delta_I = 0$.

At the same time if the two kinds of ordering are "very" dissimilar, namely $|E_{R_I}| = |E_{R_c}| \Rightarrow \max \delta_I = 1$ thus

$$(2.23) \qquad 0 \leq \delta_I \leq 1$$

3. Analysis of non-cumulative investigations

In the following we are going to examine what the situation is in those cases where the cumulativity relation /R_c/ does not exist between the test-pairs. Markings introduced in the previous points will be adopted continually, and let G_i, G_j denoted the spanning partial graph in Γ_i and Γ_j respectively, by the variables equally adopted in the T_i and T_j investigations.

Now the table here below contains all possible relations of the structure of the T_i, T_j investigations. /Under the structure of an investigation we understand the structure of variables adopted in the investigation, according to definition 1.1./ Numbers indicated in the table serve only for subsequent reference.

	$V_i = V_j$	$V_i \subset V_j$	$V_i \cap V_j \neq \emptyset$	$V_i \cap V_j = \emptyset$
$G_i = G_j$	1 $T_j \bar{R}_c T_i$	2 $T_j R_c T_i$	3 $T_j \bar{R}_c T_i$	4 ------
$G_i \cap G_j \neq \emptyset$	5 $T_j \bar{R}_c T_i$	6 $T_j \bar{R}_c T_i$	7 $T_j \bar{R}_c T_i$	8 ------
$G_i \cap G_j = \emptyset$	9 $T_j \bar{R}_c T_i$	10 $T_j \bar{R}_c T_i$	11 $T_j \bar{R}_c T_i$	12 $T_j \bar{R}_c T_i$

$\bar{R}_c$ denotes the opposite of relation R_c.

Development of the cases indicated in the table, in the sequence of reference numbers:

1. In this case $G_i=G_j \Rightarrow \Gamma_i = \Gamma_j$ /since $V_i=V_j$/. Thus, the structure of the two investigations is isomorphyc. Then there is no cumulativity to be mentioned, but in spite of this such investigations may have a very important function, since this case is nothing else than a confirmation of the T_i investigation through T_j. May we point out however, that this case is comprised in case No.2., i.e. that of cumulativity, consequently this latter is of much higher efficiency, since confirmation and building up /continued recognition/ takes place within one stage.

2. It is the case of cumulativity, which has been dealt with in detail in chapter 2.

3. In this case the conditions of cumulativity are not fulfilled however, the two investigations contain a common partial investigation, in respect of which both are cumulative. Let T_{ij} denoted this common partial investigation. Assuming the set of variables of the partial investigation T_{ij} as $V_{ij} = V_i \cap V_j$ then:

$$\left.\begin{array}{lll} V_{ij} = V_i \cap V_j & \Rightarrow & V_{ij} \subset V_i \\ G_i = G_j & \Rightarrow & \Gamma_{ij} = G_i \subset \Gamma_i \end{array}\right\} \Rightarrow T_i R_c T_{ij}$$

$$\left.\begin{array}{l} V_{ij} = V_i \cap V_j \Longrightarrow V_{ij} \subset V_j \\ G_i = G_j \Longrightarrow \Gamma_{ij} = G_j \subset \Gamma_j \end{array}\right\} \Longrightarrow T_j R_c T_{ij}$$

Figure 10. shows the recognition graph of this case.

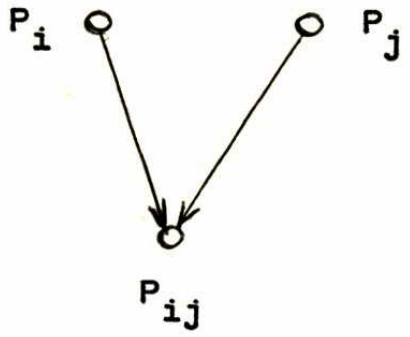

Figure 10.

4. This case cannot take place, since G_i and G_j partial graphs do not exist due to $V_i \cap V_j = \emptyset$.

5. In this case the situation is similar to that of case 3., but the determination of the partial investigation is more difficult. Now we give the process for tracing the V_{ij} set of variables of the common partial investigation T_{ij} and the conditions required for its existence. As will be seen in the following, here T_{ij} fulfilling the conditions does not necessarily exist, in contradistinction to the case 3.

3.1. DEFINITION

Let $\Gamma=(P,E)$ an arbitrary directed graph and $p_i \in P$. The graph $\Gamma'=(P', E')$ called the <u>partial graph within Γ generated by the vertex</u> p_i, if

$$(3.1) \qquad p_j \in P' \iff (p_i p_j) \in E$$

$$(3.2) \qquad (p_j p_k) \in E' \iff p_i, p_k \in P', \quad (p_j p_k) \in E$$

Thus, assume in this case $G_i \cap G_j = G$ and the set of variables V adequate to the vertex spanning of G. The elements of V only those variables, where the following condition is fulfilled in respect of the vertices corresponded to them within G:

$$(3.3) \qquad v \in V \iff k(p) + b(p) \neq 0$$

/p is the vertex in G corresponded to the variable v/

Then the variables in V_{ij} will exactly be those variables in V, for which the partial graphs in Γ_i and Γ_j respectively, generated by the vertices representing these variables are parts of G. In a formalized description this means that, if $p \in G$ and Γ_i^p, Γ_j^p in sequence are the partial graphs in Γ_i and Γ_j respectively, generated by the vertex p, then the following condition must be fulfilled in respect of variable v represented by vertex p:

$$(3.4) \qquad v \in V_{ij} \iff \Gamma_i^p = \Gamma_j^p \subseteq G$$

It follows from this condition that, although in case 5. the condition $V \neq \emptyset$ is secured, there still easily may occur $V_{ij} = \emptyset$. To illustrate this, Figures 11./a-k. show such a case to be readily controlled.
There are to be seen on the figures the same symbols as are adopted in the text.

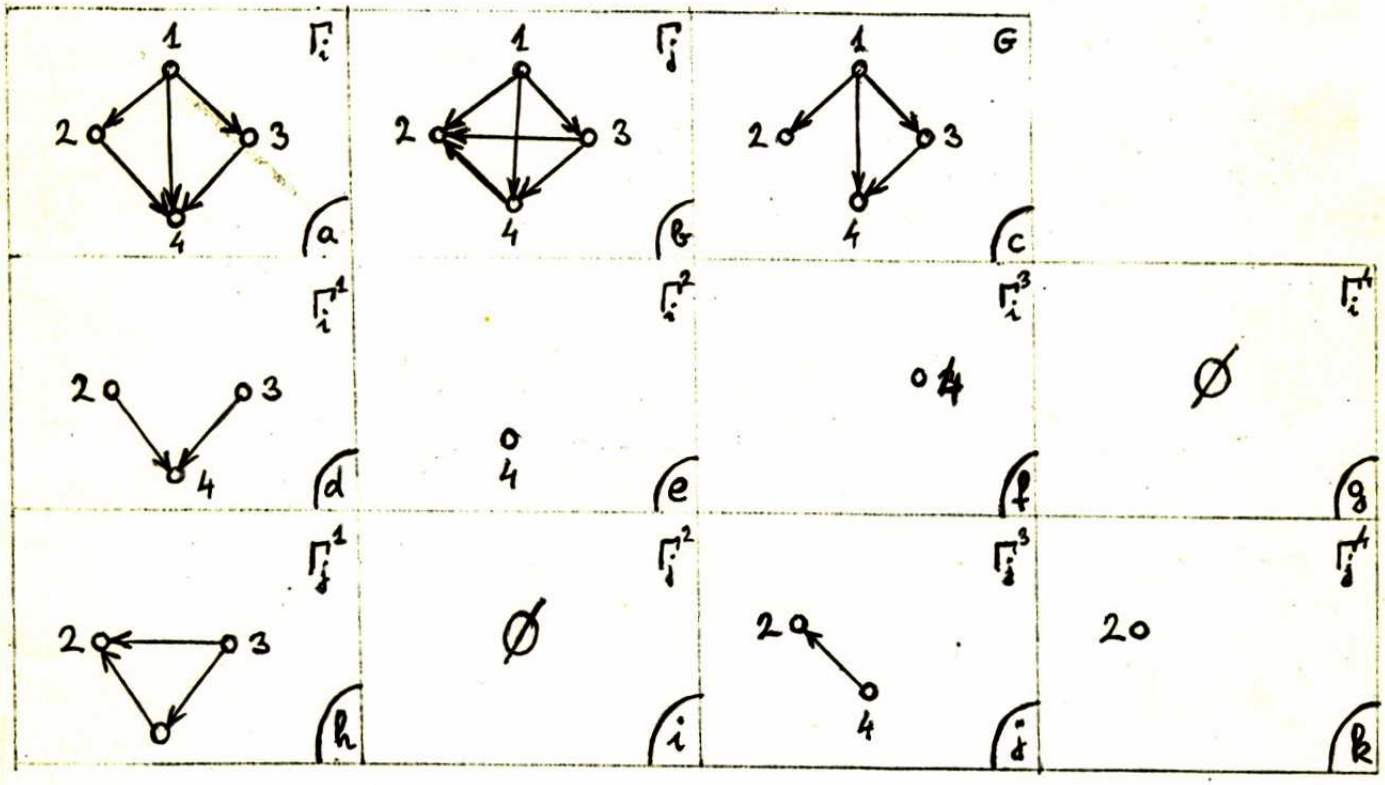

Figure 11.

6., 7. In these two cases the proceedings are identical to those performed in case 5.

8. What has been said for the case 4. are equally valid here too.

9., 10.

11., 12. In these cases one cannot speak of cumulativity, not even up to the level of partial investigations, due to the condition $G_i \cap G_j = \emptyset$. Consequently in these cases the investigations may be considered as completely isolated ones.

REFERENCES

[1] C. Berge: Graphs and Hypergraphs
North-Holland, 1973.

[2] T.Dénes : Graph theoretical approach to structural representation of systems
Proceedings of the Fourth International Conf. for Pattern Recognition, held at Kyoto, Japan, 1978.

[3] T.Dénes, P.Gelléri: On the use of mathematics to sociology today
In: Sociology of Science and Research
Akadémia kiadó, megjelenés alatt

[4] O'Muircheartaigh, C.Payne /editors/ : Model fitting
J.Wiley, 1977.

[5] O'Muircheartaigh, C.Payne /editors/ : Exploring data structures
J.Wiley, 1977.