Appendix B Subtype Matrices

This appendix provides additional discussion on subtype graphs and sample populations, and assumes that Section 6.5 on subtyping has already been read. In general, access to the domain expert is required to resolve any doubts about subtyping requirements. In the absence of the subject matter expert, the subtype graph can be determined by *subtype matrix analysis* if a sample population is provided that is significant with respect to this graph. This appendix considers two ways of performing this analysis.

A subtype graph tells us which nodes are subtypes of which, but it does not provide the subtype definitions. Since infinitely many definitions are consistent with any finite set of data, the domain expert is needed to confirm the correct definitions. Moreover, in practice it is unrealistic to assume that a sample population is significant with respect to the subtype graph. Nevertheless, subtype matrices can be useful for the following tasks: deriving a tentative subtype graph; creating a population that is significant with respect to a known subtype graph; checking that a population is significant with respect to a known subtype graph.

A *matrix* is a rectangular array of elements. For example, a drawn game of "noughts and crosses" comprises a 3×3 array of "0" and "X" marks. Two kinds of matrix may be used to determine the subtype graph. We focus on the *partition-details matrix* technique, as developed by Dr. Eckhard Falkenberg. The other matrix is briefly discussed later.

Suppose that we have isolated some object type *A* for which a subtype graph is required. We begin by dividing *A* into groups where each member of any given group has exactly the same roles recorded. These groups are mutually exclusive, and collectively exhaustive of *A* (i.e. they form a *partition* of *A*). The following Euler diagram pictures *A* being partitioned into five groups.

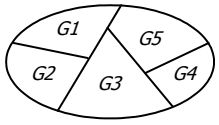

Figure B.1 *A* is partitioned into five groups.

Table B.1 A partition-details table indicates the details recorded for each group.

	Detail 1	*Detail 2*	*Detail 3*	*Detail 4*	*...*
Group 1					
Group 2					
Group 3					
...					

The list of details recorded for each member of A is the *recording pattern* for that member. Although members of the same group have the same recording pattern, different groups must have different recording patterns. Having partitioned A into groups on the basis of recording patterns, we display this on a partition-details table (see Table B.1).

As row headings we list the groups that comprise the partition. We may list these groups simply as "(1)", "(2)" etc. or, if their nature is obvious, we may use a descriptive name for each. As column headings, we list all the details that are to be recorded for at least some groups. We name each detail by asking what kind of information may be recorded for the group members (e.g., weight, age, sports played, sports enjoyed).

If there are n groups and m details we have an $n \times m$ grid of cells which we now fill in with either a "1" or a "0" using the following rule:

1 ≡ this detail may be recorded for this group
0 ≡ this detail must not be recorded for this group

Here "this detail" and "this group" mean the detail and group for that cell's column and row respectively. Thus the recording pattern for each group is indicated by the pattern of 1s and 0s on its row. When the whole table is filled in, we have an $n \times m$ matrix of 1s and 0s. As an example, consider the media survey application discussed in Section 6.5. The output report for this is reproduced in Table B.2. As usual, the mark "–" means "inapplicable because of some other data".

The roles relating to favorite channel, favorite newspaper, and preferred news source are optional. Our task is to determine the subtype graph for these optional roles. We first place the people into groups according to their recording pattern. For example, the persons with form numbers 5001 and 5002 go into the same group because they have exactly the same details recorded. Equivalently, they have exactly the same details not recorded. Clearly, *rows with the same pattern of "–" values go into the same group*.

Table B.2 An output report from the media survey.

Person	*Age (y)*	*Television (h/week)*	*Newspaper (h/week)*	*Favorite channel*	*Favorite paper*	*Preferred news*
5001	41	0	10	–	The Times	–
5002	60	0	25	–	The Times	–
5003	16	20	2	9	The Times	–
5004	18	20	5	2	Daily Mail	TV
5005	13	35	0	7	–	–
5006	17	14	4	9	Daily Sun	–
5007	50	8	10	2	Daily Sun	NP
5008	33	0	0	–	–	–
5009	13	50	0	10	–	–

Table B.3 The array of 1s and 0s is the partition-details matrix for the media survey.

	Age	*TVhours*	*NPhours*	*FavChannel*	*FavNP*	*PrefNews*
(1)	1	1	1	0	1	0
(2)	1	1	1	1	1	0
(3)	1	1	1	1	1	1
(4)	1	1	1	1	0	0
(5)	1	1	1	0	0	0

Though not demonstrated in this example, simple nulls (applicable but unknown) are treated as ordinary values when deciding the recording pattern (since the detail *may* be recorded). Using some obvious abbreviations, the partition-details table for our example is shown in Table B.3. The partition-details matrix itself is just the pattern of 1s and 0s. Here persons 5001 and 5002 form group (1); 5003 and 5006 form group (2); 5004 and 5007 form group (3); 5005 and 5009 form group (4); and 5008 forms group (5). Confirm this partition and the recording patterns for yourself. As you assign a member to a group, enter the group name beside its row in the output report. This provides a check that we haven't missed any rows. With our example, this gives:

(1) 5001 ...
(1) 5002 ...
(2) 5003 ...
(3) 5004 ...
(4) 5005 ...
(2) 5006 ...
(3) 5007 ...
(5) 5008 ...
(4) 5009 ...

Having obtained the matrix, we now find the subtype graph as follows. **Each node in the subtype graph corresponds to a distinct column pattern in the matrix.** By "column pattern" we mean the sequence of 1s and 0s in the column. We *label the different column patterns* as "A", "B" etc. Matching columns correspond to the same node, and are given the same label. Work from left to right: if a column pattern is new, write its label two spaces below the column; if the column matches an earlier one, write the earlier label one space below the column. With our example only the first three columns match, so this yields Table B.4.

Table B.4 The subtype nodes are labelled A .. D.

	Age	*TVhours*	*NPhours*	*FavChannel*	*FavNP*	*PrefNews*
(1)	1	1	1	0	1	0
(2)	1	1	1	1	1	0
(3)	1	1	1	1	1	1
(4)	1	1	1	1	0	0
(5)	1	1	1	0	0	0
		A	A			
	A			**B**	**C**	**D**

To understand why this procedure yields the nodes of the subtype graph, recall that the head of the graph has the common role(s) attached, and each subtype node has some specific role(s) attached. With our example, the column pattern for Age, TVhours, and NPhours shows that these are recorded for all the groups. So these three details correspond to the common roles attached to the head of the graph. So A is the head node.

Looking at column B we see that FavChannel is recorded *only for* groups (2), (3), and (4). So provided there is a *well-defined* way of determining membership in the union of groups (2), (3), and (4), we must make a proper subtype out of these groups. Assuming the output report is significant in this respect, this indicates that there is a subtype node with favors-TVchannel attached. So B is one of the subtype nodes.

Similarly, we can prove that there is a subtype node specifically to record favorite newspaper, and another subtype node to record preferred news source. So C and D are the other subtype nodes. Note that in general, **the details recorded for the nodes are indicated by the column headings for the nodes**.

We still have to *determine the subtype relationships between the nodes*. This can easily be done by comparing the column patterns for the nodes. Clearly, the groups included in a node are those heading the rows where the node's column value = 1. With our example, node A includes groups (1)–(5), node B includes groups (2) – (4), node C includes groups (1) – (3), and node D includes group (3). One node is a subset of another node just in case all its groups are included in the other node. So, given two nodes *X* and *Y*, it follows that *X is a subset of Y if and only if, for every row where X = 1, Y = 1*.

We use this rule to determine the proper subtype relationships between the nodes. To be systematic, we work our way through all such relationships, beginning with the largest node. With our example, this yields the following results: A has proper subtypes B, C, D; B has proper subtype D; C has proper subtype D. Check this for yourself. These relationships are portrayed in Figure B.2. To help you understand the example, the groups included in each node are also listed in the figure.

The next step is to **delete all transitively implied subtype relationships**. Our graph includes an arrow to explicitly show that D is a subtype of A. But this is transitively implied, since D is a subtype of B, and B is a subtype of A. So we delete this redundant link from the graph. At last we have the subtype graph for our example. With larger examples there may be more than one transitively redundant link to delete. Since we already know which details are recorded for each node, we can sketch these in too, as shown in Figure B.3. For compactness, the details are displayed as attributes.

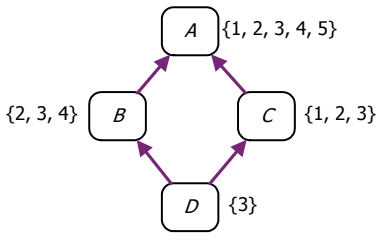

Figure B.2 The subtype graph obtained from Table B.4.

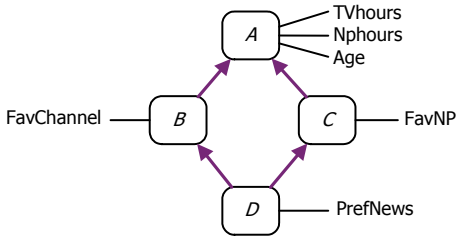

Figure B.3 The implied subtype link is deleted, and details are added for the nodes.

As discussed earlier, the head node A is the entity type Person. But we still need to **provide definitions and meaningful names for the subtypes**. To do this we go back to the original output report and ask what criteria are used to determine whether a specific subtype role is recorded. For any given subtype these criteria must be expressed in terms of one or more roles attached to its supertype(s).

By analyzing the data we can usually make good, educated guesses; but these must be confirmed with the domain expert. For example, what is the relevant difference between the people 5004 and 5007, for whom preferred news source is recorded, and the people 5003 and 5006? By inspecting Table B.2 we might note that one difference is that 5004 and 5007 are at least 18 in age (i.e., they are adults), whereas 5003 and 5006 are below 18 in age. Another difference is that 5004 and 5007 read newspapers for at least 5 hours per week. You may like to suggest other possible criteria.

Alternative hypotheses may often be rejected by testing them against other sample populations. As discussed in Section 6.5, input forms sometimes implicitly contain the definitions. However, in the absence of such documentation, the only safe way to determine the subtype definitions is to check with the domain expert.

Refer back to Figure 6.53 for the detailed conceptual schema. Note that the nodes in the subtype graph are quite different from the groups that were formed in the original partition. It is obvious from the value constraints and subtype definitions that Viewer and Reader are neither exclusive nor exhaustive. If desired, a check on exclusion and exhaustion constraints can be made by examining exclusion and exhaustion between the column patterns of the subtype nodes in the matrix: this optional check depends on the sample being significant with respect to these constraints.

Assuming a significant population is provided, the subtype matrix procedure just discussed may be summarized thus:

1. *Construct the partition-details matrix for the object type A:*
   List the details recorded for any members of *A* as column headings.
   Partition *A* into groups according to their recording patterns,
   and list these groups as row headings.
   Enter the recording pattern for each group
   (1 = detail may be recorded, 0 = detail must not be recorded)

2. *Sketch the subtype graph and attached details:*
   Label the different column patterns as nodes *A*, *B* etc.
   Determine the subtype relationships between the nodes as follows:
   node *X* is subset of node *Y* iff for each row where $X = 1$, $Y = 1$.
   Draw the subtype graph and delete any transitively implied arcs.
   Attach to each node the detail(s) indicated by its column heading(s).

3. *Provide definitions and meaningful names for the subtypes:*
   Define each subtype in terms of one or more roles attached to its supertype(s).

4. *Complete the conceptual schema:*
   Fill in the fine details of the schema diagram, including subtype definitions.

Given a significant population, an alternative matrix technique developed by Dr. Dirk Vermeir may also be used to determine the subtype graph. The associated matrix is called a *details-details matrix* because the details recorded for any member of the head supertype are listed both as row and column headings. To fill in the values of the matrix, we ask this question for each cell: *for each object instance for which we record the row detail, must we record the column detail? If the answer is Yes we enter "1", else we enter "0".*

As an example, the details-details matrix for the media survey example is shown in Table B.5. It should be obvious that the diagonal running from the top left corner to the bottom left corner must always be filled in with 1-values (Why?). As with the earlier matrix, matching columns correspond to the same node, and different column patterns correspond to different nodes in the subtype graph.

In a similar manner to the other matrix, subtype relationships between the nodes are determined by comparing the column patterns of the nodes, and the details recorded for the nodes are given by the column headings for the nodes. Check for yourself that this matrix gives the same result as the earlier one.

Unlike the partition-details matrix, the details-details matrix cannot be used to check on exclusion and exhaustion constraints. It is also somewhat less intuitive. However it is quite simple to compute, and it is idempotent with respect to Boolean multiplication (i.e., when multiplied by itself using logical rather that arithmetic operations, the product is identical to the original matrix). This property may be used as a check on the matrix. Testing this property is tedious and is best handled by automation.

Table B.5 A details-details matrix for the same application.

	Age	*TVhours*	*NPhours*	*FavChannel*	*FavNP*	*PrefNews*
Age	1	1	1	0	0	0
TVhours	1	1	1	0	0	0
NPhours	1	1	1	0	0	0
FavChannel	1	1	1	1	0	0
FavNP	1	1	1	0	1	0
PrefNews	1	1	1	1	1	1
		A	A			
	A			**B**	**C**	**D**

To use the details-details matrix instead of the partition-details matrix, replace steps 1 and 2 in the subtype matrix procedure with the method just described. Whichever kind of matrix is used, remember that each assumes the sample is significant with respect to the subtype graph.

Exercise B1

1. For a department, information of the kind indicated by the following output report is to be maintained. Departmental offices are identified by a room number. If you are unfamiliar with the metric system, note that 1 inch is about 2.5 cm (e.g. 5' 10" = 175 cm). The population shown is significant. The mark "–" means "inapplicable". Schematize this UoD, using a subtype matrix to help decide on the subtype graph.

Member	*Gender*	*Smoker ?*	*Starsign*	*Height (cm)*	*Office*	*Sports*
Adams	F	Y	Aquarius	–	–	–
Brown	M	Y	–	170	–	–
Collins	M	N	–	190	308	basketball
Davis	F	Y	Gemini	–	–	–
Evans	M	Y	–	190	–	–
Fomor	M	N	–	180	505	basketball, tennis
Gordon	F	N	Aquarius	–	406	–
Hastings	M	Y	–	165	–	–
Iveson	M	N	–	165	305	–
Jones	M	N	–	179	502	–

2. Use a subtype matrix to determine the subtype graph for Exercise 6.5, Question 5.