Electronic Journal of Statistics
Vol. 9 (2015) 2202–2254
ISSN: 1935-7524
DOI: 10.1214/15-EJS1067

Survey questionnaires and graphs

Ioana Schiopu Kratina[*,†]

Dep. of Mathematics and Statistics
University of Ottawa
585 King Edward, Ottawa, Ontario, Canada K1N 6N5
e-mail: ioanakratina@gmail.com

Christina Maria Zamfirescu[‡]

Dep. of Computer Science and Mathematics
Hunter College and Graduate Center-CUNY
New York, NY 10065 and 10036, USA
e-mail: zamfichris@gmail.com

Kyle Trépanier[*,§]

Dep. of Mathematics and Statistics
University of Ottawa
585 King Edward, Ottawa, Ontario, Canada K1N 6N5
e-mail: kyle_trepanier@hotmail.com

and

Lennon Marques[‡,¶]

EDI Options, Inc.
Mineola, NY, 11501, USA
e-mail: lennon.marques@macaulay.cuny.edu

Abstract: We introduce a special type of graphs, which we use as tools for designing and improving survey questionnaires. While the idea of representing questionnaires as graphs is not new, our graphs constitute novel theoretical and practical tools, which could turn a complex questionnaire into a questionnaire that is easier to visualize, test and analyze.

MSC 2010 subject classifications: 94C15, 60C05, 62D05.
Keywords and phrases: Questionnaire design, directed acyclic graphs, statistical analysis, finite probability space, expected number of questions.

Received September 2014.

[*]Ioana Schiopu Kratina and Kyle Trépanier were partly supported by a Natural Sciences and Engineering Research Council of Canada grant.

[†]Ioana Schiopu Kratina's work was partly supported by Statistics Canada.

[‡]Support for Christina Zamfirescu and Lennon Marques' work was provided by a PSC-CUNY Awards, jointly funded by The Professional Staff Congress and The City University of New York.

[§]Kyle Trépanier was partly supported by an Undergraduate Student Research Award, from the Natural Sciences and Engineering Research Council of Canada.

[¶]Support for Lennon Marques' work was also provided by Internships/Undergraduate Research Opportunities Fund from Macaulay Honors College while an undergraduate at Hunter College of The City University of New York.

1. Introduction

1.1. Background and objectives

Questionnaires are essential instruments of data collection used by statistical agencies all over the world, in both public and private sectors. The success of a survey, evaluated in terms of costs and quality of the collected data, depends in a large measure on how well the survey questionnaire was designed.

Data collection is always the most expensive part of running a survey. For computer assisted interviews, it may include the costs of testing and programming the questionnaire, interviewer training, interviewing respondents and converting nonrespondents, and the cost of editing and correcting errors at every stage. The quality of the published data also depends on the design and length of the questionnaire. In a well designed questionnaire, the content of questions is clear and each question is addressed only to its target population (called the *coverage* of the question). Furthermore, for each surveyed subpopulation, the questions are sequenced in a logical order, to ensure a high response rate (Section 9.91, p. 292, Handbook of Survey Research, 2010). A high response rate is generally an indicator of good quality data, whereas low response rates distributed unevenly among different surveyed subpopulations point to potential bias in the analysis. To foster complete responses, the questionnaire should also be as short as possible.

For a large survey of general interest, the content and number of questions is usually determined by a group of analysts, who may represent different organizations and are backed by different sponsors. Consequently, the range of topics covered by the questions could be quite wide. The design of the questionnaire must accommodate the requests set by the analysts, and must also comply with basic requirements. For instance, it should be apparent from the questionnaire "who is asked what". The difficult part of designing the questionnaire consists of placing these diverse questions in the "proper" order. One can view the design of the questionnaire as a solution of a puzzle, where each question with its coverage is a piece of the puzzle, and the challenge is to optimally place these pieces "on a board", so as to satisfy some logical constraints. The optimality criterion essentially requires that the questionnaire inflict a minimum response burden.

We illustrate our results on questionnaires designed for computer assisted telephone interviews, but the approach and the methods we advocate here apply to any type of survey questionnaire.

The growing capability of computers has opened the door to the development of large and complex questionnaires. Such questionnaires are split into thematic sections called *modules*, which, to a large extent, can be programmed and analyzed separately. All our examples are based on a simplified version of the module "Most Recent Employment" (henceforth abbreviated **EM**), one of the 47 modules of the "Access and Support to Education and Training Survey 2008, Questionnaire" (ASETS) http://odesi1.scholarsportal.info/documentation/ASESTS2008/asets2008_Questionaire.pdf (see Appendix A for **EM**).

The length of the ASETS modules varies from the 4 questions of the module "Information on learning" to the 24 questions of the module "Parental Savings". We note that not all questions have full coverage, i.e., some questions are addressed to subpopulations of the surveyed population. The questionnaire for such complex surveys would have to be very well designed to extract a good response rate and to enable the analysis of the data. The design is generally done by experts, who essentially rely on common sense and their extensive experience. Practical rules for structuring a questionnaire do not seem to have been prescribed. We introduce a structural approach to designing questionnaires using what we call *survey charts*, patterned after the *flow charts* used in survey practice. An example of a flow chart is **EM From Survey with condition nodes**, a visual representation of the simplified form of the questionnaire of **EM** (both in Appendix A). The corresponding survey chart is **EM From Survey** (flow count-down or flow count-up, both in Appendix A).

The survey chart is a tool with many uses (see Jabine (1985)). In the first place, it helps verify that the analytical requirements are met (the text and coverage of each question is correct). In the second place, visualizing the *paths* taken by different subpopulations within the survey chart helps program these paths, which should all reach the end of the questionnaire. Last but not least, the survey chart is a graphical aid in the analysis of the collected data. Survey charts, which we undertook to study here, are *graphs* with specific properties. Given a survey chart, we propose various transformations to it, so that the questionnaire represented by the transformed survey chart better fulfils its role, especially as an analytical tool.

1.2. Some previous related work

There is a vast literature that deals with the cognitive aspects of questionnaire design. A succinct presentation is found in chapter 9 of Handbook of Survey Research (2010). A classical guide on question formulation and complexity is Payne (1949). While cognitive research is of paramount importance to questionnaire design, we are interested here in the structure of questionnaires represented by graphs. Nonetheless, as we point out in several places, one should be aware of the connection between question complexity, the ordering of questions and the role of screening questions, on the one hand (Section 9.9 of Handbook of Survey Research, 2010), and our proposed transformations, on the other hand. Picard (1965) was the first author to view and study questionnaires as graphs, whereupon he identified the questions of a questionnaire with the nodes of the associated graph. In his set-up, the information obtained from the questionnaire partitions the surveyed population into disjoint categories, called *éventualités* (events), and each such event is assigned a positive probability of occurrence. Thus, each questionnaire generates a finite probability space. Picard's main objective was to define and construct an optimal questionnaire corresponding to a given finite probability space. To attain this goal, he performed a series of changes on graphs associated with questionnaires. These changes consist of

switching the position of questions or clusters of questions within the graph, based solely on technical considerations. His optimal questionnaire has a minimum number of expected questions, a measure of the efficiency with which the information is collected, defined in Picard (1965). Picard's algorithm for obtaining an optimal questionnaire is similar to the algorithm devised by Huffman (1952) for use in coding theory, which is also discussed in Parkhomenko (2010).

So far, Picard's seminal work on the design of optimal questionnaires has not been applied. It could be because, on the one hand, his problem-setting is technically too restrictive, while, on the other hand, the content of questions and the logic inherent in sequencing them is totally ignored in his approach. Indeed, the sole restriction that Picard imposes on his transformations is that the number of possible answers to each question remains unchanged.

Despite Picard's pioneering work on the connection between questionnaires and graphs and the proliferation of complex questionnaires, only a small number of papers that exploit this connection have been published since, some of which are listed next. McCabe (1976) defines a measure of complexity of the structure of a graph, the cyclomatic complexity. This number is used primarily to identify graphs which are hard to program. For most surveys designed by statistical agencies, the cyclomatic complexity is small and it does not help identify questionnaires that need to be redesigned.

Jabine (1985) brings forth the importance of using flow charts, which are standard graphical tools used at various stages in the development of the questionnaires. However, he suggests no systematic way in which flow charts could be used in designing questionnaires. Parkhomenko (2010) modeled organizational hierarchies (and questionnaires) as graphs, with managers viewed as nodes of these graphs. Picard's technical restrictions on moving nodes with their branches within the graph are not required here. Furthermore, these transformations are fully justified in the context of actual organizational hierarchies. Parkhomenko associates costs with each node, which allows for even more flexibility in modeling real life situations.

A different approach to questionnaire design is possible using decision trees, which are tools that can quickly classify data using relatively simple tests on the data set (Murphy, 1998, p. 345). At each node, a decision to split into new branches is taken, based on selecting the split with optimal value of a goodness measure (Murphy, 1998, p. 347). The goodness measure is used to compare numerically all possible splits, according to a predetermined criterion. The choice of the criterion to use (e.g., simplicity of questions, minimizing the expected number of questions etc.) has been covered in several papers (see Kotsiantis, 2013). Discussion on the optimal decision tree design can be found in Safavian and Landgrebe (1991). To each questionnaire one can associate a graph with questions as nodes. At each node, the coverage of the question splits into branches, determined by the response given to the question. The questionnaire ultimately classifies the surveyed population into small categories, which can be determined by considering all possible answers to all questions in the questionnaire. We can associate conditional probabilities between nodes as described

in our Proposition 1. Thus, we can view a questionnaire as a deterministic or stochastic decision tree. This descriptive approach can help in the analysis of a questionnaire after the data has been collected. Conversely, starting with all small categories the questionnaire should produce and selecting a goodness measure, one could "build up" a decision tree and then a questionnaire, following the steps in Murphy (1998), p. 349. However, this does not help us solve the problem at hand, which is to place predetermined questions in a suitable order. The reason is that the questions generated by the building up process could be quite awkward and may have very little in common with the original questions (see the example in section 2.3 of Fenn, 2015, or our section 7).

Graphs have also been used for the purpose of storing and testing electronic questionnaires. In Bethlehem and Hundepool (2004) the capabilities of the system TADEQ (Tool for the Analysis and Documentation of Electronic Questionnaires) is described. It is primarily used for storing and generating questionnaire documentation. In Elliott (2012), graphs are used mostly for testing electronic questionnaires. Other than the work of Parkhomenko (2010), little seems to have been done on the important topic of systematically designing a questionnaire which incurs a minimum response burden.

1.3. Formal definitions related to graphs

We start with definitions from graph theory, for which we refer to Harary (1969) and Picard (1965). A directed graph, or *digraph*, D, consists of a finite set V of *points* or *nodes* (used interchangeably, as has been done in the literature) and a set of ordered pairs of distinct points. Any such pair (x,y) is called an *arc*, and is denoted xy. The arc xy goes from x to y, and we say that x is *adjacent to* y, and y is adjacent *from* x. In addition, we say that xy is incident with both x and y, and x and y are the endpoints of xy.

The *outdegree* of a point x is the number of points adjacent from x. Equivalently, it is the number of arcs that go from x to some other points. The *indegree* is the number of points adjacent to x. Equivalently, it is the number of arcs that go to x from some other points. A point of indegree zero is called *source*.

The concept of being planar, generally defined for undirected graphs, can be easily defined for digraphs. A *planar* graph is a graph that can be embedded in a plane, in other words, it can be drawn on the plane in such a way that its arcs intersect only if they have a common endpoint.

A *path* in a digraph D is a sequence $x_1x_2 \dots x_n$ of distinct points of D such that x_ix_{i+1} is and arc in D, for all $i = 1, 2, \dots, n-1$. When also x_nx_1 is and arc in D, the sequence is also called a *cycle*.

A digraph T with vertex set V is called a *directed tree* of *root* $x_0 \in V$, when x_0 is a source, the indegree of x is 1, for all $x \in V, x \neq x_0$, and T contains no cycles. Since all graphs in our paper are directed, we will call directed trees simply *trees*.

A digraph is called *acyclic* when it contains no cycles.

Let T be a tree with the root x_0. It can be easily seen from the definition that every point x in T is on a single path from x_0 to x. The number of arcs in this path is called the *rank* of x.

To each acyclic digraph with only one source one can attach an associated tree by repeating the nodes with indegree larger than 1 (see figures 3 and 4).

The next definitions apply to all acyclic digraphs. Note that since all graphs in our paper are directed, unless otherwise specified, we will simply call them graphs.

When there is a path from x to y we call x an *ancestor* of y, and y a *descendent* of x.

For all points x in an acyclic graph A, if xy is an arc in A, then the point x is called a *parent* of y, and y is called a *child* of x. In a tree, the root has no parent, while any other node has exactly one parent and may have several children. In an acyclic graph which is not a tree, a node may have any number of parents, and a parent may be the parent of another parent. Similarly, a child may be the child of another child. For this reason, in the literature, when dealing with acyclic graphs which are not trees, a parent is often called a *direct ancestor*, and a child a *direct descendent*. However, for the sake of simplicity, in this paper we will call them parent and child. A node with no children is called *terminal*.

1.4. Questionnaires as survey charts

Consider now a questionnaire q, which consists of questions with their coverage and rules that define the order of asking these questions. Each question Q in q is characterized by its *content* and its coverage. The content of a question Q is expressed in its text, and the coverage is the subpopulations of the surveyed population which is asked Q. The text of Q solicits *information* from the surveyed individuals who constitute the coverage of Q. The questions of **EM** can be found in Appendix A (the bold text is what is read to surveyed individuals). These questions with their labels have been taken from the actual survey. In q, the questions are sequenced in the order in which the interview takes place, for each subpopulation of the surveyed population. This is not a complete order, and the rules for labeling the questions are quite loose.

A *survey chart* A associated with q is a directed acyclic graph with the additional properties that it has a unique source, called root, and denoted R (or Q_0), and a unique terminal node denoted END_A. The nodes of A are the questions of q. The arc Q_aQ_b, $a < b$ represents the fact that a subpopulation of the coverage of Q_a (i.e., which was asked Q_a) is asked Q_b next. The population that "travels" through Q_aQ_b is defined by the respondents' answer to Q_a and possibly by answers to some ancestors of Q_a.

In A we have the following *connectedness* property, introduced in Bethlehem and Hundepool (2004):

For each node x in A, R is an ancestor and END_A is a descendent. This means that every question Q in q is asked (the coverage of Q is not empty) and that the individuals that make up the coverage of Q "travel" along a path of A from the beginning to the end.

An interesting property of survey charts not shared by graphs in general, is that they may contain empty paths, i.e., paths that are not "travelled". For instance, in **EM From Survey** there is a path from Q_{10A} to Q_{18}, but no population travels through this path, because the coverage of Q_{10A} and the coverage of Q_{18} are disjoint sets. Expanding survey charts as trees (see Figures 3 and 4) is not a better alternative. The tree corresponding to large questionnaires like that of ASETS is huge, and clever storage methods for this graph and the graphs of its modules has to be devised, as described in Bethlehem and Hunderpool (Bethlehem and Hundepool, 2004). This expansion defies the very purpose of the survey charts-to visually and succinctly represent questionnaires. Thus, having to deal with empty paths is the price we have to pay for "visually packing" the questionnaire in a survey chart rather than a tree.

The flow charts used in survey practice differ from the survey charts (which we promote), in that flow charts also contain *conditions nodes*, e.g., C_7, C_{18} and C_{24} in **EM From Survey with condition nodes** (Appendix A). These condition nodes are mainly used for programming, but they are misleading when determining the paths different populations actually take within the flow chart. For instance, C_{24} introduces the "empty" arc $C_{24}Q_{25}$, which should actually coincide with the arc $Q_{23}Q_{25}$. To create the survey chart $A =$ **EM From Survey** from **EM From Survey with condition nodes**, we removed the nodes C_7, C_{18} and C_{24} and replaced all their incident arcs with arcs incident to questions, which are actually traveled by surveyed individuals.

To simplify the presentation, we did not take nonresponse into account in **EM**. In practice, there is nonresponse to every question in the questionnaire and the design has to account for it. Nonrespondents must be treated just as any other subpopulation of the surveyed population, i.e., be assigned arcs on which they can travel. So, there is no loss of generality in our ignoring nonrespondents as yet another specific subpopulations of the surveyed population. Nevertheless, we present in Section 4.4 an application specific to nonrespondent subpopulations, which results in a decrease in the response burden of the questionnaire.

1.5. Our contribution

In this paper we introduce survey charts, graphical tools originating from survey practice, which we use to represent and transform survey questionnaires. Although our set-up is close to that of Picard (1965) or Parkhomenko (2010), our approach differs from theirs in some fundamental ways, specifically in the purpose of our research, the definitions that we give, and the procedures that we propose and follow. We succinctly present these differences here, with details given in section 7. Our survey charts are graphs with special properties, and cannot generally be identified with trees (the type of graphs used by Picard (1965) or Parkhomenko (2010)), either before or after we apply our transformations. We endow each survey chart with a finite probability space, richer than Picard's (1965) in that it has "points", or *analytical outcomes*, which capture the entire

information collected by the questionnaire and are indispensable for a complete analysis. We propose our transformations with several objectives in mind. We intend to ease the verification of the coverage of questions ("who is asked what"), to streamline the questionnaire, so that the important analytical paths are easier to discern, and to reduce, whenever possible, the expected number of questions as defined in Picard (1965). Unlike Picard (1965), we term *information* what is extracted from the responses (or nonresponse) to each question in the questionnaire. This information, collected over the entire questionnaire, is what drives and also restricts the type of transformations that we propose. We also pay special attention to the problem of nonresponse. In addition to minimizing, via repeated transformations, the expected number of questions, we dedicate section 4.4 of the paper to the applicability of a proposed transformation to reducing nonresponse.This paper is addressed primarily to the technical personnel responsible for the impact of the questionnaire on the collected data. The key players are: the subject matter specialists, who represent the analysts, the programmers, the interviewers' representative and the project manager, who is also responsible for the statistical methodology used. Ultimately, this paper is addressed to practitioners, in the hope that it will help them modify a questionnaire in its early stages of development and make it more amenable to testing and statistical analysis. Our proposed transformations are illustrated by several examples based on an actual questionnaire.

The remainder of the paper is organized as follows. Section 2 defines the probability space we associate with survey charts and the expected number of questions. Section 3 gives an overview of our transformations, detailed in sections 4–5. Section 6 motivates the design of the survey chart **EM Analytical**, which we recommend as the result of our proposed transformations. In section 7 we compare various methods. In section 8 we draw conclusions and suggest future work. Appendix A presents the text of the questionnaire **EM** studied here, its flow chart, and graphs associated with its survey chart. Appendix B contains an algorithm that counts the number of flows required in the calculation of probabilities of questions. Appendix C contains technical details that complement sections 2 and 4. Appendix D presents and analyzes the survey chart **EM Analytical**, as well as the flow chart **EM Simplified**, which results from **EM Analytical** by eliminating some questions. Finally, Appendix E illustrates Parkhomenko's algorithm A1 on a simple example.

2. The probability space and the expected number of questions

2.1. Analytical outcomes and elementary events

Assume that we have a questionnaire q, which we identify with a survey chart A. Assume the questions are labeled $Q_j, j = 0, \ldots, M-1$, where the labelling is such that, if Q_iQ_j is an arc in A, then $j > i, i = 0, \ldots, M-2$. Each questionnaire generates a set Ω_A of *analytical outcomes*.

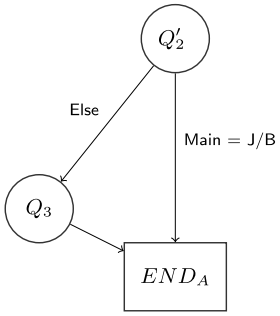


FIG 1. *Example 1*

Definition 1. An *analytical outcome*, denoted by ω, is a category of individuals defined by completing the questionnaire, which cannot be divided into further categories based on the information collected by the questionnaire. The set Ω_A of analytical outcomes constitutes the *analytical potential* of the questionnaire.

Examples of analytical outcomes appear throughout the paper, e.g., in Examples 1 and 1′.

We base all our examples on **EM** (Appendix A), which collects information on the most recent job or business (*MRJ/B*) that selected individuals held during the reference period (*R.P.*). To obtain **EM**, we eliminated from the original version of "Most Recent Employment" question Q_{17} and retained all others, along with their original text and labels. We use the abbreviation *J/B* for *job or business*. As is customary, bold letters indicate that the text is read to the respondent.

In Example 1 we present several analytical outcomes associated with a very short questionnaire.

Example 1. Consider a questionnaire q with two questions, Q_2' and Q_3, where Q_2', the root, is a simplified version of Q_2 in **EM**, and Q_3 is taken from **EM** (see Appendix A):

Q_2': **During the *R.P.*, what was your main activity?: working at a *J/B*, doing volunteer work, going to school, taking care of family or household responsibilities, other...**

Q_3: **Did you work at a *J/B* at any time during the *R.P.*?**

If the answer to Q_2' is *J/B*, no further question is asked; otherwise, the last question Q_3 is asked. The survey chart associated with q is presented in Figure 1. Let $\omega_1 = \{Q_2' = J/B\}$ be the set of surveyed individuals whose main activity during the *R.P.* was to work at a *J/B*, and ω_1^c its complement, i.e. the set of individuals whose main activity during the (*R.P.*) was not working at a *J/B*. While ω_1 is an analytical outcome, neither $\omega_1^c \cap \{Q_3 = yes\}$ nor $\omega_1^c \cap \{Q_3 = no\}$ is. In fact, $\omega_1^c \cap \{Q_3 = yes\}$ is a union of the analytical outcomes ω_i, $i = 2, 3, 4, 5$

succinctly represented as:

$$\begin{aligned}
\omega_2 &= \{Q_2' = \text{volunteer work}\} \cap \{Q_3 = yes\} \\
\omega_3 &= \{Q_2' = \text{going to school}\} \cap \{Q_3 = yes\} \\
\omega_4 &= \{Q_2' = \text{taking care of family/household}\} \cap \{Q_3 = yes\} \\
\omega_5 &= \{Q_2' = \text{other}\} \cap \{Q_3 = yes\}
\end{aligned}$$

Altogether there are 9 analytical outcomes that form the analytical potential of q.

The survey chart **EM From Survey** (Appendix A) is obtained from the original flow chart of **EM** by eliminating the conditions as nodes. The survey chart **EM Analytical** (Appendix D) is obtained from **EM From Survey** by applying the transformations described in Sections 3–4. It has the same analytical potential as **EM From Survey**.

We now construct a probability space appropriate for the analysis of the information collected by the questionnaire. We assign to each $\omega \in \Omega_A$ a probability $P_{analytical}(\omega)$, and consider the probability space $(\Omega_A, \wp(\Omega_A), P_{analytical})$, where $\wp(\Omega_A)$ is the field of all subsets of Ω_A (the power set), i.e., the field of events generated by the analytical outcomes. One can define $P_{analytical}(\omega), \omega \in \Omega_A$ to be the proportion of the surveyed population classified into the category represented by ω, and then calculate the probability of all other events in $\wp(\Omega_A)$. Estimates of $P_{analytical}(\omega), \omega \in \Omega_A$ are not very accurate, since many analytical outcomes could consist of a small number of individuals. Estimating the probabilities of larger sets using real proportions would give better results (see sections 2.2 and 3).

Our purpose here is to analyze the structure of A in order to simplify it, before the questionnaire is tested and programmed. We therefore work on a different probability space, more in line with Picard's (1965). Unlike Picard (1965), we have a set Ω_A of analytical outcomes as above, but the field of events is now generated by *elementary events* $F_i, i = 1, \ldots, N$ (called éventualités in Picard's (1965)), or *flows*.

Definition 2. A *flow* f_i of length $n_i + 1$ in A consists of a string of questions (nodes) connected by arcs, which starts at the root R and ends with a last question Q_{in_i} before reaching END_A, $i = 1, \ldots, N$. Two flows are different if there is at least one node that belongs to one of the flows, and not to the other.

The conditions that define all *the arcs of a flow* are logically consistent, which means that, theoretically, there exist a subpopulation that can "travel" along that flow. In practice, though, one may find at the end of collection that some well-defined flows have not been traveled, either because the selected sample did not contain the "right" individuals, or because of nonresponse.

We write a flow as $f_i = RQ_{i1} \ldots Q_{in_{i-1}} Q_{in_i} F_i$, where $R = Q_{i0}$, $Q_{ij}Q_{i(j+1)}$ is the arc connecting the nodes Q_{ij} and $Q_{i(j+1)}, j = 0, \ldots, n_i - 1$, Q_{in_i} is the last question before END_A on the flow f_i, and F_i is an *elementary event*, $i = 1, \ldots, N$. The arc $Q_{ij}Q_{i(j+1)}$ (or the population that travels through it) is uniquely identified by the respondents' answers to question Q_{ij} and possibly to

questions that precede Q_{ij}. The *elementary event* F_i stands for the category of individuals defined by their answers to all questions on the flow $f_i, i = 1, \ldots, N$ and is part of END_A.

We identify each flow f_i with the corresponding elementary event F_i, $i = 1, \ldots, N$. In a sense, the order in which questions along a flow are asked is irrelevant in defining a flow. In modifying a survey chart, we can change the order of questions on a flow in the initial survey chart, to obtain a more suitable representation of this flow in the new survey chart.

There is an important difference in a survey chart between flows and paths. Although every arc on the survey chart can be traveled, some longer paths on the survey chart cannot be traveled by any subpopulation, because the conditions that define different arcs on the paths are contradictory. In **EM From Survey** there is no flow that passes through the path from Q_8 (addressed to self employed workers in *MRJ/B*) to Q_{18} (addressed to employees in *MRJ/B*) because these two populations are disjoint. Thus, every flow follows a path of A that starts at R and continues to END_A, but not every path of A lies on a flow in the survey chart. A path on a survey chart can be on any number of flows, including zero (see **EM From Survey**, flow count-down or flow count-up, Appendix A).

The measurable space for analyzing the structure of A is $(\Omega_A, \sigma_A(F_i, i = 1, \ldots, N))$, where $\sigma_A(F_i, i = 1, \ldots, N)$ is the field of *events* generated by the elementary events $F_i, i = 1, \ldots, N$. We define the event $END_A = \Omega_A = \bigcup_{i=1,\ldots,N} F_i$. On the survey chart, END_A is a terminal node, and so, on the survey chart, every F_i, $i = 1, \ldots, N$, is identified with this terminal node. We note that the elementary events F_i, $i = 1, \ldots, N$ form a partition of Ω_A.

The difference between analytical outcomes and elementary events is illustrated in Example 1′. Two possible probability measures on $\sigma_A(F_i,\ i = 1, \ldots, N)$ are discussed in Section 2.2.

Example 1′. In Example 1, an analytical outcome in **EM From Survey** is

$$\omega = \{Q_2' = \textit{volunteer work}\} \bigcap \{Q_3 = no\}.$$

While $\{\omega\} \in \wp(\Omega_A)$, $\omega \notin \sigma_A(F_i,\ i = 1, \ldots, N)$, because ω is only part of the elementary event $F_1 = \{Q_2' = J/B\}^c \bigcap \{Q_3 = no\}$. Note that F_1 represents the flow $f_1 = RQ_3F_1$, where $R = Q_2'$. The other flow is $f_2 = Q_2'F_2$. Thus, in this example, we have 9 analytical outcomes, of which 8 form the elementary event F_1 and 1 forms F_2.

2.2. Probabilities of flows and questions

Consider now the measurable space $(\Omega_A, \sigma_A(F_i, i = 1, \ldots, N))$, where each elementary event F_i is assigned a probability p_i of occurrence, $i = 1, \ldots, N$, $\sum_{i=1,\ldots,N} p_i = 1$. Events that are unions of elementary events (flows), e.g., $\bigcup_{i \in I(E)} F_i$, have probability of occurrence $P_A(E) = \sum_{i \in I(E)} p_i$. The END_A event and the root R corresponds to the union of all flows and have probability 1 of occurrence. We also have $P_A(RQ_{i1} \ldots Q_{in_{i-1}} Q_{in_i} F_i) = p_i$ for the flow corresponding to the elementary event $F_i, i = 1, \ldots, N$.

We explore two probability measures on $(\Omega_A, \sigma_A(F_i, i = 1, \ldots, N))$:

1. Probability measure 1 assigns to each elementary event of the surveyed population its true proportion in that surveyed population. These proportions are unknown, but good estimates can sometimes be found, when the elementary events consist of large populations. For small populations, estimates from censuses can be used. However, census data is often out of date, and the nonsampling errors may accumulate when these estimates are added. Alternatively, one could use sample estimates from similar surveys to estimate directly the probabilities of questions. Probability measure 1 can also be used with the probability space $(\Omega_A, \wp(\Omega_A))$ defined in Section 2.1.
2. Probability measure 2 assigns equal probabilities to all elementary events. We use this probability measure, also explored in Picard (1965), when no reliable estimators for the probabilities in 1 are available. Probability measure 2 is helpful in understanding the structure of the questionnaire. Its use is justified when the number of elementary events is large, but could be misleading otherwise.

The definitions and results below apply to any probability measure on $(\Omega_A, \sigma_A(F_i,\ i = 1, \ldots, N))$.

For the remainder of this section, we assume that each path in A is part of a flow. The general case will be treated in Section 2.3.

While paths leading from R to a question Q and paths leading out of Q to END_A can be visualized on the survey chart, we need to express them as events. Given a question Q and a flow that contains it, say $f_i(Q) = RQ_{i1} \ldots Q \ldots Q_{in_i} F_i$, where $Q = Q_{ij}$, we associate the events $f_i(Q+)$ and $f_i(Q-)$. The event $f_i(Q+)$ is the set of all flows formed by completing $RQ_{i1} \ldots Q$ to a flow, i.e., all flows of the type $f = RQ_{i1} \ldots QQ_{k(j+1)} \ldots F_{ik}$, for some $k = 1, \ldots, N$. Note that any flow $f \in f_i(Q+)$ contains Q and $f(Q+) = f_i(Q+)$. When $i \neq j$, the events $f_i(Q+)$ and $f_j(Q+)$ might be disjoint, but generate identical paths from Q to END_A, because we assumed here that every path is part of a flow. Let $OUT(Q)$ be the set of all indices associated with these paths, and write $card(OUT(Q))$ for its cardinality, which is also the number of flows in $f_i(Q+)$, for any flow f_i that contains Q, $i = 1, \ldots, N$. Note that the outdegree of Q, i.e., the number of arcs in A that leave Q, is generally smaller than $card(OUT(Q))$, since several flows may travel through the same arc. We define $f_i(Q-)$, $IN(Q)$ and $card(IN(Q))$ in a similar way. The indegree of Q in the survey chart is generally smaller that $card(IN(Q))$.

The equality $f_i(Q) = f_i(Q+) \bigcap f_i(Q-), i = 1, \ldots, N$ expresses the fact that any flow containing Q consists of part of an incoming flow from R to Q followed by part of an outgoing flow from Q to END_A.

In Example 2 we appeal to Figure 4 to visualise the theoretical concepts introduced above.

Example 2. Consider the survey chart on the left of Figure 4 and let us focus on Q_8. The flow $f_1(Q_8) = Q_7Q_5Q_8F_1$ gives $f_1(Q_8^+) = f_1(Q_8)$, $f_1(Q_8^-) =$

$f_1(Q_8) \cup Q_7Q_5Q_6Q_8F_3$ and $f_1(Q_8^+) \cap f_1(Q_8^-) = f_1(Q_8)$. The other flow that contains Q_8 is $f_3(Q_8) = Q_7Q_5Q_6Q_8F_3$ with $f_3(Q_8^+) = f_3(Q_8)$, and $f_3(Q_8^-) = f_3(Q_8) \cup f_1(Q_8) = f_1(Q_8^-)$. We note that, although $f_1(Q_8^+) \cap f_3(Q_8^+) = \emptyset$, $f_1(Q_8^+)$ and $f_3(Q_8^+)$ contain one flow each, so $card(OUT(Q_8)) = 1$. Similarly, $card(f_1(Q_8^-)) = card(f_3(Q_8^-)) = card(IN(Q_8)) = 2$. There is 1 flow that exits Q_8 and 2 that enter it.

Any path $Q_a \dots Q_b$ on a flow can be represented by an event, with a probability assigned to it, as described below. Let $p_{ij}(Q_a \dots Q_b)$ be the probability of a flow which starts with the flow labelled i in $IN(Q_a)$, "travels" from R to Q_a, continues through $Q_a \dots Q_b$ and then reaches END_A on the flow labeled j in $OUT(Q_a)$. For $Q_a \dots Q_b$ in general and for a single question Q in particular, we have:

$$P_A(Q_a \dots Q_b) = \sum_{\{i \in IN(Q_a)\}} \sum_{\{j \in OUT(Q_b)\}} p_{ij}(Q_a \dots Q_b),$$
$$P_A(Q) = \sum_{\{i \in IN(Q)\}} \sum_{\{j \in OUT(Q)\}} p_{ij}(Q) \tag{1}$$

We also obtain from (1):

$$P_A(RQ_{i1} \dots Q) = P_A(f_i(Q+)), P_A(Q \dots Q_{in_i}) = P_A(f_i(Q-)) \tag{2}$$

The *coverage* of the question Q in A is the event:

$$cover_A(Q) = \bigcup_{i \in IN(Q)} f_i(Q+) = \bigcup_{i \in OUT(Q)} f_i(Q-) = \bigcup_{\{i: Q \in f_i\}} f_i, \tag{3}$$

i.e., the union of all flows containing Q. Each elementary event F_i, which corresponds to the flow f_i, can be defined by a set of consistent attributes. Likewise, the coverage of Q is also defined by a set of consistent attributes. This set represents the subpopulation of the surveyed population which is asked question Q.

Example 3. We first illustrate formulas (1) and (2), where A is the survey chart in Figure 4, in which all paths are travelled. We take P_A to be probability measure 2, i.e., each of the 4 elementary events associated with A has probability $\frac{1}{4}$ of occurrence. We have $P_A(Q_5Q_6) = \frac{1}{2}$, as there is 1 flow entering Q_5 and 2 flows leaving Q_6. We have $P(Q_5) = 1$, as there is 1 flow entering Q_5 and 4 flows that exit it, i.e., 4 flows that contain Q_5. Furthermore, $P_A(RQ_5) = P_A(Q_7Q_5) = 1$, $P_A(Q_6Q_4Q_{10B}) = \frac{1}{4}$. To illustrate definition (3), we see that $cover_A(Q_{10B}) = Q_7Q_5Q_{10B}F_2 \cup Q_7Q_5Q_6Q_4Q_{10B}F_4$. The attributes that define the population that is asked Q_{10B}, i.e., $cover_A(Q_{10B})$ are: surveyed individuals who were employed or worked at a family business during their MRJ/B.

2.3. The expected number of questions

In this section, we adapt to our survey charts Picard's expected number of questions in a questionnaire, which measures the response burden incurred when the

questionnaire is administered. Ideally, one should minimize the expected number of questions, while still obtaining the information one seeks when administering the questionnaire, thus preserving the analytical potential of the questionnaire. Note that, in general not all paths that appear in the survey chart are "travelled", so we have to define this concept in terms of flows alone.

The *expected number of questions* in the questionnaire represented by a survey chart is:

$$E_A = \sum_{i=1,\ldots,N} n_i p_i \tag{4}$$

In (4), n_i is the number of questions on flow f_i, with associated probability p_i. Let us write:

$$P_A(Q_j) = P_A(cover_A(Q_j)), j = 0, \ldots, M-1 \tag{5}$$

for the probability that question Q_j is asked when the questionnaire is administered (see (3)). For the root R, we have $P_A(R) = 1$, and for every other question Q in the questionnaire $P_A(Q) \leq 1$.

When the number of elementary events is very large, we may opt for probability measure 2, under which $p_i = 1/N, i = 1, \ldots, N$ and then (5) gives:

$$P_A(Q_j) = N_j/N, \tag{6}$$

where N_j is the number of flows going through question $Q_j, j = 0, \ldots, M-1$. The formula (6) is further expanded in Proposition 1. One can show that (see Appendix C)

$$E_A = \sum_{j=0,\ldots,M-1} P_A(Q_j) \tag{7}$$

Calculating (4) then reduces to finding (5) for each question in the questionnaire.

Remark 1. More generally, we could minimize instead an expression similar to (7), in which each term is multiplied by the cost c_j associated with question j, $j = 0, \ldots, M-1$. The cost of a question could represent its real cost, e.g., in a health survey, when results of some medical tests are required. It could also represent the difficulty in answering the question, or its complexity. We consider here that all questions have equal costs and deal with expression (7) only. One should estimate (7) before the questionnaire is programmed, so that changes to the questionnaires could be made to reduce response burden, should this be the case.

We consider now the general case, where the questionnaire is addressed to K disjoint subpopulations, denoted $pop_k(A), k \in K$. To each $pop_k(A)$, we uniquely associate a *category of flows* as described below. Each subpopulation $pop_k(A)$ generates an induced subgraph A_k of A, $k \in K$, defined as follows. The vertices (nodes) of A_k are all questions Q of A such that $cover_A(Q) \supseteq pop_k(A)$ and the arcs of A_k are all the arcs of A, which are traveled by subsets of $pop_k(A)$, $k \in K$ (see Appendix A, the graphs of **EM From Survey**, Category 1, 2).

The flows of A_k constitute a *category of flows*, if every path of A_k is part of a flow of $A_k, k \in K$, or, equivalently, if every path from R to END_A is a flow. It is always possible to write all flows as a union of disjoint categories of flows. For instance, each individual flow could constitute its own category of flows. In order to efficiently count the number of flows that contain a specific question, it is preferable to have as few categories of flows as possible. In the case of only one category, we could use formula (1) to calculate this number. On the other hand, reordering questions in order to minimize the number of flow categories may increase the indegree of some questions, which makes the coverage of such questions hard to verify. Therefore, one has to balance attaining a small number of categories against the simplicity of the questionnaire. Categories of flows often correspond to analytically important populations.

There are three subpopulations that appear to be important in **EM From Survey**. These are: employees in *MRJ/B*, abbreviated *e.*, self employed in *MRJ/B*, abbreviated *s.e.*, and respondents who worked for a family business in *MRJ/B*, abbreviated *f.b.* They do not necessarily determine three categories of flows.

An algorithm that uncovers all flows in a survey chart, as well as the logical contradictions which lead to the formation of categories can be found in Trépanier (2013). We illustrate it in Example 4.

Example 4. Consider the survey chart in Figure 3 and let us uncover all its travelled paths and categories. Starting with Q_4, the population of the survey travels to Q_5, where it splits into three subpopulations: $\mathscr{S}_1 = \{$ s.e. with 1 J/B $\}$, $\mathscr{S}_2 = \{$e. or f.b. with 1 J/B$\}$, $\mathscr{S}_3 = \{$ e., f.b. or s.e. with > 1 J/B $\}$. We follow $\mathscr{S}_1$ to Q_8 and F_1 in END_A, and uncover the flow represented by F_1 in Figure 3 (right). We return to Q_5 to follow $\mathscr{S}_2$, which can only travel through Q_7 to Q_{10B} to complete the flow represented by F_2. We note that the path $Q_5Q_7Q_8$ cannot be travelled since the sets $\{$s.e.$\}$ and $\{$e. or f.b.$\}$ are disjoint and must travel within different categories of flows. We write $F_1 \in$ Category 1, $F_2 \in$ Category 2 and note that we have completely examined $\mathscr{S}_1$ and $\mathscr{S}_2$. We return to Q_5 to examine $\mathscr{S}_3$, which travels through Q_6 to Q_7 and splits into $\mathscr{S}_{3,1}$, and $\mathscr{S}_{3,2}$. The population $\mathscr{S}_{3,1}$ then goes through Q_8 to END_A and completes the flow $F_3 \in$ Category 1. The population $\mathscr{S}_{3,2}$ goes through Q_{10B} to F_4, with $F_4 \in$ Category 2. Thus, Category 1 $= \{F_1, F_3\} = \{s.e.\}$, Category 2 $= \{F_2, F_4\} =$ {e. or f.b.}.

We notice that **EM from survey with condition nodes** (Appendix A) has three categories of flows. An unnecessary category is introduced by condition C_7, viewed as a node. It generates two arcs adjacent to it, Q_5C_7 and Q_6C_7, and two arcs adjacent from it, C_7Q_8, C_7Q_7. No population can travel the path $Q_6C_7Q_8$, and so the populations *s.e.* with (abbreviated *w.*) 1 *J/B* and *s.e. w.* > 1 *J/B* must be in different categories. We also note that C_{24} as node creates 16 spurious flows that travel from it to Q_{25}, which have been already accounted for through the arc $Q_{23}Q_{25}$.

In **EM From Survey** when the conditions have been eliminated (see Appendix A), we have only two categories of flows: Category 1 contains *e.*, *f.b.*, and

respondents who held no job or business during the reference period. Category 2 consists of *s.e.* only.

Proposition 1 below shows how to calculate conditional probabilities and probabilities of questions in the presence of more categories. These probabilities are important to analysts, but also for calculating the expected number of questions before finalizing the questionnaire.

Let $C^{(k)}, k = 1, \ldots, K$ be the categories of flows that partition all flows of the survey chart. For each question Q and in each category k, we define $f^{(k)}(Q+) = \bigcup_{i \in C^{(k)}} {f_i}^{(k)}(Q+)$ and $f^{(k)}(Q-) = \bigcup_{i \in C^{(k)}} {f_i}^{(k)}(Q-)$. With the obvious meaning for $IN^{(k)}(Q)$ and $OUT^{(k)}(Q), k = 1, \ldots, K$, we have:

Proposition 1. *For each question* Q, $P(Q) = \sum_{k=1,\ldots,K} \sum_{\{i \in IN^{(k)}(Q)\}} \sum_{\{j \in OUT^{(k)}(Q)\}} {p_{ij}}^{(k)}(Q)$, *where* ${p_{ij}}^{(k)}(Q)$ *is the probability of a flow in category* k *, which contains* Q*, and is formed by putting together the incoming flow labelled* i *in* $IN^{(k)}(Q)$ *and the outgoing flow labeled* j *in* $OUT^{(k)}(Q), k = 1, \ldots, K$.

Let ${f_i}^{(k)}(Q)$ be a flow which contains $Q = Q_{ij}^{(k)}$. We have:

$$P(Q/{Q_{i(j-1)}}^{(k)} \ldots {Q_{i1}}^{(k)} R) = P({f_i}^{(k)}(Q+))/P({f_i}^{(k)}(Q_{i(j-1)}+)).$$

When $p_i = 1/N$, $i = 1, \ldots, N$,

$$P(Q) = N^{-1} \sum_{k=1,\ldots,K} card(IN^{(k)}(Q)) \times card(OUT^{(k)}(Q)),$$

$$P(Q/{Q_{i(j-1)}}^{(k)} \ldots {Q_1 1}^{(k)} R) = card(OUT^{(k)}(Q)) \times card(OUT^{(k)}(Q_{i(j-1)})).$$

where $P(Q|Q_{i(j-1)}^{(k)} \ldots Q_{i1}^{(k)})$ is the conditional probability of asking question Q, given that all questions on the string $RQ_{i1}^{(k)} \ldots Q_{i(j-1)}^{(k)}$ have been asked. The proof of this result can be found in Appendix C.

In Example 5 below we illustrate how to calculate probabilities of questions (in probability measure 2) when two categories are present.

Example 5. We apply the results of Proposition 1 on the survey chart A in Figure 2, which is reproduced twice. On the left graph in Figure 2, we are counting the flows of A from root to END_A, using the algorithm described in Appendix B. On the right graph, we count the flows of A from END_A to root. We recall from Example 4 that A has two categories of flows, Category 1 = { s.e. } and Category 2 = { e. or f.b. }, and A has N = 2 + 2 = 4 flows. To calculate probabilities of questions in probability measure 2, we use the survey chart at the left to count the flows into each question Q, and the survey chart at the right to count the flows out of Q. For Q_7, the scalar product of (1,2) and (1,1) is 3, so $P(Q_7) = \frac{3}{4}$. If we ignored the existence of categories, with 1 + 2 = 3 incoming flows and 1 + 1 outgoing flows, we would erroneously count 6 flows that go through Q_7, out of N' = 6, so $P(Q_7)$ would be erroneously calculated as 1. This is so because we counted as flows the paths $Q_4 Q_5 Q_7 Q_8 END_A$ and $Q_4 Q_5 Q_6 Q_7 Q_8 END_A$, which are not travelled.

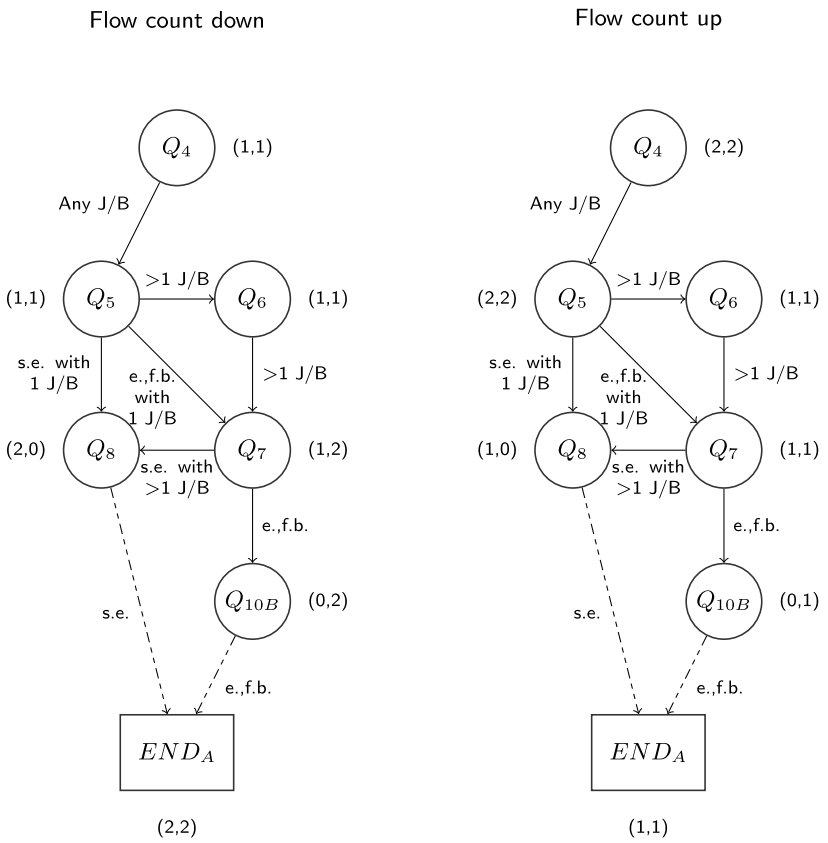


FIG 2. *Example 5*

3. Overview of transformations

Transformations 1 and 2, described in detail in sections 4–5, turn a survey chart A into a survey chart A', which is easier to test and use for analysis. Survey charts A and A' have the same analytical potential and A' has, at the most, the same expected number of questions as A. The transformations are illustrated by way of examples based on **EM**.

The next result shows the link between the shape of the survey chart and the ordering of the coverage of questions in the questionnaire.

Proposition 2. *Consider a survey chart A. The following statements are equivalent:*

(i) A *is a tree*
(ii) $indegree_A(Q) = 1$, *for all nodes* $Q \neq R$
(iii) For any node Q_a *and any of its descendents* Q_b, *we have* $cover_A(Q_a) \supseteq cover_A(Q_b)$

Proof. We first recall definition (3). The equivalence of (i) and (ii) follows directly from the definition of a tree in our context (see Section 1.3). Since from (ii), every descendent of Q_a in A has only one parent, it follows that (ii) implies (iii). Conversely, assume that (iii) holds. If (ii) did not hold, there would be a node Q_c with at least two parents, say Q_a and Q_b, and a flow $f_{ac} \in cover_A(Q_c)$, $Q_a \in f_{ac}$. If Q_b is not on f_{ac}, then $f_{ac} \notin cover_A(Q_b)$, which contradicts (iii). If Q_b is on f_{ac}, it is either an ancestor of Q_a or a descendent of Q_c. The latter cannot occur, because Q_b is also parent of Q_c and A is acyclic, so Q_b is an ancestor of Q_a. There exists then an arc Q_bQ_c, and thus a flow f_{bc}, which does not contain Q_a. We have that $cover_A(Q_a)$ does not include $cover_A(Q_c)$, which contradicts (iii) and proves the result. □

Transformation 1 reduces the expected number of questions by switching the position of two related questions, so that, in A', the question requiring more information from the surveyed population precedes the question that requires less. The coverage of the former question increases in A', while the coverage of the latter decreases. Transformation 1 is reminiscent of rule R2 of Picard (1965), which moves questions with higher probabilities closer to the root (i.e., they are assigned a lower rank in a tree). Rule R2 cannot be readily applied here, though, since it does not allow for the root of a questionnaire to be moved. Transformation 1 seems to bring a survey chart structurally closer to a tree (see Proposition 2). Transformation 2 does not decrease the expected number of questions, but does place questions with higher indegree (and probability) closer to the root. It, too, turns A into a survey chart A' structurally closer to a tree.

We note that our transformations do not create questions of higher complexity than that of the two original questions that underwent the transformation. Indeed, transformation 1 essentially inverts the order of the two questions. Transformation 2 creates a new question with essentially the same text as that of the original questions; it is only its coverage that is larger than the coverage of each of the original questions.

Remark 2. In a survey chart, to move a question Q "closer to the root" (i.e. Q becomes the ancestor of one of its former ancestors), one must take two things into consideration. First, one must make sure that all previous information required to ask Q is still available at the new location of Q. Second, the arcs adjacent to and from Q in A must be redirected. The latter requirement can always be achieved, since the descendents of Q can still receive the required information from Q at its new location.

The populations surveyed in A and A' are the same. This means that, if probability measure 1 is used, the probability measures in A and A' remain the same. On the other hand, if, as a result of a transformation, the number of elementary events changes and probability measure 2 is used, the probabilities in A and A' are different.

To evaluate the effect of a transformation on the expected number of questions (4), we have to estimate the difference between two values of (4). With

probability measure 1, it suffices to estimate only the probabilities of four questions (see (8)).

4. Transformation 1

4.1. Description

Throughout this section, A is a survey chart, Q_a is an ancestor of Q_b in A, and (Ω_A, P_A) is the probability space associated with A. Transformation 1 starts with the pair (Q_a, Q_b) in A and redefines it as a pair (Q'_a, Q'_b) in a new survey chart A', with associated probability space $(\Omega_{A'}, P_{A'})$, when (Q'_a, Q'_b) collects the same information as (Q_a, Q_b), and the expected value in (4) decreases, i.e., $E_{A'} < E_A$. The analytical potential of A is preserved, i.e., $\Omega_{A'} = \Omega_A$.

Transformation 1 switches the order of Q_a and Q_b, and should be considered when we can logically extend the coverage of Q_b to the coverage of Q_a, and when Q_b completes, in some sense, the information collected by Q_a. It consists of placing Q_b, the more complete question in A, which is renamed Q'_a in A', in the first position in A', and Q_a,the less complete question in A, renamed Q'_b in A', in the second position in A', while aiming at obtaining $P_{A'}(cover_{A'}(Q'_b)) < P_A(cover_A(Q_b))$. In other words, in A' we try to obtain as much information as possible from the first question, so that more respondents to this first question are spared the burden of having to respond to the second. We create A' from A as follows. The text of the new question Q'_a is based on the text of Q_b, while the information sought by Q'_a may be more detailed than the information sought by both Q_a, Q_b. In A', we place Q'_a where Q_a used to be in A so that $cover_{A'}(Q'_a) = cover_A(Q_a)$. Next, we retain the text of Q_a, which becomes Q'_b and replaces Q_b in A', then send to Q'_b a subpopulation of $cover_{A'}(Q'_a)$ so as to achieve $cover_{A'}(Q'_b) \subset cover_A(Q_b)$. Other than Q_a, Q_b, Q'_a, Q'_b the survey charts A and A' have the same questions Q and $cover_{A'}(Q) = cover_A(Q)$. To achieve these goals, the arcs of A may have to be modified to create the arcs of A'. In probability measure 1, $P_{A'}(Q) = P_A(Q)$ for all questions Q other than Q_a, Q_b, Q'_a, Q'_b (see Section 3). Then the difference between the expected number of questions before and after transformation 1 is:

$$E_A - E_{A'} = P_A(cover_A(Q_a)) + P_A(cover_A(Q_b)) - P_{A'}(cover_{A'}(Q'_a)) - P_{A'}(cover_{A'}(Q'_b)). \tag{8}$$

From the description of transformation 1 (here probability measure 1 is used), $P_{A'}(cover_{A'}(Q'_a)) = P_A(cover_A(Q_a))$. Consequently, (8) becomes,

$$E_A - E_{A'} = P_A(cover_A(Q_b)) - P_A(cover_{A'}(Q'_b)) \tag{9}$$

The idea is to apply transformation 1 when (9) is positive. In what follows we illustrate such instances with examples based on **EM From Survey**.

4.2. Simple order-inversion

We start with a simple example, when changing the order of two questions may decrease (4) when the probability measure 1 is used. This is so because the relative sizes of the populations asked these questions change with the reordering of questions. In the new ordering, we attempt to obtain the most information from the first question and then ask a subset of the coverage of this first question an additional question.

Example 6. Starting with the survey chart $A =$ **EM From Survey**, we generate the survey chart A' by inverting the order of the questions $Q_a = Q_{18}$, $Q_b = Q_{19}$ with $cover_A(Q_a) = \{Q_7 = e.\}$, and $cover_A(Q_b) = \{Q_{18} = no\}$. We recall these questions, where the job referred to is the *MRJ/B*:

Q_{18}: **In this job, were you a union member?** $\{Q_{18} = yes\}$ go to END_A
Q_{19}: **Were you covered by a union contract or collective agreement?** asked if $\{Q_{18} = no\}$

In A', the text of Q'_a is the text of Q_{19}, the text of Q'_b is the text of Q_{18}, and we have $cover_{A'}(Q'_a) = \{Q_7 = e.\}$, $cover_{A'}(Q'_b) = \{Q'_a = yes\} = \{Q_{19} = yes\}$. The other questions and their coverage are the same in A and A', as are the arcs joining them. We perform this transformation when, with probability measure 1, the difference in (9) is positive, which happens when, in the *MRJ/B*, the proportion of employees covered by a union, i.e., $\{Q_{19} = yes\}$ is smaller than the proportion of employees not covered by a union, which account for the larger part of the population defined by $\{Q_{18} = no\}$.

4.3. The general case

As described above, the transformation consists of switching the order of two questions Q_a and Q_b of A, when the conditions described in Section 4.2 apply. In this more general situation, we may combine the information provided by Q_b with information collected from ancestors of Q_b other than Q_a to ensure that $P_{A'}(cover_{A'}(Q_{b'})) < P_A(cover_A(Q_b))$. We consider applying transformation 1 when the information collected from Q_b and its ancestors answers Q_a for a sufficiently large category of respondents so that (9) is positive. We illustrate this more general situation with Examples 7 and 7′.

Example 7. Consider the sub-graph of **EM From Survey** that contains the nodes $Q_a = Q_4(root), Q_5, Q_6, Q_b = Q_7, Q_8, Q_{10B}$. This corresponds to a questionnaire with six questions, represented by the survey chart A in Figure 3. The second graph in Figure 3 is the representation of A as a tree. We recall the questions Q_4 to Q_8:

Q_4: **During the reference period, were you self-employed at any time?**
Q_5: **Did you work at more than one job/business *(J/B)* during the reference period?**

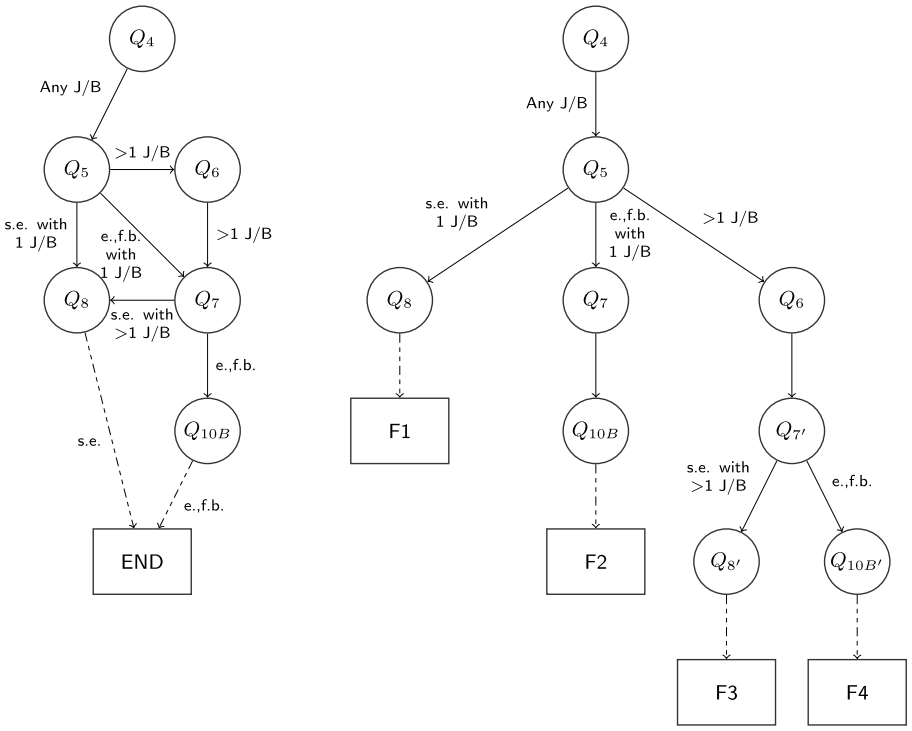


FIG 3. *Example 7*

Q_6: **How many jobs/businesses *(J/B)* did you have during the reference period?** (if $Q_5 = yes$)

Q_7: **In your most recent job/business, were you: an employee, self-employed, working in a family business without pay?** (exclude $\{Q_4 = yes\} \bigcap \{Q_5 = no\}$)

Q_8: **Did you have any employees?** All *s.e.* are asked Q_8, then go to END_A.

Here we can combine the information provided by Q_4 and Q_5, both ancestors of Q_7. For the category of respondents $\{Q_4 = yes\} \bigcap \{Q_5 = no\}$, i.e., those who held only one job during *R.P.* (necessarily the *MRJ/B*), we conclude that they were self-employed in *MRJ/B*, or *s.e.*, so they skip Q_7 in A, which has indirectly been answered. Thus, combining information from Q_4 and Q_5 gives the answer to Q_7 for a subset of $cover_A(Q_5)$. From the tree representation in Figure 3, we see that there are four elementary events and $E_A = 3p(F_1) + 4p(F_2) + 5p(F_3) + 5p(F_4)$.

Example 7′. We apply transformation 1 to the survey chart A in Example 7, so that Q_7, the more informative question (when combined with Q_5), is asked first in the transformed chart A' and becomes Q'_a, while Q_4, which becomes Q'_b, is asked after Q_7, Q_5, Q_6 in A' (see Figure 4). To complete the description of A', we note that we need an arc from Q_6 to Q_8, and one from Q_5 to Q_{10B}, which

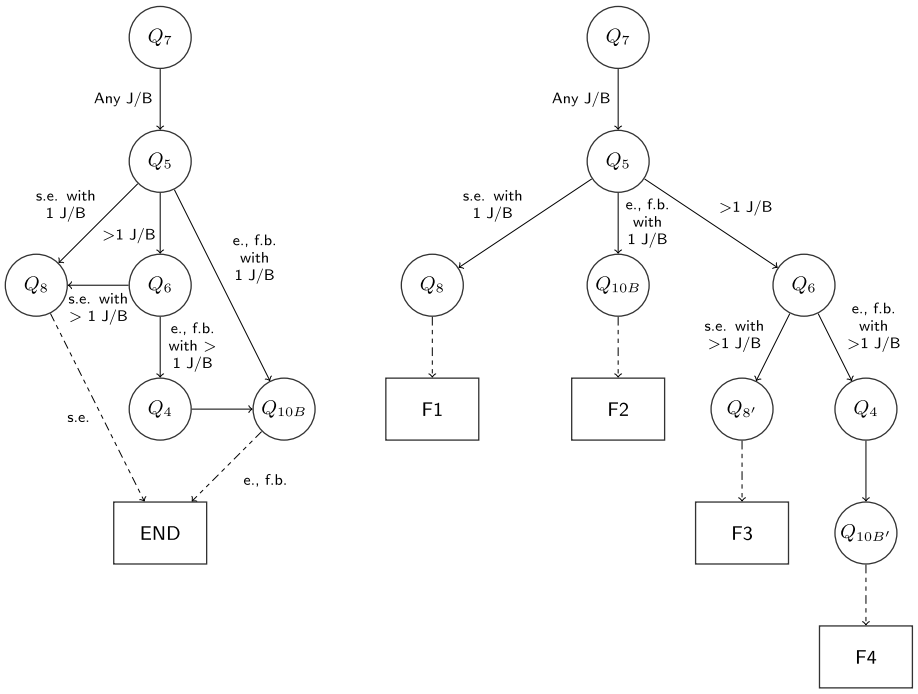


FIG 4. *Example 7′*

did not exist in A. On the other hand, there is no arc from Q_5 to Q_7 or from Q_7 to Q_8 in A'. From Figure 4, we see that there are still four elementary events and that the probability space is still that of Example 7, even with probability measure 2. We have $E_{A'} = 3p(F_1) + 3p(F_2) + 4p(F_3) + 5p(F_4)$.

To show informally that the difference in (9) is positive, we argue that the category of respondents that skips Q'_b in A' is larger than the category of respondents that skips Q_b in A, once Q_5 is asked. Indeed, the respondents that had to skip Q_7 in A were *s.e.*, who held only one *J/B* during *R.P.* (see Example 7). With the new ordering of questions in A', a larger category of respondents must now skip Q_4 in A, once Q_7 and Q_5 have been asked. It consists of *all s.e.* and all other respondents who held only one *J/B* during the *R.P.* (which is necessarily the *MRJ/B*), since we already know from Q_7 their status in *MRJ/B*. A formal argument is presented in Appendix C. Alternatively, using the tree representation in figures 3–4, we have $E_A - E_{A'} = p(F_2) + p(F_3) > 0$ in any probability measure.

Remark 3. In **EM Simplified** (see Appendix D), Q_4 was eliminated. The information it collects says nothing about *MRJ/B*, the main unit of analysis. Furthermore, it complicates the structure of the graph, since an arc has to be created to redirect a subpopulation away from Q_7.

4.4. Transformation 1 and nonresponse

In practice, every question generates a nonresponse category $\{Q = (Dk, Rf)\} = \{Q = Dk\} \bigcup \{Q = Rf\}$. Here Dk stands for "Don't know" and Rf for refusal to answer the question Q. Let $\{Q = Res\} = \{Q = (Dk, Rf)\}^c$ be the category of respondents to Q. In this section we are dealing with a design which takes nonresponse into account, i.e., nonrespondents are part of the coverage of some questions and there are arcs specifically designed for nonrespondents to travel through. Transformation 1 applied to nonresponse can turn the questionnaire A into a questionnaire A', which has, in some sense, a richer analytical potential than A and a lower value of (4).

Description of transformation 1 for nonresponse. With transformation 1 in mind (sections 4.1 and 4.3), we proceed as follows. We assume that we can combine the contents of Q_a and Q_b, which results in the text of Q'_a, and place Q'_a in A' where Q_a used to be in A. The information solicited by the more complete question Q_b in A is now solicited by Q'_a in A', and we set $cover_{A'}(Q'_a) = cover_A(Q_a)$. The question Q'_b has the content of Q_a, it is placed in A' where Q_b used to be in A, and we set $cover_{A'}(Q'_b) = \{Q'_a = (Dk, Rf)\}$. All other nodes of A are nodes of A', and A' has no other nodes. The coverage of all questions, other than Q_a and Q_b, is the same in A and A', while the arcs of A may have to be adjusted in A'.

Consider a questionnaire associated with the survey chart A, and assume that the probability measure 1 is used. Then (9) adapted to this situation gives:

$$E_A - E_{A'} = P_A(cover_A(Q_b)) - P_A(\{Q'_a = (Dk, Rf)\}) \tag{10}$$

An examination of the right hand side of (10) shows that, with a good follow up strategy, the negative term can be negligible. In particular, if there is no nonresponse, the negative term is zero and (10) is strictly positive, i.e., transformation 1 in this case reduces (4).

Placing the more detailed question before a more general one is not uncommon in survey practice. The entry module of many surveys requests first the individual's date of birth, and then the age, but only of individuals who refused to give their date of birth. We illustrate this procedure with examples based on **EM from Survey**.

Example 8. We define the survey chart A, which represents a questionnaire with two questions. The nodes of A are the root $R = Q_{20}$ and a final question Q_{21} , both taken from **EM**. We assume that the coverage of R consists of adults who held at least one job during the *R.P.* This population could be obtained, for instance, from a labour force survey, which collected information on all jobs held during our *R.P.* We recall Q_{20} and Q_{21}, which refer to the *MRJ/B*:

Q_{20}: **Was this job permanent, or is there some way in which it was not permanent? For example, seasonal, temporary, term or casual?**

Q_{21}: **In what way was this job not permanent?** The possible categories: seasonal job, temporary, term or contract job (non seasonal), casual job, other, are not read to respondents. (asked of $\{Q_{20} = not\ permanent\}$)

We define the arcs as follows. Respondents in $\{Q_{20} = \textit{not permanent}\}$ are first asked Q_{21}, then sent to END_A, while individuals in the complementary category $\{Q_{20} = \textit{permanent}\} \bigcup \{Q_{20} = (Dk, Rf)\}$ are sent directly to END_A. The analytical potential of A is $\Omega_A = A_1 \bigcup A_2 \bigcup A_3 \bigcup A_4$, where: $A_1 = \{Q_{20} = \textit{permanent}\}$, $A_2 = \{Q_{20} = \textit{not permanent}\} \bigcap \{Q_{21} = Res\}$, $A_3 = \{Q_{20} = \textit{not permanent}\} \bigcap \{Q_{21} = (Dk, Rf)\}$, $A_4 = \{Q_{20} = (Dk, Rf)\}$. Here A_1, A_2 contain the full responses, A_3 the partial responses, and A_4 the total nonresponse. The survey chart A produces two elementary events, $F_1 = A_1 \bigcup A_4$, and $F_2 = A_2 \bigcup A_3$. While A_1 is itself an analytical outcome, A_2, A_3, A_4 are unions of analytical outcomes, e.g., $A_4 = \{Q_{20} = Dk\} \bigcup \{Q_{20} = Rf\}$.

Example 8′. In this example we apply Transformation 1 to the survey chart A in Example 8. In the transformed survey chart A', all the necessary information is collected in one question Q'_{20}, which is slightly more complex than Q_{21}. Respondents to Q'_{20} are sent to END_A and nonrespondents to Q'_{20} are asked Q'_{21}:

Q'_{20}: **Was this job permanent, or was it seasonal, temporary, term or contract (non seasonal), casual job, other?**
Q'_{21}: **Was this job permanent, or is there some way in which it was not permanent?** (if $Q'_{20} = (Dk, Rf)$).

Remark 4. The questions in Examples 8 and 8′ refer to the same concepts. In Example 8, Question Q_{20} has 22 words, while Q_{21} has 8. At 17 words, question Q'_{20} is less complex than Q_{20}, while Q'_{21} , with 15 words, is more complex than Q_{21}. The new questions have less than 20 words, which is the acceptable upper bound for the complexity measure set by Payne (1949).

In view of the coverage of Q'_{21} and the information requested by Q'_{20}, Example 8′ could be modified by retaining Q'_{20} and replacing Q'_{21} with the simpler question:

Q''_{21}: **Was this job permanent?**

The set of analytical outcomes $\Omega_{A'}$ in Example 8′ is described in Appendix C, where we also compare Ω_A and $\Omega_{A'}$. It appears that $\Omega_{A'}$ gives more information on nonresponse patterns than Ω_A.

Essentially, two analytical outcomes of A', namely $\{Q'_{20} = Dk\} \bigcap \{Q'_{21} = \textit{permanent}\}$, and $\{Q'_{20} = Rf\} \bigcap \{Q'_{21} = \textit{permanent}\}$ cannot be generated by A. An individual classified to $\{Q'_{20} = Dk\} \bigcap \{Q'_{21} = \textit{permanent}\}$ is a "reluctant respondent", who had to be asked twice to give their complete information. This opportunity is not provided by the questionnaire represented by the survey chart A. Perhaps some of these "reluctant respondents" have a lower propensity to respond than the more eager respondents classified in A_1 in Example 8. Should this be the case, the information from A', in conjunction with additional information on these respondents (e.g., information on their income) is valuable in studying nonresponse patterns. Alternatively, some "reluctant respondents" could have had problems responding to the complex formulation of question

Q'_{20} (or Q_{20}), in which case A' gives valuable information for designing questionnaires. Either way, A' provides more information that A on this subject. We reach the following conclusion:

Remark 5. The survey chart A' in Example 8′ may provide more information on nonresponse patterns then survey chart A in Example 8. Furthermore, with a good follow up strategy, $E_{A'} < E_A$,when the probability measure 1 in Section 2.2 is used. On the other hand, if probability measure 2 is used, $E_{A'} = E_A$

Proof. The first statement follows from the discussion above and the details in Appendix C. To prove the second statement, we note that (10) is, in this case:

$$E_A - E_{A'} = P_A(\{Q_{20} = not\ permanent\}) - P_A(\{Q'_{20} = (Dk, Rf)\})$$

Since the proportion of employees with non permanent jobs is sizeable and we can reduce the negative term with a good nonresponse strategy, we conclude that A' creates less response burden than A. To prove the last statement of Remark 5, we note that the charts A, A' are identical, so, using probability measure 2, $E_{A'} = E_A$. □

Remark 6. Example 8′ illustrates the fact that probability measures 1 and 2 can give different results when transformation 1 is applied with nonresponse. The inadequacy of probability measure 2 to measure the true response burden on the surveyed individuals should come as no surprise. Probability measure 2 on Q completely disregards the number of individuals that have to answer question Q.

When a question is followed by a more detailed question on the same topic, it is tempting to combine the content of both questions into one, more complete question, which, in the transformed graph, would replace the two original questions. This, however, reduces the analytical potential in the transformed questionnaire, as illustrated by the following example.

Example 8″. We form A'' from A' in Example 8′, by retaining Q'_{20} and discarding Q'_{21}. In this case, $\Omega_{A''} \subset \Omega_A$ and $E_{A''} < E_{A'}$ regardless of the probability measure used. Details are provided in Appendix C.

Remark 7. In **EM Simplified** (see Appendix D), we retained Q'_{20} and discarded Q'_{21}, both defined in Example 8′. The cumulative effect of eliminating relatively redundant questions may lead to a substantial reduction in response burden. On the other hand, we gain no insight into the response mechanism with this reduced questionnaire.

In Example 9, we refer to Q_5, Q_6, which were stated in Example 7. The coverage of Q_5 consists of all surveyed individuals who held at least one J/B during R/P, $cover_A(Q_6) = \{Q_5 = yes\}$, and Q_6 solicits more information than Q_5.

Example 9. We start with a survey chart A consisting of two questions: $R = Q_a = Q_5$, and $Q_b = Q_6$. Respondents to Q_a are sent to Q_b and non-

respondents to END_A. Applying transformation 1 for nonresponse, we produce A', in which $Q'_a = Q'_5$ has the text of Q_6, $Q'_b = Q'_6$ has the text of Q_5, $cover_{A'}(Q'_5) = cover_A(Q_5)$, and $cover_{A'}(Q'_6) = \{Q'_5 = (Dk, Rf)\}$. To achieve this, some adjustments have to be made to the arcs of A.

Remark 8. Question Q_5 is a screening question: only respondents who answer "yes" to it are asked Q_6. The impact such screening has on motivating respondents to answer further questions is unclear. On the one hand, it gently leads the respondents in $\{Q_5 = \text{yes}\}$ to Q_6. On the other hand, respondents may learn that answering "yes" often leads to follow-up questions and decide to answer "no" instead, which could eventually bias the results (Section 9.9.1, Handbook of Survey Research 2010, p. 291). Eliminating Q_5 will reduce response burden, but some information on nonresponse patterns will be lost. On the other hand, Q_2 is an important screening question that should be retained.

Remark 9. In **EM Simplified**, Q_5 was eliminated. We felt that simplifying the graph and eliminating a superfluous question was worth losing some information on nonresponse patterns. Furthermore, since we are dealing with people who held at least one job during the reference period, asking them directly how many jobs they held during the same period is neither offensive nor illogical.

5. Transformation 2

5.1. Description

We start with a survey chart A, which contains two similar questions Q_a and Q_b, asked of two disjoint populations a and b. Transformation 2 creates a new question $Q_{a \bigcup b}$, which represents these similar questions and has $cover(Q_{a \bigcup b}) = c = a \bigcup b$, then places $Q_{a \bigcup b}$ in a transformed survey chart A', so that its indegree is reduced. The coverage of the $Q_{a \bigcup b}$ is generally easy to verify in A'. While A' has the same expected number of questions as A, it is easier to use for statistical analysis, as it has a simpler structure. The first phase of transformation 2 consists of creating the common question $Q_{a \bigcup b}$. In general, it is a good idea to place $Q_{a \bigcup b} = Q'_a = Q'_b$ in the transformed survey chart close to one of the original questions, to preserve the logical flow of the questionnaire. After the completion of the first phase of transformation 2, the indegree of $Q_{a \bigcup b}$ is larger than 1. The second phase of transformation 2 attempts to lower the indegree of $Q_{a \bigcup b}$ in a transformed survey chart A'. This is achieved by moving $Q_{a \bigcup b}$ closer to the root in A'. The second phase can actually be applied as a stand-alone transformation to any question Q_c with indegree larger than 1, e.g., to $Q_c = Q_{25}$ of **EM From Survey** (see also Example 14 below). Technically, the coverage of any question Q_c with indegree larger than 1 is of the form $c = a \bigcup b$, where a and b are disjoint populations, so we can always assume that two identical questions asked of disjoint populations have been joined to create Q_c.

5.2. Examples

We now illustrate the first and the second phase of transformation 2 with examples.

Example 10 below illustrates the first phase of transformation 2. We replace questions Q_{10A} and Q_{10B} in $A =$ **EM From Survey** (see Appendix A) by a common question Q'_{10} in A'. The two original questions have a similar content, requiring the name of the most recent business (Q_{10A} is asked of $a = \{s.e.\}$), or the name of the most recent employer (Q_{10B} is asked of $b = \{e.\} \bigcup \{f.b.\}$). This information is needed to appropriately address all questions from Q_{11} to END_A. We recall that the categories *e.*, *f.b.*, and *s.e.*, represent the status of every respondent in their *MRJ/B*. This preliminary information comes from two sources in A. The first source is Q_7 for *e.*, *f.b.*, and the *s.e.* who held more than one *J/B* during the reference period. The second source identifies the *s.e.* who held only one *J/B* during the reference period by combining information provided by Q_4 and Q_5. Due to the structure of this survey chart, we can simply merge Q_{10A} and Q_{10B} to create a common question Q'_{10}, which reads:

Q'_{10}: **What was the name of your most recent employer/business?**

Example 10. We start with $A =$ **EM From Survey** and construct a survey chart A' as follows. We "merge" Q_{10A} and Q_{10B} to form a node Q'_{10} in A' and set $cover_{A'}(Q'_{10}) = \{e.\} \bigcup \{f.b.\} \bigcup \{s.e.\}$. All other nodes and arcs in A' are identical to those of A.

We discuss now a more complex example of the first phase of transformation 2. In $A =$ **EM From Survey**, we notice that Q_9(coverage $a = \{s.e.\}$), and Q_{22}(coverage $b = \{e.\}$), both ask the number of employees at the *MRJ/B*, but are situated at different locations in A. In this survey chart, we cannot simply "merge" Q_9 and Q_{22}, because a cycle is formed in the new graph, which generates an endless loop when the questionnaire is programmed. Fusing Q_9 and Q_{22} would also create a question with a high indegree. As in Example 10, the preliminary information is the classification information, which comes in A from Q_7 or from Q_4 combined with Q_5. In principle, the required information for a similar question should be identical for the populations *e.* and *s.e.* We notice, however, that Q_8 collects additional information for Q_9 from the population *s.e.*, so we should not place $Q_{a \bigcup b}$ before Q_8 in A'. This lack of symmetry complicates the structure of the survey chart.

In **EM Analytical** (see Appendix D), we have decided against merging Q_9 and Q_{22}. As is now, in **EM Analytical** the flows carrying *s.e.*, *e.*, and *f.b.* are clearly separated, so one can easily study the information collected from each of these analytically important categories of respondents. Example 11 below illustrates the disadvantages of merging questions Q_9 and Q_{22}.

Example 11. Starting with $A =$ **EM Analytical**, we merge Q_9 and Q_{22} to create a common question Q'_{22}, which we place right before Q_{18}, after the population *f.b.* has left the survey and create a new survey chart $A' =$ **EM Analytical Merged** (see Appendix D). The common question Q'_{22} reads:

Q'_{22}: **About how many persons, other than yourself, were employed at the location of your MRJ/B?**

The survey chart **EM Analytical Merged** is a nonplanar graph (see section 1.3). Such survey charts generally present problems for programmers, which call the logic behind such charts "spaghetti", or "unstructured". Furthermore, A' contains empty paths, and so the populations *s.e.* and *e.* must belong to two different categories of flows. An example of an empty path is $Q_8Q'_{22}Q_{18}$.

Remark 10. In **EM Simplified** (see Appendix D), we were able to merge Q_9 and Q_{22}, since Q_8 and Q_4 had been removed.

Example 7′ can be viewed as a successful, albeit subtle form of transformation 2, in which only part of the coverage of Q_4 is joined to the coverage of Q_7. Here we do not discard question Q_4, because it seeks additional information not related to Q_7.

Example 12. Here A is defined in Example 7, A' in Example 7′. We have created A' first by defining a "common" question, with the text of Q_7 and $cover_{A'}(Q_7) = cover_A(Q_7) \bigcup (\{Q_4 = yes\} \bigcap \{Q_5 = no\})$

Thus, part of the coverage of Q_4 has been transferred to the coverage of the "common" question Q_7. Question Q_4 is retained in A' with a reduced coverage, because of the additional information required from respondents who held more jobs in *R.P.* and were not self employed in *MRJ/B*. The second phase of transformation 2 is completed by placing Q_7 before $Q_5Q_6Q_4$ in A', so that its indegree becomes 1.

We now give an example of the second phase of transformation 2.

Example 13. In Example 10, note that the indegree of Q'_{10} in A' is three. An "informational" ancestor of Q'_{10}, namely $Q'_a = Q_7$, is created in Example 7′. Thus, in **EM Analytical** we can place Q'_{10} right after Q_7, which reduces its indegree to 1 in A' and completes the second phase of transformation 2.

In Example 14 below, the second phase of transformation 2 is applied directly to a question with high indegree, Q_{25}. Starting with $A =$ **EM From Survey**, we move Q_{25} closer to the root, thus reducing its indegree in the new survey chart A', while preserving its coverage, which consists of all respondents who worked at a *J/B* during the reference period. One can ask Q_{25} as soon as R7 has been read, so we can actually place Q_{25} in A' somewhere between Q_{11} and Q_{18}.

Example 14. We create a new survey chart A' from $A =$ **EM From Survey** by moving Q_{25} from its position in A, right after Q_{14} in A', and then sending all its incoming arcs in A to $END_{A'}$ in A', as in **EM Analytical**. The content of Q_{25} fits logically around Q_{14}, which has the same coverage. The position and coverage of every other question are the same in A and A'. While the coverage of Q_{25} remains the same in A', its indegree in A' is now 1.

5.3. General comments on transformations 2

Transformation 2 combines questions with a similar content and places questions with larger coverage closer to the root, thus turning the initial survey chart A into a survey chart A' that is closer to a tree (see Proposition 2). Applications of transformation 2 make it easier for the analyst to see "who is asked what" in the questionnaire. For instance, in Example 12 (or Example 7′), Q_7 becomes a "roster" question in A', which is asked quite early in the game. It is much easier for the analyst to see the distribution of the type of workers (*e,. s.e., f.b.*) in *MRJ/B* using Q_7 in A' than the combination of Q_7 and (Q_4,Q_5) in A. Furthermore, placing Q_7 closer to the root is in line with the good questionnaire design practice of asking "roster" questions as early as possible. It also conforms with rule R4 of Picard (1965), which requires that nodes with high outdegree be placed closer to the root in an optimal questionnaire (the nodes from Q_7 to Q_5 in **EM Analytical** formally constitute a single node in Picard (1965)).

In Example 14 it is easier to verify the coverage of Q_{25} in A' than in A. The fact that the indegree of the terminal node increases in A' is beneficial. It makes it easier to distinguish the paths taken by various category of respondents, namely *s.e.*, *f.b.* and *e.*, as seen in **EM Analytical**. It is also easier for programmers to verify that all subpopulations that make up the population of the survey have reached $END_{A'}$.

In Example 14 it was easy to "move up" Q_{25}, because little preliminary information was needed to ask it. More generally, let us start with $Q_{a \bigcup b}$, where the disjoint populations *a,b*, travel along the inflows f_a, f_b, which carry information needed to ask $Q_{a \bigcup b}$. In principle, the required information to ask $Q_{a \bigcup b}$ should be very similar for populations a and b. This is not always the case, as seen in Example 11. Nonetheless, some common preliminary information may be needed and is available from similar questions situated along f_a and f_b. We then proceed in creating a common informational ancestor, (see Example 13). If more informational ancestors are required, we create them in the order in which the information is needed for asking $Q_{a \bigcup b}$. Although transformation 2 does not, in general, reduce the expected number of questions, it plays an important role in improving the structure of the questionnaire.

6. EM analytical and EM simplified

Based on a simplified version of the questionnaire of **EM**, we created the survey chart **EM From Survey with conditions** (Appendix A). The resulting survey chart, our benchmark **EM From Survey**, is the starting point for applying transformations 1 and 2. There are five graphs associated with **EM From Survey** in Appendix A. The first and second illustrate the count-down and count-up of flows, needed to calculate the expected number of questions with probability measure 2, which is 14.889. **EM From Survey** has 45 flows and two categories: the first category consists of flows travelled by the populations *e.* or *f.b.* or no *J/B*, colour-coded blue on the third of the five graphs. The second

category has flows travelled by *s.e.*, and it is colour-coded yellow on the fourth of these five graphs. The fifth graph contains three colours, which resulted from putting the two categories together. We note that the colours also represent the coverage of questions. Questions with coverage containing both categories are colour-coded green. We can thus see that Q_{25} is misplaced and that the question with maximum coverage could be better organized. This has been accomplished in **EM Analytical** (Appendix D), the survey chart that we recommend.

EM Analytical has the same analytical potential as **EM From Survey** and it is simpler to deal with. It has 45 flows and the expected number of questions in probability measure 2 is $14.40 < 14.889$. It has only one category, which makes the calculation of probabilities easier. To obtain $A' =$ **EM Analytical** from $A =$ **EM From Survey**, we apply successively the transformations described in sections 4 and 5.

Example 15. Starting with $A =$ **EM From Survey**, we apply transformation 1 in Example 7′ to extend the coverage of Q_7 and reposition the question. Next, we apply transformation 2 to create Q'_{10} and then reposition it right after Q_7 in A'. Questions Q_{11} to Q_{14}, which have a large coverage, are placed right after Q'_{10}. We then reposition Q_{25} between Q_{14} and Q_5 and Q_{22} before Q_{18}. The group of questions with smaller coverage are placed after Q_5 and the structure of the subgraph that follows is basically inherited from A. This completes the description of $A' =$ **EM Analytical** (see Appendix D).

EM Simplified (see Appendix D) has only one category with 7 flows, and the expected number of questions in probability measure 2 is 10.857. To obtain **EM Simplified**, we eliminated some questions. In practice, this requires negotiations and cannot be done without the approval of all stake holders, as the analytical potential of the resulting questionnaire is reduced. Our justification for eliminating some questions is given in Remarks 3, 7, 9. We also eliminated Q_2; while the main activity during the reference period could have been the object of analysis in **EM**, it is not, and the information collected by this question has no connection with any other question in the questionnaire.

Example 16. The survey chart **EM Simplified** (see Appendix D) can be obtained from **EM Analytical** first by eliminating questions Q_2, Q_4, Q_5, Q_8. Next, Q_{20} and Q_{21} are replaced by Q'_{20}. Finally, we merge Q_9 and Q_{22} to create Q'_{22} and place it after Q_6 (see Remark 10).

While identifying questions for deletion was based on judgment alone, one can develop a more structured approach to this process, which can also be used to streamline the questionnaire. To illustrate this, we consider the text of **EM** coupled with the survey chart **EM From Survey** (both in Appendix A). We start with the root, and then proceed to the next question in the pre-established order of **EM**. For each question, we determine first if it is analytically useful for the purpose of the survey. If "no", we consider removing it, but continue nonetheless with its analysis. We check next if it leads to a significant split, e.g., one that sends a subpopulation to the end, in which case we definitely keep the question (e.g., Q_3 should be kept). Next, we check, using key words, if part of the

information collected by this question is related to information solicited by other questions in the questionnaire. If "yes", care must be exercised in untangling similar concepts and defining the coverage of the questions involved so that no individual is asked the same thing more than once. Sometimes even asking very similar questions of the same individual can generate nonresponse. To exemplify, we start with the root Q_2, which, at a first glance, appears to collect useful information about employment. The key words of note are: *reference period* (abbreviated *R.P.*), *main activity*, *job or business* (abbreviated *J/B*), *family*. Setting *R.P.* aside, a search through the questionnaire finds these key words in Q_3, R_7, Q_7, Q_{10A}, Q_{11}, Q_{13}. A comparison of Q_2 and Q_3 reveals that holding a *J/B* in Q_2 refers to the *J/B* being the main activity during *R.P.*, whereas in Q_3 it refers to holding a *J/B* at any time during *R.P.*, two different concepts that may coincide for some individuals, thus the split after Q_2. In R_7, it becomes clear that the concept of interest is *the most recent J/B* (*MRJ/B*), and Q_2 is more in line with it, as *J/B* at any time in *R.P.* includes *MRJ/B*. All questions from R_7 on refer to *MRJ/B*, which is the unit of analysis, as revealed in the title of the module. We can therefore remove Q_2, and the coverage of Q_3 can be extended to all surveyed individuals, which is what we did in **EM Simplified**. A similar analysis concludes that the concepts of being self employed at any time during *R.P.*, introduced in Q_4, and being self employed during *MRJ/B*, (abbreviated *s.e.*), are similar, but not identical. Thus, asking Q_4 might be more trouble than it is worth. For this reason, we eliminated Q_4 in **EM Simplified**. We could proceed in a similar manner until all questions in **EM** will have been analyzed.

7. Comparison with other methods

In this paper we have discussed a specific problem, often encountered in practice. We start with a questionnaire, with predetermined questions and analytical potential. If need be, we perform transformations on the survey chart associated with this questionnaire, to decrease the expected number of questions (in probability measure 2, which is always available), and make it more amenable to analysis. Attaining the last objective essentially means that the transformed survey chart is closer to a tree than the original survey chart (see Proposition 2). It is apparent from the description of our transformations (sections 3–5) that the text of our transformed questions differs minimally from the text of the original questions.

We now look at the work of our best contenders, i.e., Picard/Parkhomenko (henceforth called Parkhomenko, 2010) and decision theory, to see how our problem could be solved with the tools they provide. In both of these approaches, questionnaires are viewed as trees, where every non terminal node (question) is a parent. Therefore, analytical outcomes appear as children of the "last" questions. We illustrate these two approaches with simple examples.

7.1. Parkhomenko (2010)

Example 17. Consider the sub-graph A induced by the survey chart in Figure 2, where only the nodes Q_4, Q_5, and Q_7 are retained. The arcs are: Q_4Q_5, Q_5END_A (s.e. w. 1 J/B) , Q_5Q_7, and Q_7END_A (all but s.e. w. 1 J/B). There are 8 analytical outcomes associated with the corresponding questionnaire q, all appearing as end nodes in the first graph of Appendix E. This first graph represents q as a tree (see also figures 3 and 4) in the set-up of Parkhomenko (2010).

Parkhomenko's algorithm A1 (p. 1128 of Parkhomenko, 2010) with our definitions and probability measure 2 is illustrated on Example 17 and in Appendix E. It uses two lists: At the onset (Step 1), List 1 has the genuine analytical outcomes placed in the left column, in increasing order of their probabilities (in the right column), and randomly within groups of equal probability. Here, all analytical outcomes have probability 1/8 of occurrence. At the onset, List 2 has in the left column all question nodes in increasing order of their outdegree (in the right column). Recall that the survey chart has been turned into a tree, so there are two copies of Q_5 and three copies of Q_7.Within the same recorded outdegree, the ordering of questions on List 2 is random (Q_4 is first on List 2). The first two analytical outcomes on List 1 are assigned to Q_4, then deleted from List 1, and Q_4 moves from List 2 to the bottom of List 1 of "analytical outcomes", and is assigned probability 2/8 in Step 2. A new tree starts being built, with Q_4 as root and the two, now deleted analytical outcomes as its children. The procedure continues until the two lists are empty and a final tree is built (Step 7). This final tree has Q_7'' as root and the other copies of Q_7 as two of its three children. Clearly, this cannot represent a questionnaire, let alone a questionnaire with prescribed questions. It can be shown (see Appendix E) that, through some "neutral" transformations, we can obtain a bona fide questionnaire q' with Q_7 as root and questions that do correspond to the original questions. The change from q in Example 17 to q' in Appendix E is very similar to our transformation 1 on Example 7, but far more complicated.

Clearly, some of our transformations do not have a counterpart in the work of Parkhomenko (2010), e.g., the first phase of transformation 2, which puts together two questions with similar content and disjoint coverage.

7.2. Decision trees

The examples below illustrate that reducing the number of questions using decision trees may create complex questions, which do not match the original questions set by analysts.

Example 18. Consider the population of employees whose most recent job during R.P. was not a family business. The questionnaire consists of the following questions, adapted from **EM**.

Q_{18}: **In this job, were you a union member?** $\{Q_{18} = \text{yes}\}$ go to Q_{20}, else continue.

Q_{19}: **Were you covered by a union contract or collective agreement?**
Q_{20}: **Was this job permanent, or is there some way in which it was not permanent? For example, seasonal, temporary, term or casual?** $\{Q_{20} =$ permanent $\}$ go to END, else continue.
Q_{21}: **In what way was this job not permanent?** The possible categories: seasonal job, temporary, term or casual job, other, are not read to respondents. After Q_{21}, all individuals go to END.

There are 15 analytical outcomes associated with this questionnaire. One such outcome is: $\{Q_{18} = yes\} \cap \{Q_{20} =$not permanent$\} \cap \{Q_{21} =$seasonal$\}$. In the next example, adapted from Fenn (2015), section 2.3, we start with this analytical potential, and build up a questionnaire using as goodness measure the minimum number of questions. We constrain the number of questions to 2. Reducing the questionnaire to one question will make it difficult for the surveyed individuals to give accurate answers.

Example 18′. Consider the population surveyed in Example 18 and the corresponding analytical potential. A questionnaire with two questions that attains the analytical potential of Example 17 is:

Q_1: **In this job, were you either a union member, not a union member in a unionized job, or was this job not covered by a union contract or collective agreement?**
Q_2: **Was this job permanent, seasonal, temporary, term or casual, other?**

The first question is a complex question and is not close to any of the questions in Example 18. There are other differences between our approach and the approaches presented above. Due to our requirement of preserving, inasmuch as possible, the initial questions, some survey charts could never be turned into trees, which is the only type of questionnaires that Picard considers (e.g., the triangle $Q_{18}Q_{19}Q_{20}$ in **EM Analytical** must stay on). A survey chart, even an optimal one, could still have empty paths. As mentioned before, this disadvantage is offset by the fact that survey charts are a visually effective and succinct mode of storing the information potentially contained in a questionnaire.

8. Conclusion and future work

In this paper we introduced a new type of graphs, the survey charts, which we used as aids in designing questionnaires. A comparison of the graphs **EM From Survey** (or **EM From Survey with condition nodes**, both in Appendix A), and **EM Analytical** (Appendix D), shows that the questionnaire based on **EM Analytical** is a lot more amenable to analysis. The analyst interested in specific questions (a form of cross-sectional analysis), would find Q_7 in **EM Analytical** much easier to deal with than the combination of Q_4, Q_5 and Q_7 in **EM From Survey**. It is also apparent that **EM Analytical** is the more useful tool in verifying the coverage of questions and that the string of questions addressed

to each of the three important categories of workers (e., s.e., f.b.) can be easily read-off the flows of **EM Analytical**.

To arrive at **EM Analytical**, we defined and performed a series of transformations which brought the structure of **EM From Survey** closer to the structure we prefer: a road-map to a questionnaire that is easy to analyze and has a minimum expected number of questions. Complex questionnaires other than **EM** could help uncover other useful transformations. Perhaps more fruitfully, we could establish a closer connection between our work and that of Picard (1965) and Parkhomenko (2010). Their ideas could guide us through defining the concept of an optimal survey chart. A comparison of the concept of information used in Picard (1965) and our informal concept of information, which is related to the number of analytical outcomes, could be illuminating. To uncover the logical connection between questions, we could establish a way of sequencing questions and designing the arcs of the survey chart in a way that reflects the order in which the information is gathered. This could help us deal with another difference between our survey charts and Picard's questionnaires. While each question creates splits in Picard's questionnaires or in decision trees, our survey charts allow for unbroken stretches of questions (e.g, Q_7–Q_5 in **EM Analytical**), some of which do not contribute information required to form new branches in the survey chart (e.g., the five questions Q_{11}–Q_{25} in **EM Analytical**). Such questions could be theoretically lumped together, and, using the more general approach of Parkhomenko (2010), their contribution to the expected number of questions could be given a cost equal to their number. Introducing a cost in formula (7) would also allow us to capture the complexity of each question. Alternatively, we could use decision trees in designing a questionnaire with a given analytical potential, by selecting an appropriate goodness measure and placing a cap on the complexity of questions, as in Example 18′. When retaining predetermined questions is required, we could experiment with various goodness measures, and incorporate a distance between the text of two questions in the function we want to minimize. Test data could be obtained from previous, similar surveys.

An alternative way of streamlining a survey chart is to recursively analyse the questions, as outlined in Section 6. This approach also helps in deciding which questions could be modified or simply eliminated, when the requirement for preserving the analytical content of the questionnaire can be relaxed.

In conclusion, this paper proposes a structural approach to designing questionnaires, which opens many theoretical possibilities for further development. We also hope that practitioners will benefit from the examples that we provide here in order to improve on questionnaires at early stages in their development.

Appendix A: A simplified version of the questionnaire of the module "Most recent employment" of ASETS (EM)

This is the questionnaire of **EM**, our simplified version of the questionnaire of the ASETS module "Most recent employment". Bold letters indicate that the

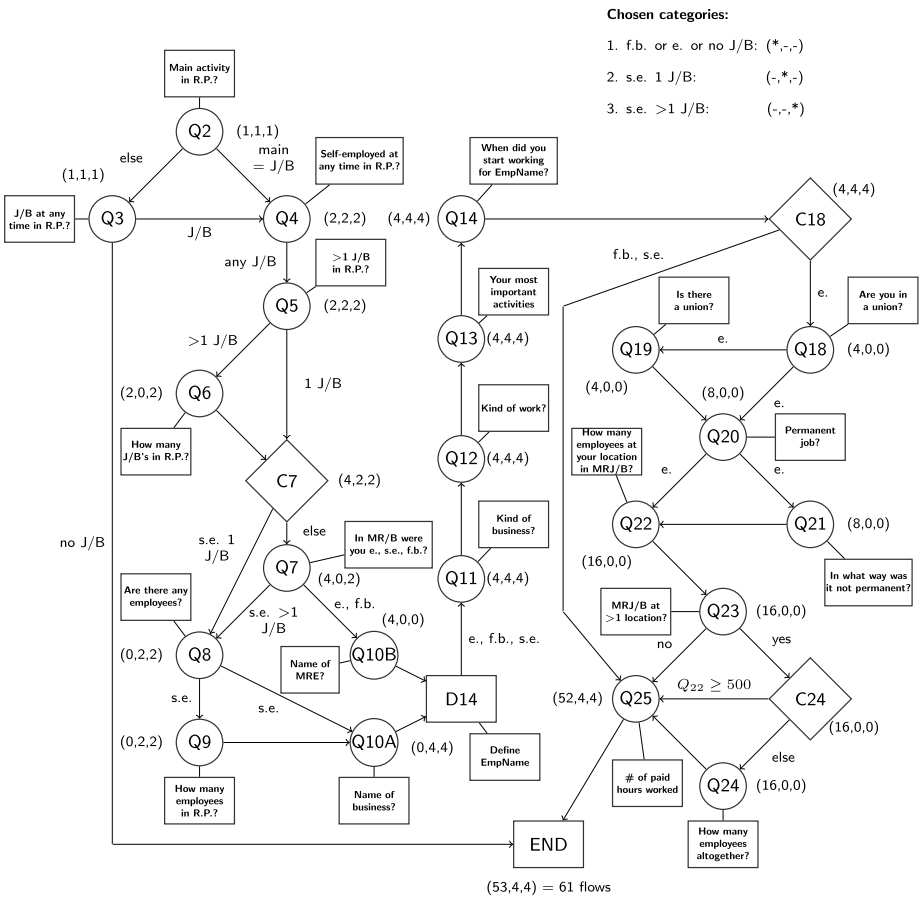


FIG 5. ***EM From Survey with condition nodes***. *Three categories of flows. Flow count-down.*

text is read to respondents. We use the acronyms: *Reference period = R.P.*, the time between July 1, 2007 and June 30, 2008; *Job or business = J/B*; *Most recent job or business = MRJ/B*; *employee = e.*, employed in MRJ/B; *self employed = s.e.*, self employed in *MRJ/B*; *family business = f.b.*, when *MRJ/B* is a family business. *Most recent employer* (only for *e.* and *f.b.*) = *MRE*; *Employer name = EmpName* (covers *e.*, *s.e.*, *f.b.*). When no condition is specified after a question, it is understood that the population moves to the next question or condition.

Q2: **During the *R.P.*, what was your main activity?** (categories are read to respondents): **working at a job or business (*J/B*), doing volunteer work, looking for work, going to school, taking care of family or household responsibilities, retired, long term illness or disability, other**... {Q2=*J/B*} go to Q4, all others to Q3.

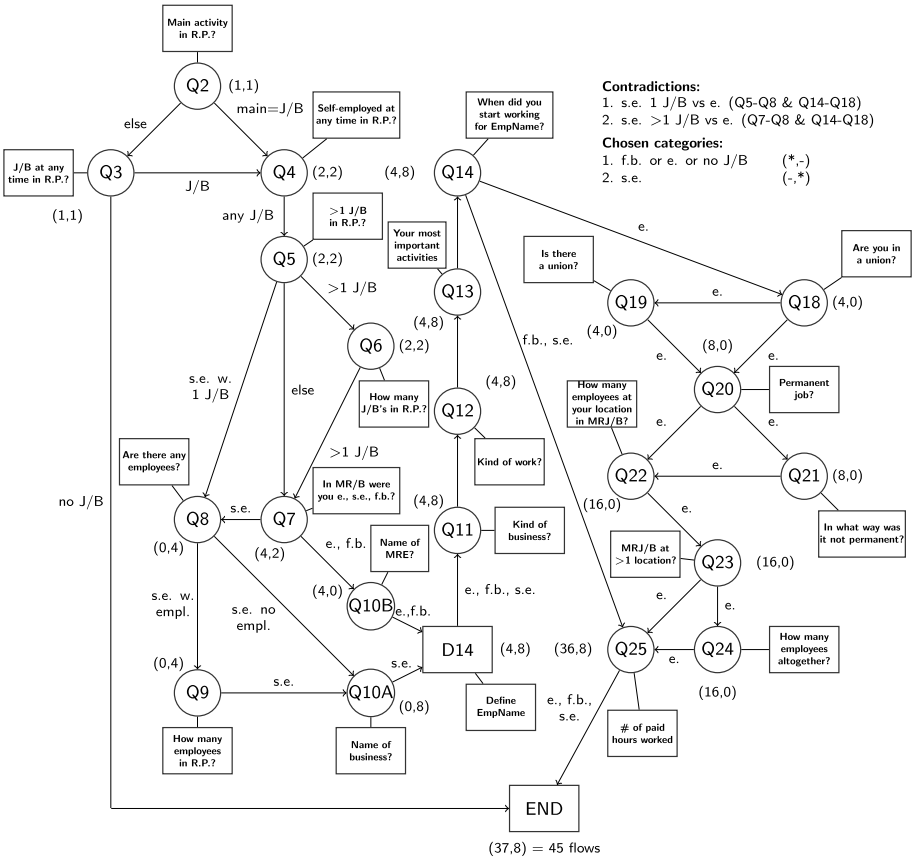


FIG 6. ***EM From Survey***. *Two categories of flows. Flow count-down.*

Q3: **Did you work at *J/B* at any time during the *R.P.*?** {Q3=*yes*} continue to Q4, {Q3 = *no*} leave the survey.

Q4: **During the *R.P.*, were you self employed at any time?**

Q5: **Did you work at more than one *J/B* during the *R.P.*?** {Q5 = *yes*} go to Q6, {Q5 = *no*} skip to condition C7.

Q6: **How many *J/B*s did you have during the *R.P.*?**

C7: This condition is used for programming. As a node, it gathers all flows and redistributes the populations as follows: {Q4 = *yes*} and {Q5 = *no*}, i.e., self employed who held only one job in *R.P.*, go to Q8, everybody else goes to Q7.

R7: **The next questions are about the most recent job/business (*MRJ/B*) worked at during the *R.P.*** This text prepares respondents for the remainder of the questionnaire.

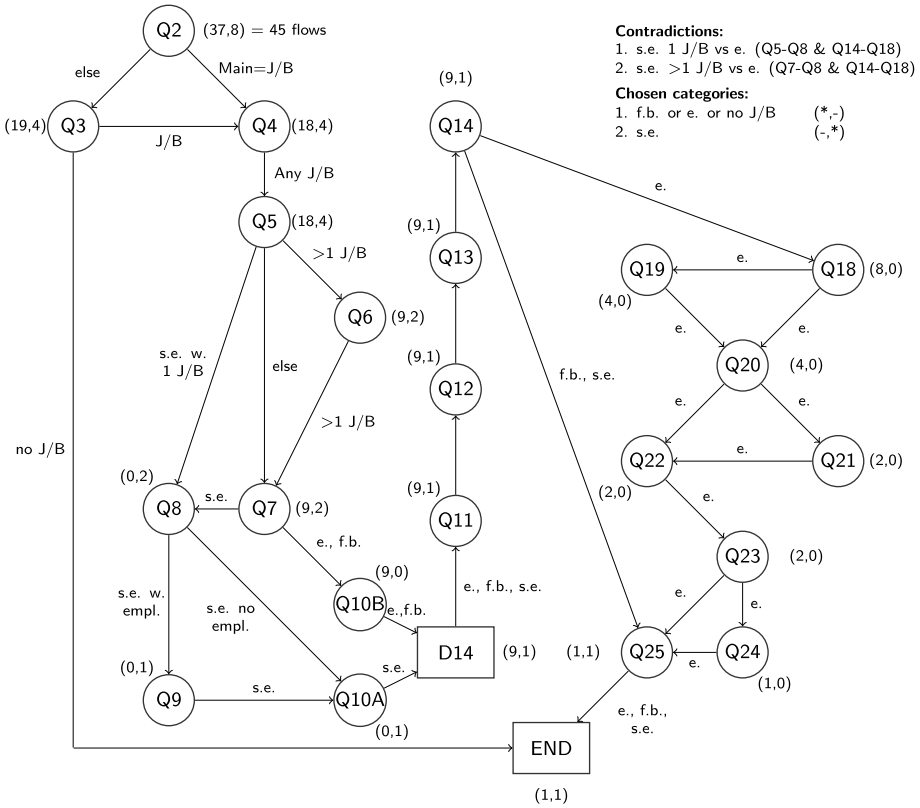


FIG 7. ***EM From Survey***. *Two categories of flows. Flow count-up.*

Q7: **In your *MRJ/B*, were you: an employee, self employed, working in a family business without pay?** Categories *e.*, *s.e.*, *f.b.* are created. Then {Q7= *s.e.*} go to Q8, everybody else goes to Q10B.

Q8: **Did you have any employees?** All *s.e.* are asked Q8. {Q8 = *yes*} go to Q9, {Q8 = *no*} go to Q10A.

Q9: **On the average, how many employees did you have during the *R.P.*?**

Q10A: **What was the name of your business?** Q10A is asked of all *s.e.*, after which they go to D14.

Q10B: **What was the name of your most recent employer (*MRE*) during the *R.P.*?** Q10B is asked of all *e.* and *f.b.*, which then go to D14.

D14: An internal definition EmpName is created for all *e.*, *s.e.*, *f.b.*.

Q11: **What kind of business, industry or service was this?** All *e.*, *s.e.*, *f.b.* are asked.

Q12: **What kind of work were you doing?**

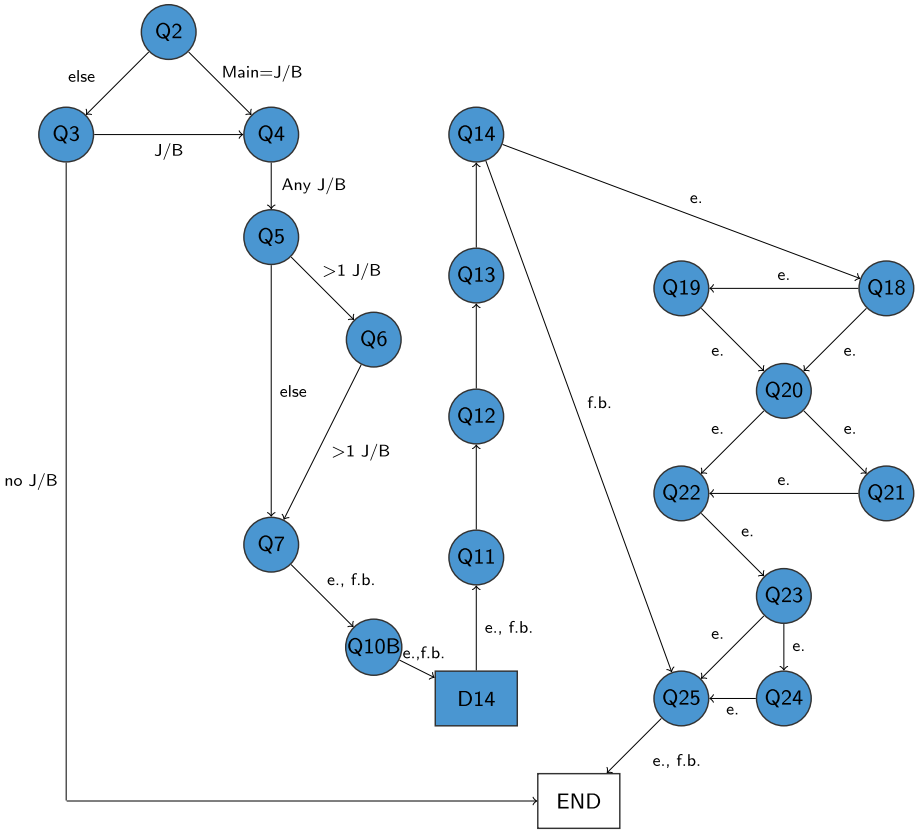


FIG 8. ***EM From Survey***. *Category 1: e. or f.b. or no J/B.*

Q13: **What were your most important activities or duties?**
Q14: **When did you start working for *EmpName*?**
C18: This condition redistributes the population as follows: *s.e.* and *f.b.* skip to Q25, *e.* go to Q18.
Q18: **In this job, were you a union member?** {Q18 = *yes*} go to Q20
Q19: **Were you covered by a union contract or collective agreement?**
Q20: **Was this job permanent, or is there some way that it was not permanent? For example, seasonal, temporary, term or casual?** {Q20 = *permanent*} go to Q22, the others go to Q21.
Q21: **In what way was this job not permanent?** The possible categories: seasonal job, temporary, term or contract job (non seasonal), casual job, other, are not read to the respondent. Next, this population goes to Q22.
Q22: **About how many persons were employed at the location were you worked for *MRE*? Would it be...?** Categories are read to respondents.

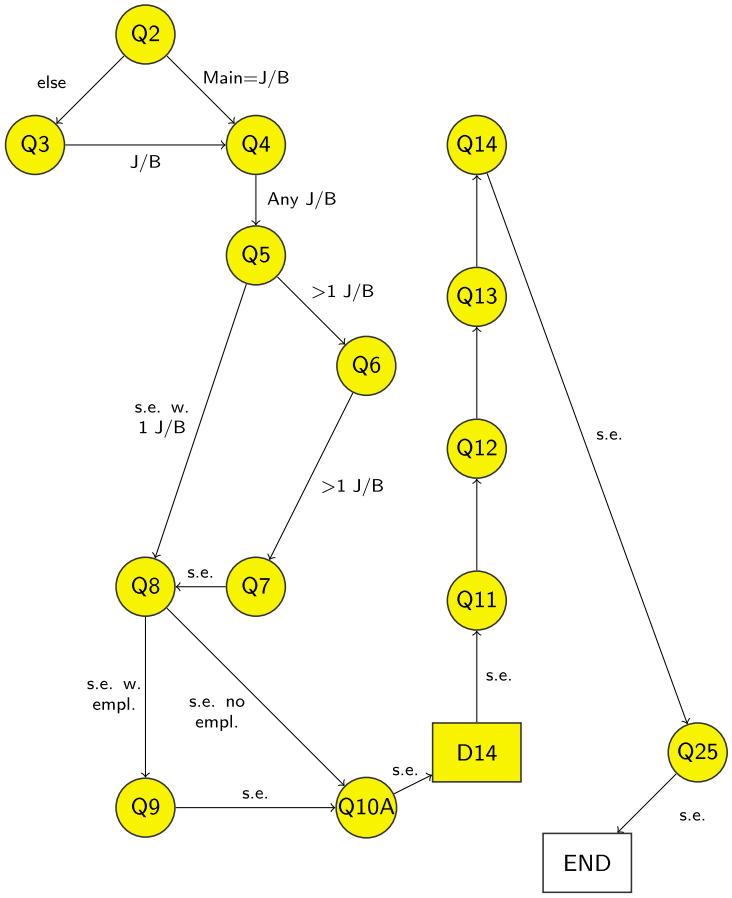


FIG 9. ***EM From Survey***. *Category 2: s.e.*

Q23: **Did *MRE* operate at more than one location?** {Q23 = *no*} goes to Q25, the others go to condition C24.

C24: This condition uses earlier information. {Q23 = *yes*} and {Q22 $\geq$ 500} go to Q25, whereas {Q23 = *yes*} and {Q22 $<$ 500} go to Q24.

Q24: **In total, about how many persons were employed at all locations? Would it be...?** Categories are read to respondents. This population goes to Q25.

Q25: **How many paid hours did you usually work per week at this *MRJ/B*?** This last question is asked of *e.*, *s.e.*, *f.b.*.

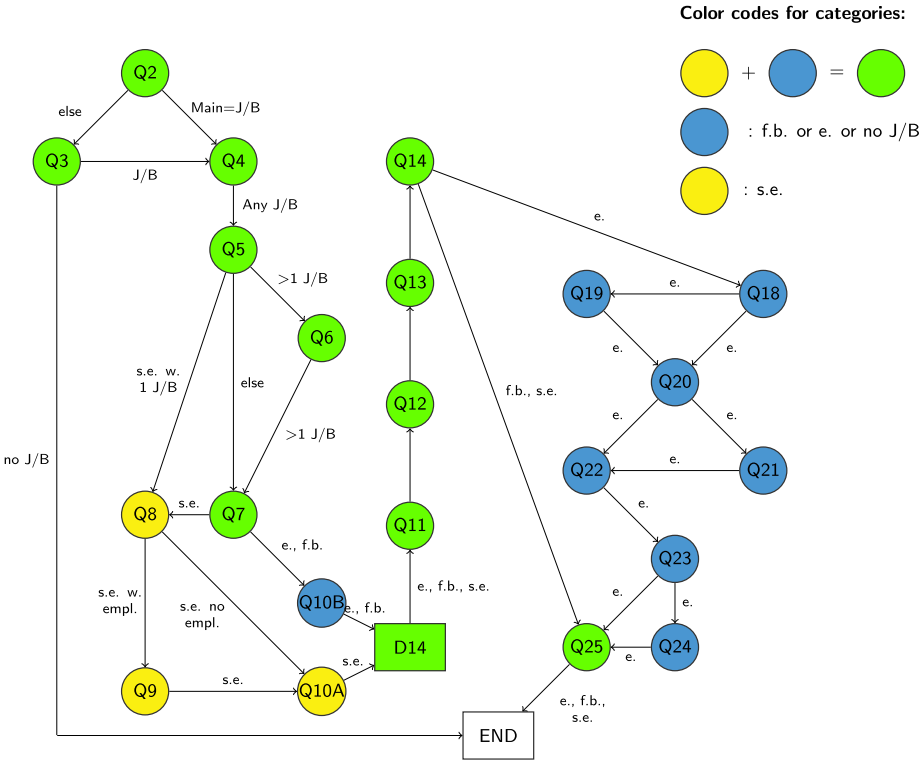


FIG 10. ***EM From Survey***. *Colour representation of all categories.*

Appendix B: Algorithm for flow count – Root to end

The algorithm determines the number of non-zero flows in a survey chart A, with K categories of flows.

Notation and rationale

Let q be a questionnaire with survey chart A and questions $Q_j, j = 0, \ldots, M$, of which Q_0 is the root and Q_M the node END_A. Let V be the associated vertex set and $C^{(k)}, k = 1, \ldots, K$ the K categories of flows. Let $parent(Q_j)$ be the set of all parents of Q_j (see 1.3), and $card(IN^{(k)}(Q_j))$ the number of all flows in category $C^{(k)}$ that enter $Q_j, k = 1, \ldots, K$ (see 2.1 and 2.2).

The input to this algorithm is A, with the additional information that the K categories are known, and each arc has annotations giving information on which categories travel that arc.[1]

[1]Every survey chart in Appendices A and D is a valid input. In Appendix A, where this algorithm has already been applied, the vectors attached to its nodes are the output and not part of the input.

For each node Q_j, we calculate the *K-dimensional* vector of non-zero flows, $v_j = (card(IN^{(1)}(Q_j)), card(IN^{(2)}(Q_j)), \ldots, card(IN^{(K)}(Q_j))) = (v_j^{(k)})_{k=1,\ldots,K}$. For this, we partition the vertex set V into two subsets: V_1, which consists of all questions for which we already computed the vectors v_j, and $V_2 = \{Q_j \in V : Q_j \notin V_1\}$, the remaining vertices. We initialize $V_1 = \{Q_0\}$, and sequentially increase V_1 by adding one question from V_2, say Q_i, for which $parent(Q_i) \subseteq V_1$. Note that since A is finite and acyclic, finding Q_i is always possible, as long as $V_1 \neq V$.

We compute $card(IN^{(K)}(Q_i))$ by adding $card(IN^{(K)}(Q_p))$ over all parents Q_p of Q_i, for which the arc Q_pQ_i, is traveled by some population in category k. Each component $v_i^{(k)}$ of the vector v_M represents the number of (non-zero) flows in $C^{(k)}$ for the entire A. The number of all non-zero flows in A is the sum of all components of v_M.

Procedure

Let $v_0 = (1, 1, \ldots, 1), V_1 = \{Q_0\}$, and $V_2 = \{Q_j \in V : j \neq 0\}$.

1. Find $Q_i \in V_2$, with $parent(Q_i) \subseteq V_1$. Increase the set V_1 by adding to it the question Q_i.
2. Compute each component $v_i^{(k)}$ of v_i, using the formula:

$$v_i^{(k)} = \sum_{\{p:Q_p \in parent(Q_i)\}} v_p^{(k)} I_p^{(k)}(Q_i),$$

where $I_p^{(k)}(Q) = 1$ if Q_pQ is travelled by a subpopulation in category $C^{(k)}$, and 0 otherwise.

Repeat steps 1 and 2 until $V_1 = V$, that is, until $Q_M = END_A$ belongs to V_1.

Appendix C: Technical details

Section 2.3 equations details

To prove (7) from (4), we write:

$$E_A = \sum_{i=1,\ldots,N} n_i p_i = \sum_{i=1,\ldots,N} \sum_{j=0,\ldots,M-1} I_j(f_i) p_i = \sum_{j=0,\ldots,M-1} \sum_{i=1,\ldots,N} I_j(f_i) p_i$$

where $I_j(f_i)$ is the indicator function associated with question j: it is 1 if Q_j is on flow f_i, and 0 otherwise $j = 0, \ldots, M-1$. Now

$$E_A(Q_j) = \sum_{i=1,\ldots,N} I_j(f_i) p_i = P_A(Q_j) = P_A(cover_A(Q_j)), j = 0, \ldots, M-1$$

Replacing this in the equation above gives (7).

Formula (6) is obtained as follows:

$$P(Q_j) = \sum_{i=1,\ldots,N} I_j(f_i)/N = N_j/N$$

We can express the probabilities in (5) in terms of transition (conditional) probabilities on flows.

For each question Q we have, using (3):

$$\begin{aligned} P(Q) = P(cover(Q)) &= \sum_{\{i \in IN(Q)\}} P(RQ_{i1} \ldots Q_{i(j-1)}Q) \\ &= \sum_{\{i \in IN(Q)\}} P(f_i(Q+)), \qquad (11) \\ &= \sum_{\{i \in OUT(Q)\}} P(f_i(Q-)),\ \textit{where } Q = Q_{ij}. \end{aligned}$$

From the second equality, we can further obtain, in terms of conditional probabilities:

$$\begin{aligned} P(Q) = &\sum_{\{i \in IN(Q)\}} P(Q/Q_{i(j-1)} \ldots Q_{i1}R) \\ &\times P(Q_{i(j-1)}/Q_{i(j-2)} \ldots Q_{i1}R) \ldots P(Q_{i2}/RQ_{i1})P(Q_{i1}) \qquad (12) \end{aligned}$$

As before, $P(Q/Q_{i(j-1)} \ldots Q_{i1}R)$ represents the conditional probability of asking question Q given that all questions on the string $RQ_{i1} \ldots Q_{i(j-1)}$ have been asked. Note that when $Q = Q_{n_i}$, the probability in the equations above is $p_i, i = 1, \ldots, N$. Likewise, if $Q = R$, both these equations give 1.

Proof of Proposition 1. Using (1) within categories, and then adding over all categories, we obtain:

$$P(Q) = \sum_{k=1,\ldots,K} \sum_{\{i \in IN^{(k)}(Q)\}} \sum_{\{j \in OUT^{(k)}(Q)\}} {p_{ij}}^{(k)}(Q).$$

The formula for conditional probabilities follows immediately from the definitions and (2), and the particular situation when $p_i = 1/N, i = 1, \ldots, N$ follows from the general case. □

Remark 11. We can consider a K-dimensional vector that stores information on the number of incoming flows per category for each question Q, namely $(card(IN^{(1)}(Q)), card(IN^{(2)}(Q)), \ldots, card(IN^{(k)}(Q)))$. An algorithm that calculates the components of this vector for each question is given in the Appendix B. We have a similar vector for the outgoing flows, $(card(OUT^{(1)}(Q)), card(OUT^{(2)}(Q)), \ldots, card(OUT^{(k)}(Q))$. We note that probability measure 2 of a question is the scalar product of these two vectors divided by N.

Section 4 details

Example 7′ details. We show directly that $P_A(cover_A(Q_b))>P_A(cover_{A'}(Q'_b))$, thus the difference in (10) is strictly positive. Let $B = \{Q_5 = no\}$ be the set of respondents who held *only one J/B* during *R.P.*, and B^c its complement in $\{held\ a\ J/B\ in\ R.P.\}$, or $B^c = \{Q_5 = yes\}$. $P_A(cover_A(Q_b)) = P_A(cover_A(Q_7)) = P_A(B^c) + P_A(B \bigcap \{Q_4 = no\})$ and $P_A(cover_{A'}(Q'_b)) = P_A(B^c \bigcap \{e.\ or\ f.b.\})$. We have $P_A(cover_A(Q_7)) > P_A(cover_{A'}(Q'_b))$, since $P_A(B^c) > P_A(B^c \bigcap \{e.\ or\ f.b.\})$. Note that $P_A(cover_A(Q_7))$ also contains a second term.

Example 8′ details. We have $\Omega_{A'} = A'_1 \bigcup A'_2 \bigcup A'_3 \bigcup A'_4$, where:

$$
\begin{aligned}
A'_1 &= \{Q'_{20} = Res\},\\
A'_2 &= \{Q'_{20} = (Dk, Rf)\} \bigcap Q'_{21} = not\ permanent\},\\
A'_3 &= \{Q'_{20} = (Dk, Rf)\} \bigcap \{Q'_{21} = permanent\} = \{\omega_1\} \bigcup \{\omega_2\},\\
A'_4 &= \{Q'_{20} = (Dk, Rf)\} \bigcap \{Q'_{21} = (Dk, Rf)\} = A_4.
\end{aligned}
$$

We also have, from Example 8, that $\Omega_A = A_1 \bigcup A_2 \bigcup A_3 \bigcup A_4 = A'_1 \bigcup A'_2 \bigcup A'_4$, if we identify $A'_1 = A_1 \bigcup A_2$, $A'_2 = A_3$, $A'_4 = A_4$. It follows that $\omega_i \notin \Omega_A, i = 1, 2$. The identification above is not fully justified, though. Although the individuals classified in A'_2 or A_3 give identical survey information, their behaviour as partial respondents is different. More importantly, the surveyed individuals in A'_1 are all "eager respondents", willing to answer a complete question right away. On the other hand, individuals in A_2 have been prepared to answer a question through a preliminary question. In some similar situations, we may expect that, as sets of individuals, $A'_1 \subset A_1 \bigcup A_2$. In our case, though, the text of Q_{20} is so lengthy, that some respondents to Q_{20} may actually refuse to answer Q_{21}. A better example of a preliminary question is Q_5 to Q_6 in Example 9. In Example 8′, transformation 1 might enhance the analytical potential of A, in addition to reducing response burden.

Example 8″ details. The second statement is obvious, as A'' has only one question. To prove the first statement, we first write:

$$\Omega_{A''} = \{Q_{20} = permanent\} \bigcup \{Q_{21} = Res\} \bigcup \{Q'_{20} = (Dk, Rf)\}.$$

Recalling the decomposition of $\Omega_A = A_1 \bigcup A_2 \bigcup A_3 \bigcup A_4$ in Example 8, we see that $\Omega_{A''} = A_1 \bigcup A_2 \bigcup A_4$. The set $A_3 \subset \Omega_A$ consists of two analytical outcomes, $\omega_1 = \{Q_{20} = not\ permanent\} \bigcap \{Q_{21} = Dk\}$ and $\omega_2 = \{Q_{20} = not\ permanent\} \bigcap \{Q_{21} = Rf\}$. However, $\omega_i \notin \Omega_{A''}, i = 1, 2$

Appendix D: EM Analytical and EM Simplified

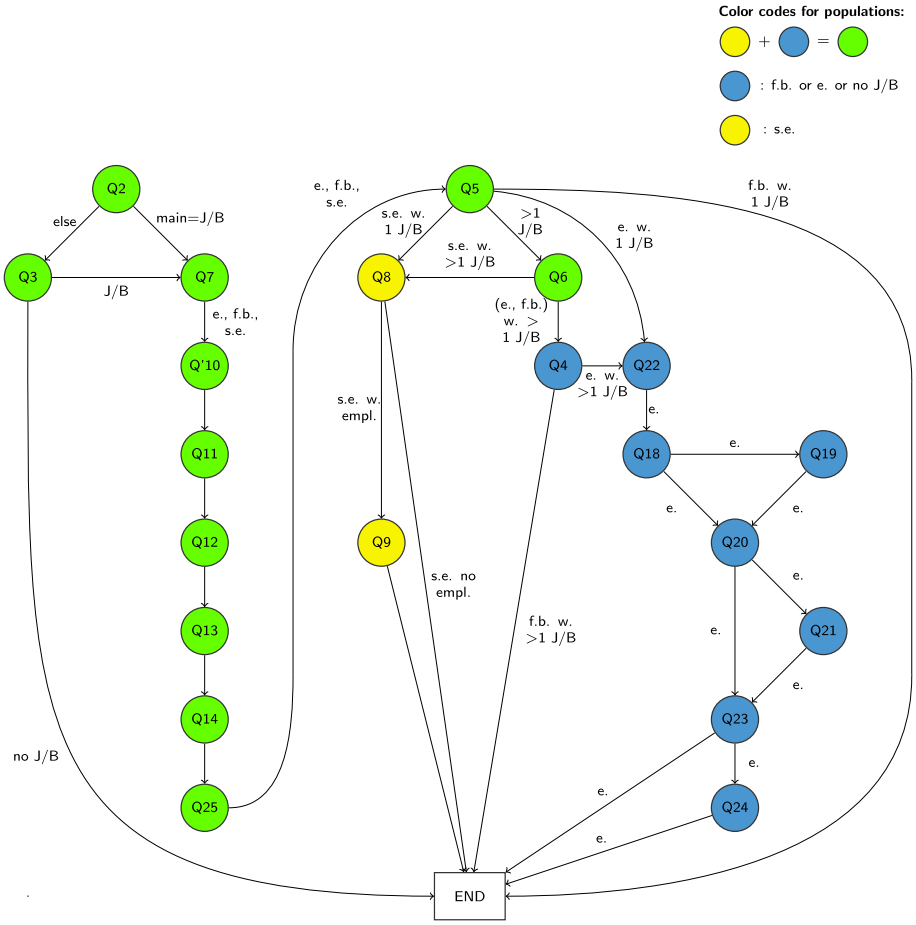


FIG 11. ***EM Analytical***. *One category of flows.*

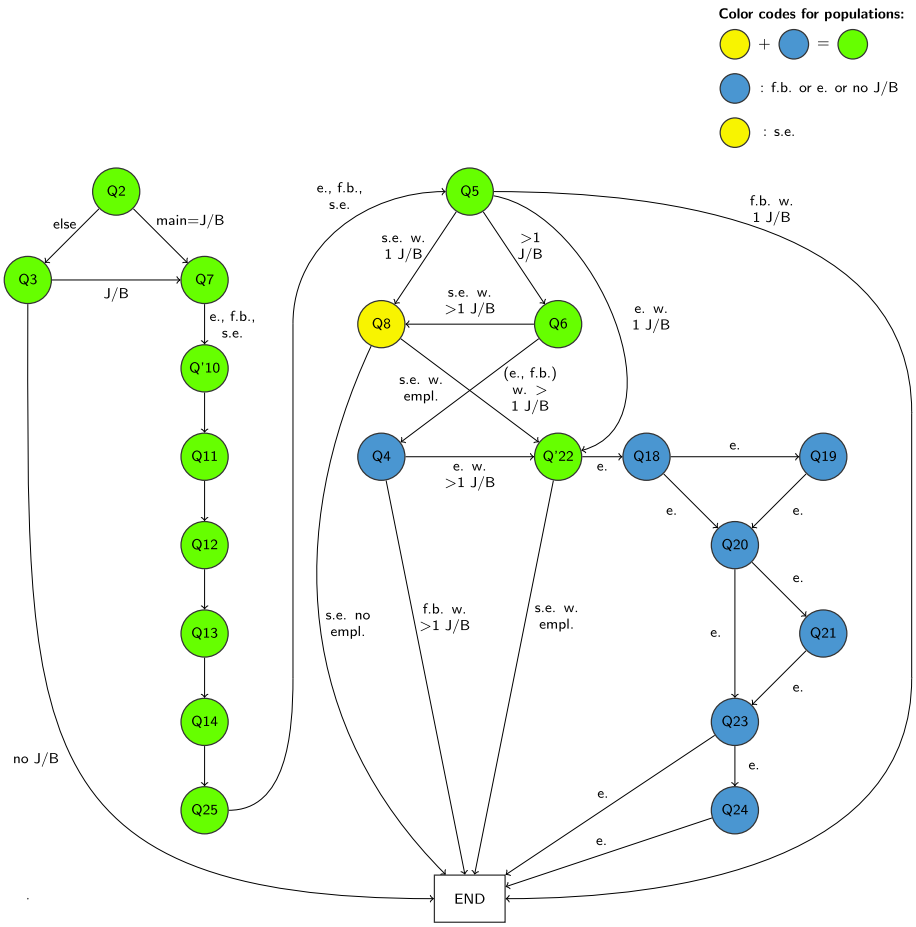


FIG 12. ***EM Analytical***. *Merged.*

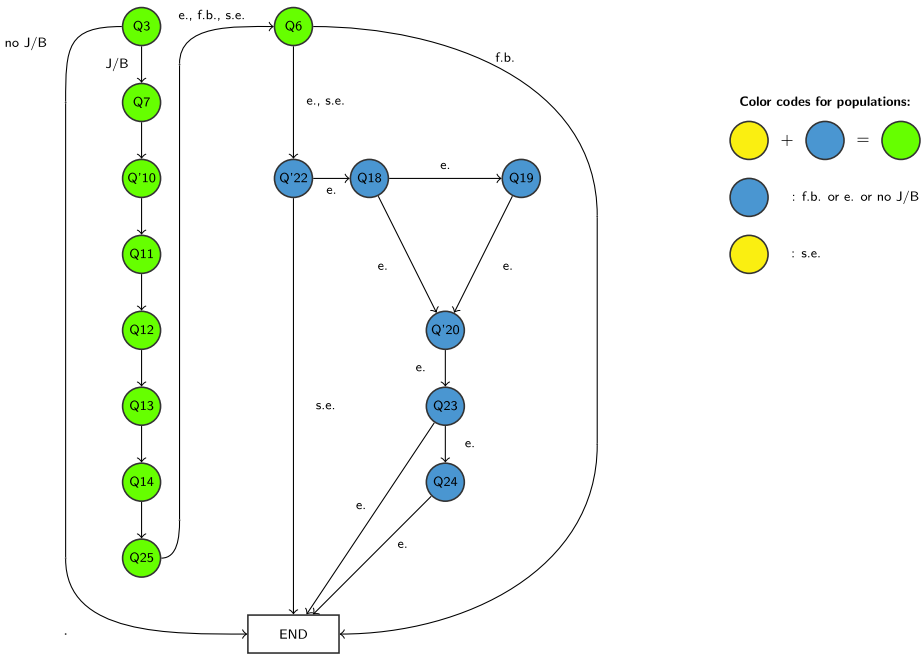
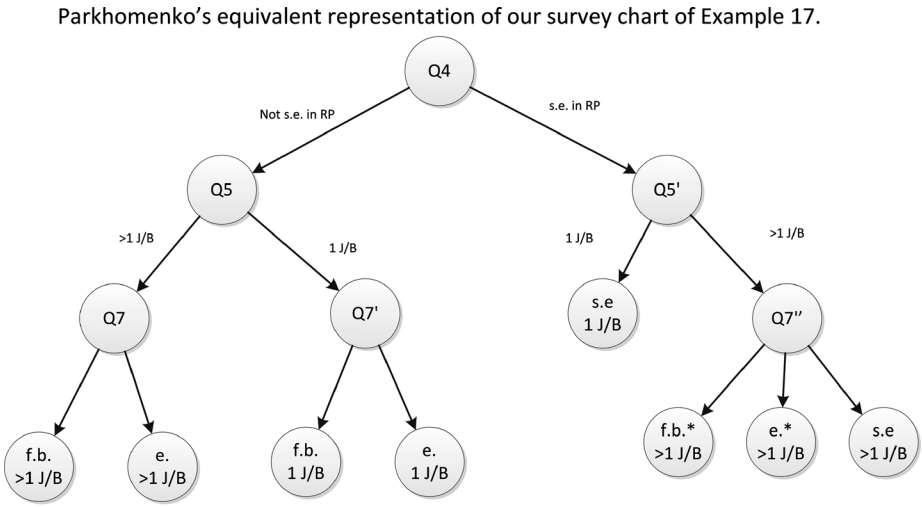
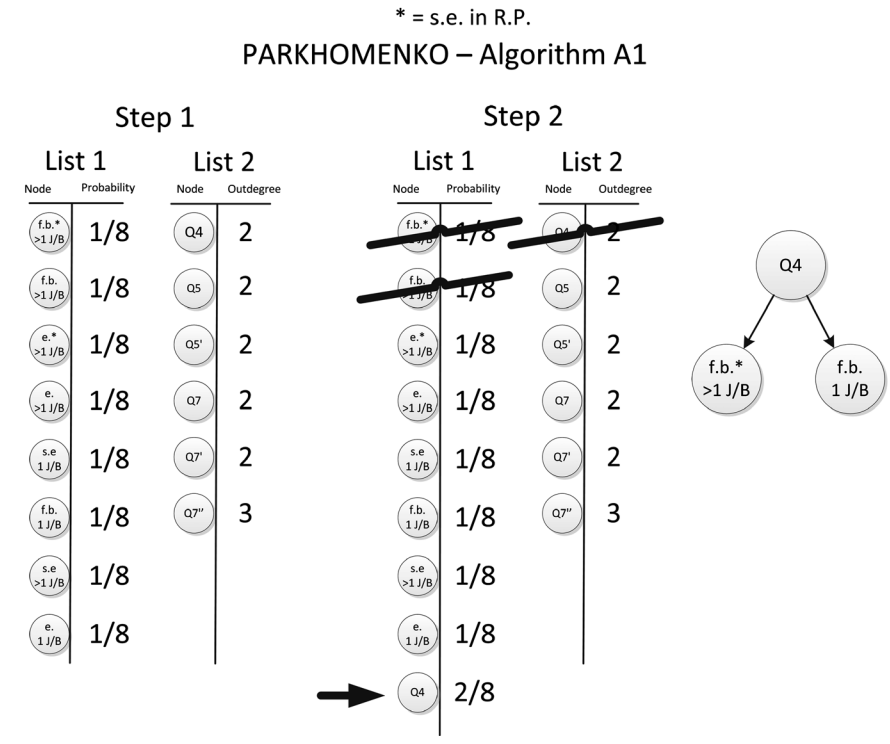


FIG 13. ***EM Simplified***. *One category of flows.*

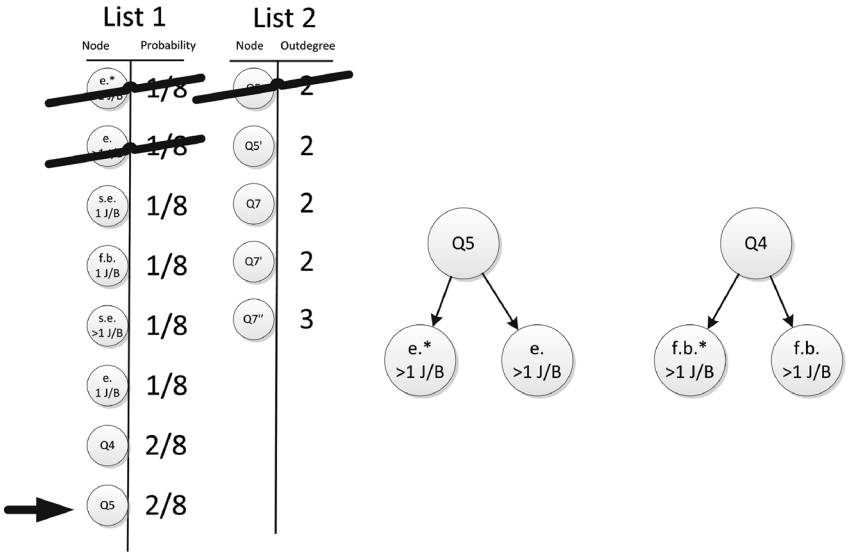
Appendix E: Parkhomenko's Algorithm A1



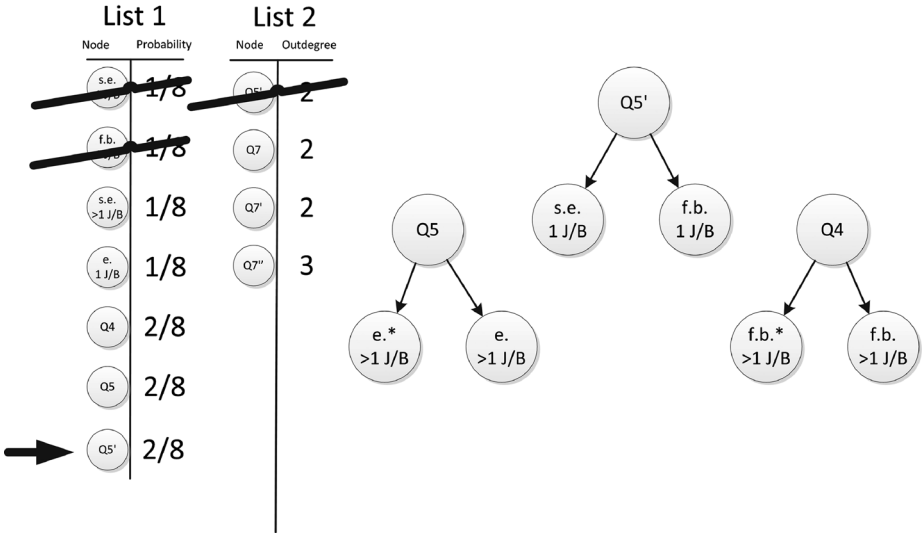


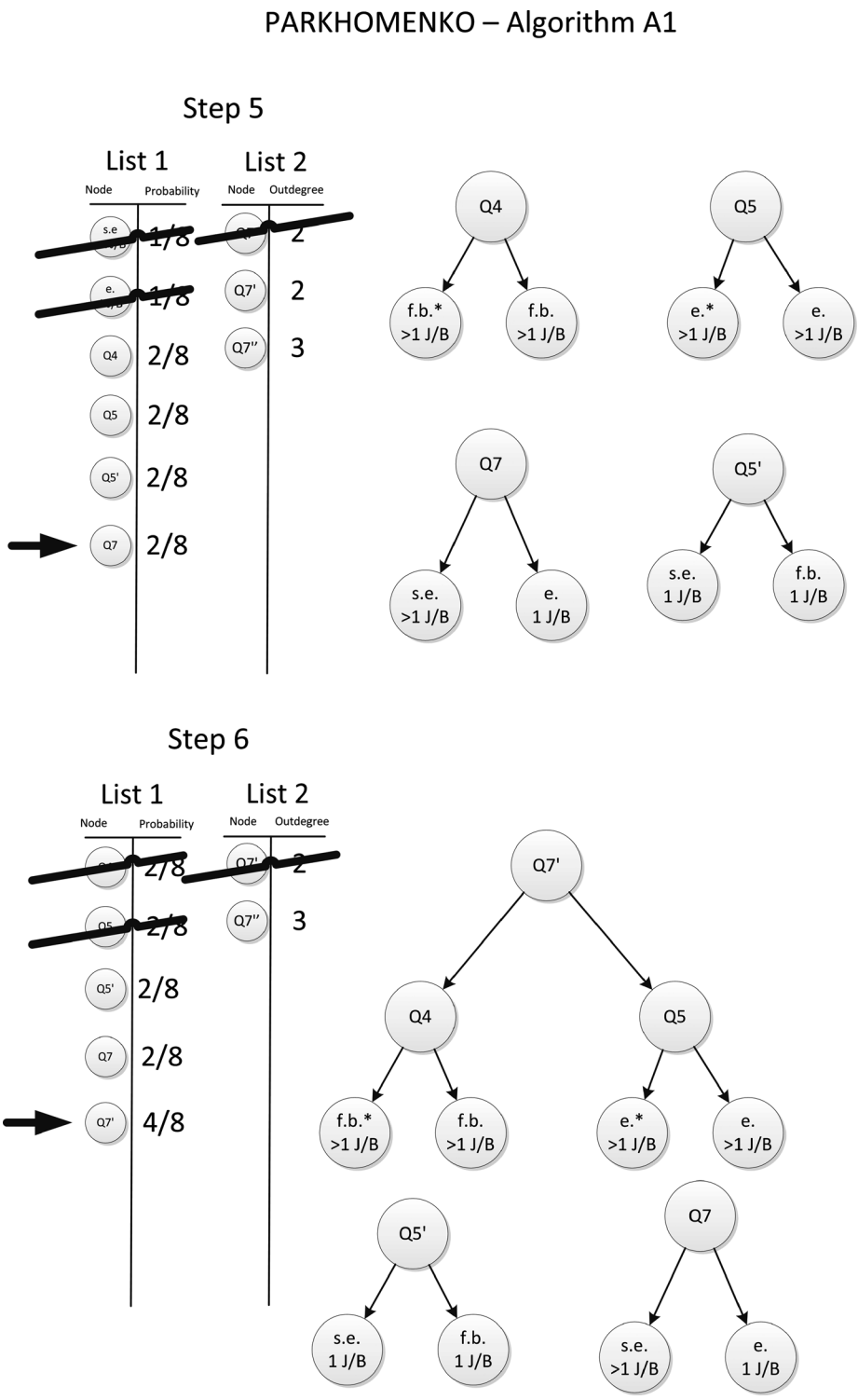
PARKHOMENKO – Algorithm A1

Step 3

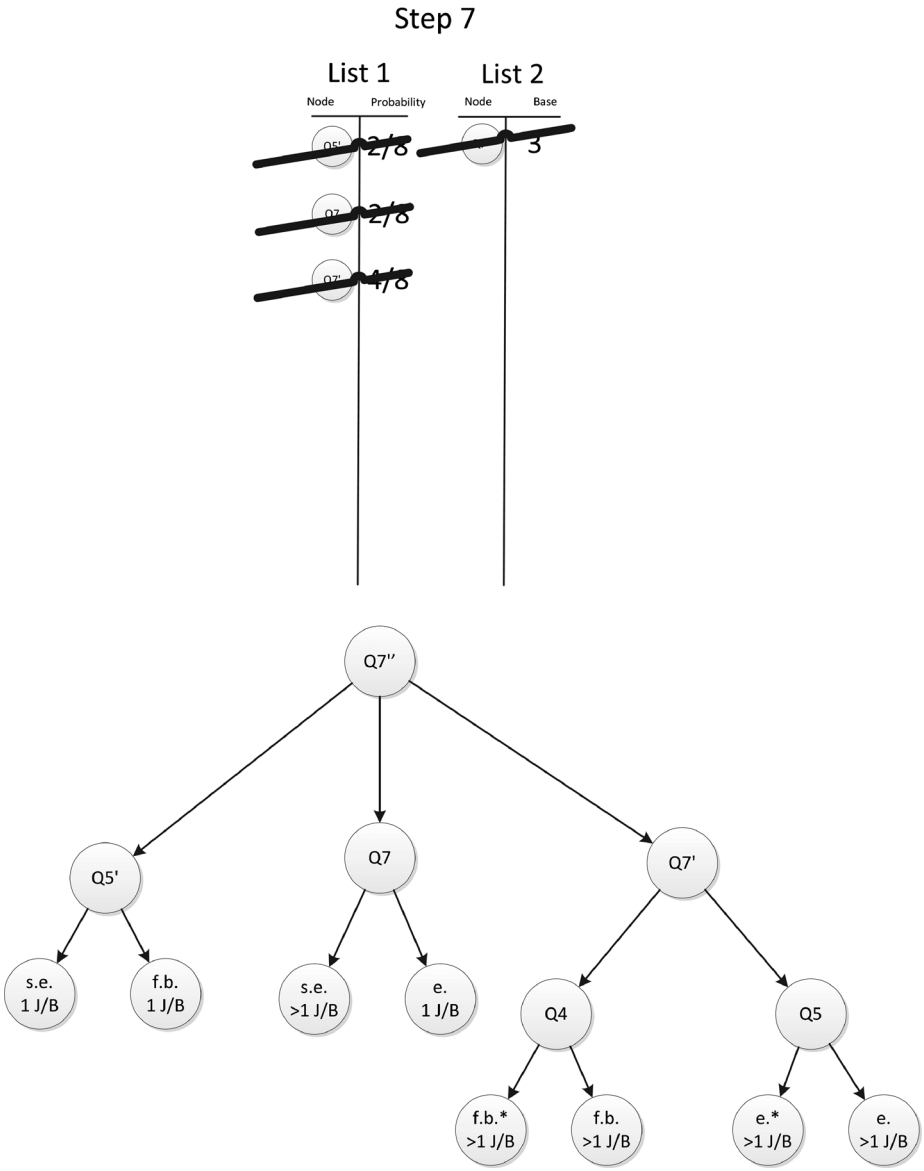


Step 4

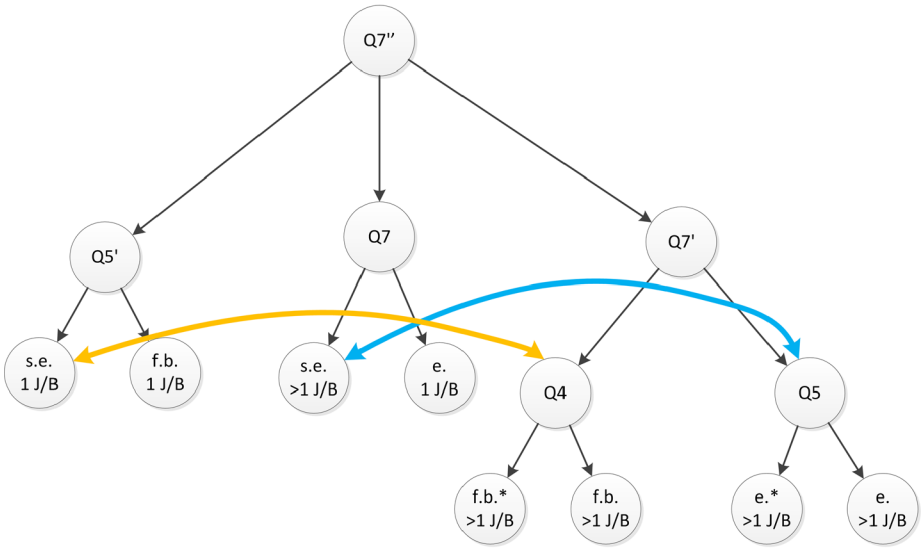


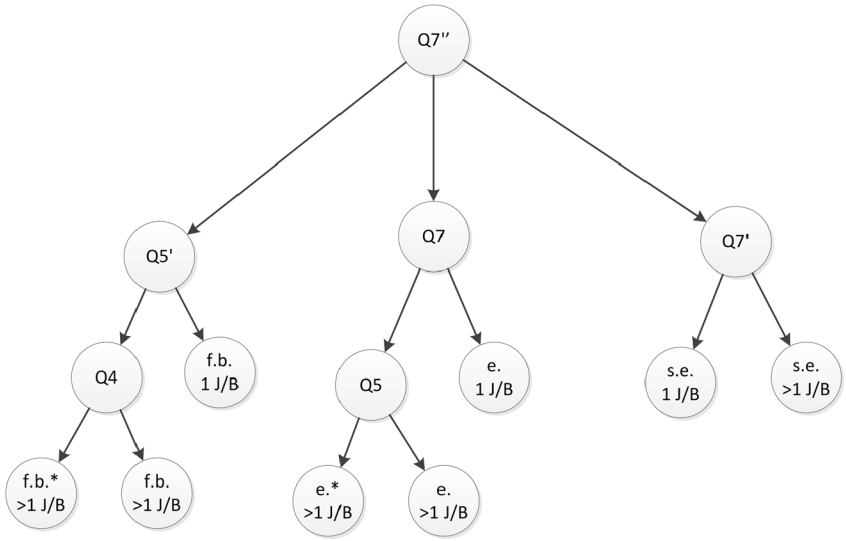

PARKHOMENKO – Algorithm A1
Step 5
List 1
Node
Probability
List 2
Node
Outdegree
Q7'
2
Q7''
3
Q4
2/8
Q5
2/8
Q5'
2/8
Q7
2/8
Q4
f.b.*
>1 J/B
f.b.
>1 J/B
Q5
e.*
>1 J/B
e.
>1 J/B
Q7
s.e.
>1 J/B
e.
1 J/B
Q5'
s.e.
1 J/B
f.b.
1 J/B
Step 6
List 1
Node
Probability
List 2
Node
Outdegree
Q7'
2
Q7''
3
Q5'
2/8
Q7
2/8
Q7'
4/8
Q7'
Q4
Q5
f.b.*
>1 J/B
f.b.
>1 J/B
e.*
>1 J/B
e.
>1 J/B
Q5'
s.e.
1 J/B
f.b.
1 J/B
Q7
s.e.
>1 J/B
e.
1 J/B

PARKHOMENKO – Algorithm A1

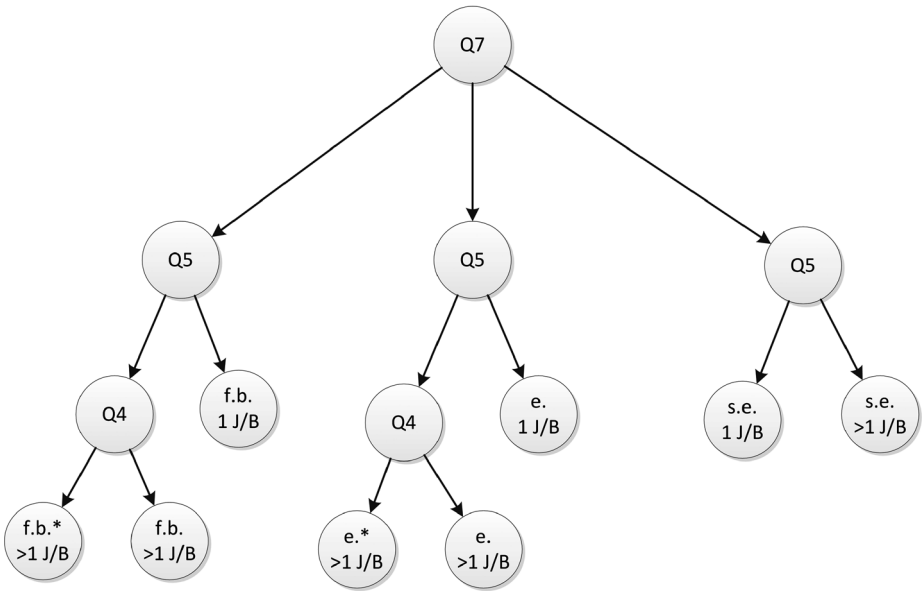


We make changes that do not affect the expected number of questions





By renaming the question nodes, we have a tree representing a questionnaire similar to that of Example 7', Figure 4.



References

Bethlehem, J. and Hundepool, A. (2004). TADEQ: A tool for the documentation and analysis of electronic questionnaires. *Journal of Official Statistics*, **20**, pp. 233–264.

Elliott, S. (2012). The application of graph theory to the development and testing of survey instruments. *Survey Methodology* **38**, pp. 11–21.

Fenn, L. (2015). Decision Trees and Surveys. *Technical report-Hunter College-CUNY.* http://larryfenn.com/decisiontrees.pdf

Harary, F. (1969). *Graph Theory.* Reading, Mass.: Addison-Wesley. MR0256911

Huffman, D.A. (1952). A method for the construction of minimum-redundancy codes. *Proc. Inst. Radio Engrs.* **9**, pp. 1098–1102.

Jabine, T.B. (1985). Flow charts: a tool for developing and understanding survey questionnaire. *Journal of Official Statistics* **1** pp. 189–207.

Kotsiantis, S.B. (2013). Decision trees: a recent review. *Intelligence Review* 39, pp. 261-283.

Krosnick, J.A., and Presser, S. (2010). Question and Questionnaire Design. In P.V. Marsden & J.D. Wright (Eds.), *Handbook of Survey Research*, (Second ed.). Emerald Group Publishing.

Marsden, P.V. and Wright, J.D. (2010). *Handbook of Survey Research, Second Edition*. Emerald Group Publishing.

McCabe, T.J. (1976). A complexity measure. *IEEE Transactions on Software Engineering* **SE-2**(4), pp. 308–320. MR0445904

Murphy, S.K. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery* **2**, pp. 345–389.

Parkhomenko, P.P. (2010). Questionnaires and Organizational Hierarchies. *Automation and Remote Control* **71**(6), pp. 124–134. MR2724789

Payne, S.L. (1949). Case study in question complexity. *Public Opinion Quarterly* **13**, pp. 653–658.

Picard, C. (1965). *Théorie des Questionnaires*, Les Grands Problèmes des Sciences (Vol. 20). Paris: Gauthier-Villars.

Safavian, S.R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions n Systems, Man and Cybernetics* **21**(3), pp. 660–674. MR1130731

Trépanier, K. (2013). A structural approach to CATI questionnaire design using graphs-revised. *Co-op term report, University of Ottawa*.