Survey questionnaires and graphs

Ioana Schiopu-Kratina, Kyle Trépanier [1] and
Christina Maria Zamfirescu[2]
[1]University of Ottawa, Department of Mathematics and Statistics

[2]Hunter College and Graduate Center-CUNY, Department of Computer science and Mathematics

QDET2
Miami, Florida
November 9-13, 2016

Overview

1.1 Motivation for this research

- Data collection is the most expensive part of running a survey.

- The quality of the published data depends on the design and length of the questionnaire.

- The use of computers has opened the door to the development of large and complex questionnaires, which should be well structured.

- Theoretical research on the structure of questionnaires was done by Picard (1965) and Parkhomenko (2010). More research is required.

1.2 The challenge

• Our examples are based on a version of the module "Most Recent Employment" (**EM**), of the 2008 "Access and Support to Education and Training Survey" (ASETS). ASETS has 47 modules with 4 - 25 questions/module.

• The content and number of questions of ASETS was determined by many analysts, who had different interests.

- The logical link among questions was not always apparent.
- It is difficult to place a variety of questions within a common questionnaire.

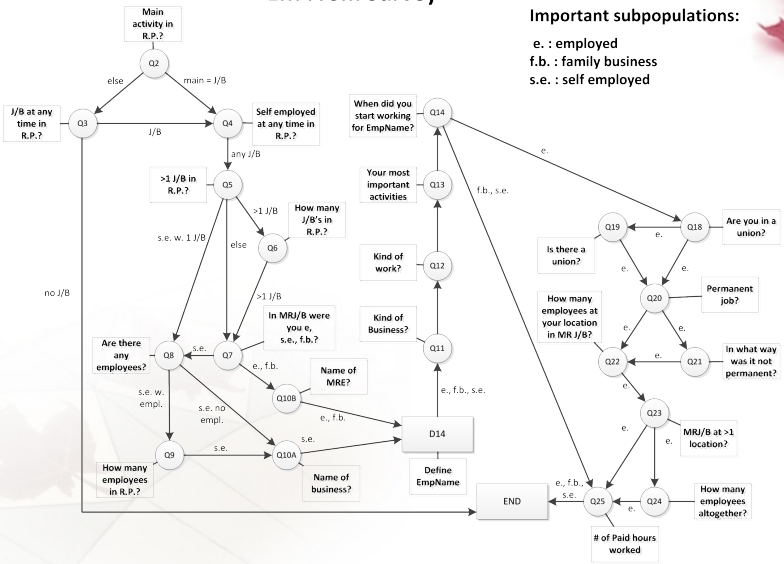
• We propose practical rules to structure questionnaires using graphs.

1.3 Our requirements

- The coverage (target) of each question should be apparent.
- For each surveyed subpopulation, the questions should be sequenced in a logical order.
- The paths followed by important subpopulations should be apparent in the questionnaire.
- While the structure may change, the content of the questions should be preserved.
- The burden on respondents should be minimal.
- We apply a set of transformations to attain the above stated goals.

1.4 Survey charts as graphs: Example 1

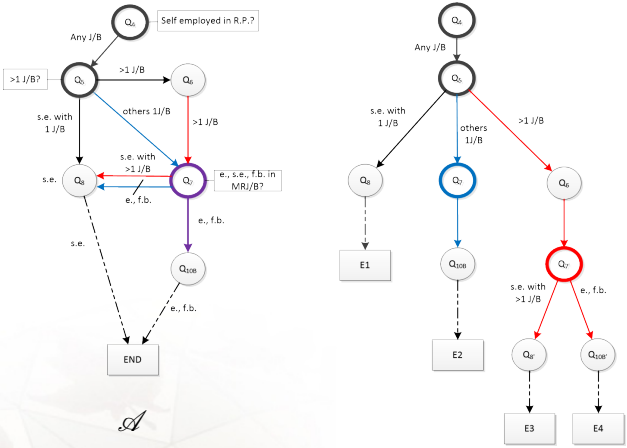
EM From Survey



Important subpopulations:

e. : employed
f.b. : family business
s.e. : self employed

1.4 Survey charts as graphs: Example 2



$\mathscr{A}$

- The survey chart $\mathscr{A}$ is a directed, acyclic graph, with root $R = Q_4$ and a node $END_{\mathscr{A}}$. **The path $Q_5Q_7Q_8$ is empty.**
- The directed tree on the right represents $\mathscr{A}$. Here Q_7, with two parents in $\mathscr{A}$, appears twice in the tree.

2.1 Basic definitions

Definition 1. A *flow* $f_i = RQ_{i1}...Q_{in}F_i$ in $\mathscr{A}$ is a string of questions (nodes) connected by arcs, which starts at the root R and continues to $END_{\mathscr{A}}$. The conditions defining all its arcs are consistent, so a surveyed population could travel through the flow.

The *elementary event* F_i stands for the category of individuals that travel through f_i. It is part of $END_{\mathscr{A}}$.

Definition 2. An *analytical outcome* ω of a questionnaire q is a minimal group of individuals defined by their specific answers to questions in a flow.

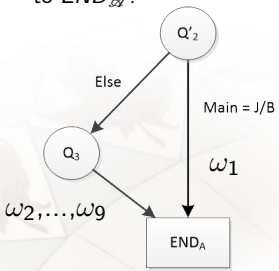
The set of all analytical outcomes Ω_q is the *analytical potential* of q.

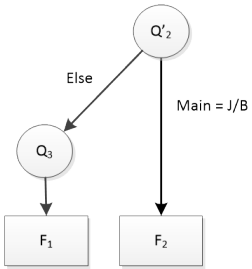
2.1 Basic definitions: Example 3

Example 3. Consider a questionnaire q with two questions:

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...** If $\{Q_2' \neq J/B\}$, go to Q_3. Else go to $END_{\mathscr{A}}$.

Q_3 **: Did you work at a J/B at any time during the R.P.?** Go to $END_{\mathscr{A}}$.





2.2 Probabilities of flows and events

- Consider $(\Omega_q,\ \sigma_A(F_i, i = 1, \cdots, N),\ P_A)$, where the elementary event F_i is assigned a probability p_i of occurrence, $\sum_{i=1}^{N} p_i = 1$.
- Important particular case: $p_i = \frac{1}{N},\ i = 1, ..., N$. This is useful when there are many elementary events, but can be misleading otherwise.

Definition 3. The *coverage* of a question Q is:
$cover_A(Q) := \bigcup_{\{1 \le i \le N : Q \in f_i\}} f_i$, and

$$(1)\ P_A(Q) = \sum_{\{1 \le i \le N : Q \in f_i\}} p_i,$$

where p_i is the probability of f_i.

2.2 Probabilities of flows and events

- If q has M questions, the *expected number of questions* is:

$$(2)\ E_A = \sum_{j=0}^{M-1} P_A(Q_j)$$

- We minimize (2). When every path can be travelled,

$$(3)\ P_A(Q) = \sum_{i \in IN(Q)} \sum_{j \in OUT(Q)} P_A(f_{ij}(Q))$$

where $f_{ij}(Q) = RQ_{i1} \cdots Q_{ik} QQ_{j,k+2} \cdots F_{ij}$. Here the flow $f_{ij}(Q)$ is put together from an "incoming" flow $f_i(Q)$ and an "outgoing" flow $f_j(Q)$.

2.2 Probabilities of flows and events

When $p_i = \frac{1}{N}$, $i = 1 \cdots N$, (4) becomes:

$$P_A(Q) = \frac{card(IN(Q)) \times card(OUT(Q))}{N}$$

• Survey charts visually pack information, but may contain empty paths.

• One must partition the flows into *categories of flows*, within which all paths are travelled.

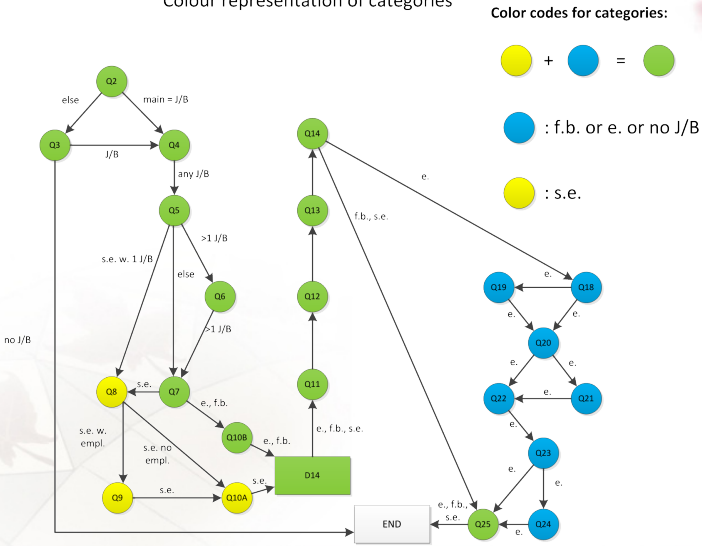
• In calculating $P_A(Q)$, one must account for categories of flows.

• The categories correspond to analytically significant subpopulations.

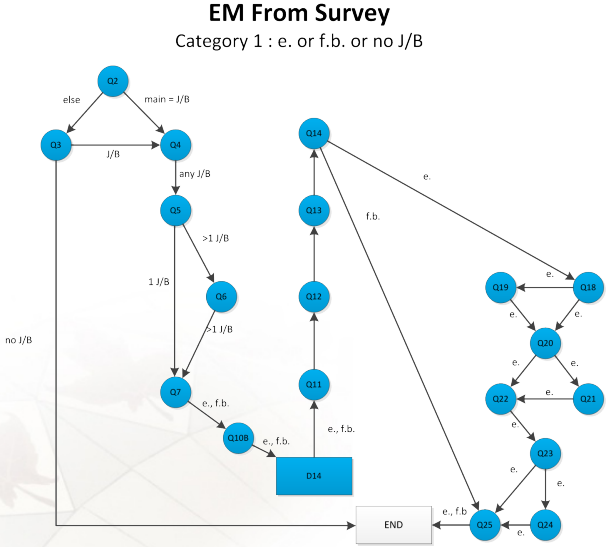
2.3 Categories of flows: Example 4

EM From Survey – 2 categories

Colour representation of categories



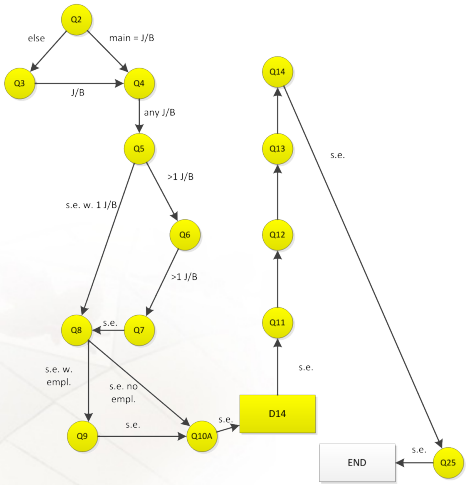
2.3 Categories of flows: Example 4



2.3 Categories of flows: Example 4

EM From Survey

Category 2 : s.e.



3.1 Overview of transformations

• Transformation 1: Reverses the order of Q_a and Q_b, when Q_b is a descendent of Q_a in $\mathscr{A}$, which collects more information than Q_a. The transformed survey chart $\mathscr{A}'$ has $E_{\mathscr{A}'} \leq E_{\mathscr{A}}$.

• Transformation 2: When possible, joins two similar questions with disjoint coverage in $\mathscr{A}$, and moves the new question "closer" to the root in $\mathscr{A}'$.

• The transformations preserve the analytical potential.

• These transformations bring a survey chart closer to a tree, so more information is collected earlier in the survey.

3.2 Consequences of transformations

- The coverage of each question is easily read off $\mathscr{A}'$.

- The number of categories of flows can be reduced and it is easier to follow important subpopulations through $\mathscr{A}'$ (s.e., e., f.b.).

- $E_{\mathscr{A}'} \leq E_{\mathscr{A}}$

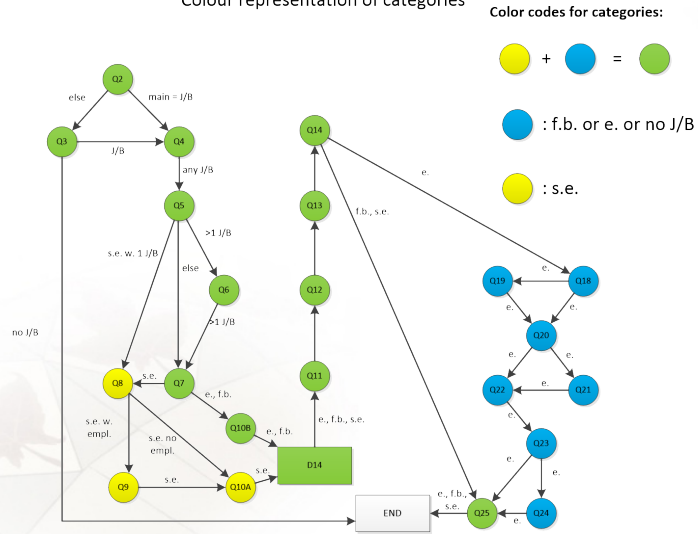
- The transformed survey chart points to questions that could be removed.

- We may move questions with high indegree closer to the root (e.g., Q_{25}) of **EM From Survey**.

3.2 Consequences of transformations: Example 4

EM From Survey – 2 categories $E_A = 14.889$

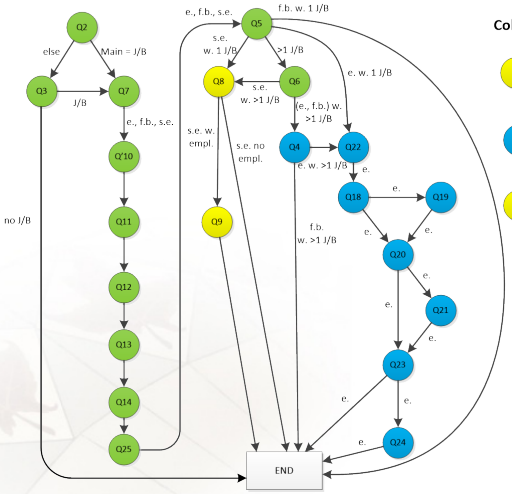
Colour representation of categories



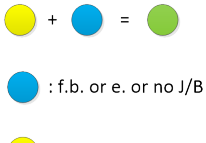
3.2 Consequences of transformations: Example 5

EM Analytical – 1 category

Colour representation of populations

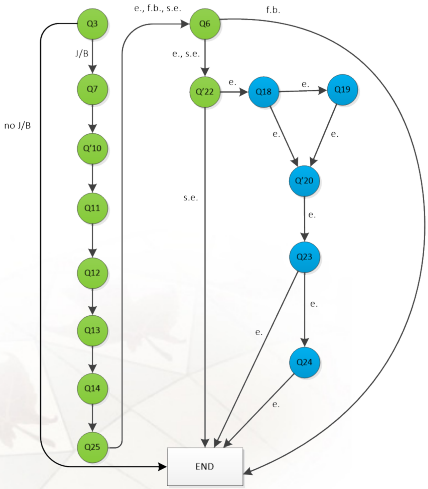


Colour codes for populations:



3.2 Consequences of transformations: Example 5

EM Simplified $E_A = 10.857$



Legend

: f.b. or e. or no J/B

: s.e.

Removed:
Q_2, Q_4, Q_5, Q_8

Combined:
Q_{20} & $Q_{21} \rightarrow Q'_{20}$
Q_9 & $Q_{22} \rightarrow Q'_{22}$

4.1 Conclusion and future work

- Starting with a flow chart used in practice, we construct a new type of graph, the survey chart.

- We perform transformations on a survey chart to make the corresponding questionnaire more amenable to analysis.

- We minimize the expected number of questions in a questionnaire, while preserving its analytical content.

- Future research includes devising an algorithm for finding categories, and collecting information more efficiently.

Thank you!

References

- Parkhomenko, P.P. (2010). Questionnaires and Organizational Hierarchies. *Automation and Remote Control* 71(6), pp 124-134.
- Picard, C., *Théorie des Questionnaires*, Gauthier-Villars Paris, Les Grands Problmes des Sciences XX, 1965.
- Schiopu Kratina, Ioana; Zamfirescu, Christina Maria; Trépanier, Kyle; Marques, Lennon. Survey questionnaires and graphs. Electron. J. Statist. 9 (2015), no. 2, 2202–2254. doi:10.1214/15-EJS1067. http://projecteuclid.org/euclid.ejs/1444308086.