Quality & Quantity
https://doi.org/10.1007/s11135-023-01680-1

A network model for multiple selection questions in opinion surveys

Stefano Benati[1] · **Justo Puerto**[2]

Accepted: 8 May 2023

Abstract
Opinion surveys can contain closed questions to which respondents can give multiple answers. We propose to model these data as networks in which vertices are the eligible items and arcs are the respondents. This representation opens up the possibility of using complex networks methodologies to retrieve information and most prominently, the possibility of using clustering/community detection techniques to reduce data complexity. We will take advantage of the implicit null hypothesis of the modularity function, namely, that items are chosen without any preferential pairing, to show how the hypothesis can be tested through the usual calculation of p-values. We illustrate the methodology with an application to Eurobarometer data. There, a question about national concerns can receive up to two selections. We will show that community clustering groups together concerns that can be interpreted in a consistent way and in general terms, such as Economy, or Security or Welfare issues. Moreover, we will show how different society groups are worried by different class of items.

Keywords Community detection · Modularity maximization · Modularity validation · Multiple choice and multiple selection questions · Public opinion's concerns · Eurobarometer

1 Introduction and model motivation

Some surveys contain closed questions in which respondents are proposed with a list of items among which they can elicit more than one answer, but with a maximum number of choices. For example, see European Commission (2018), the Eurobarometer standard survey formulates a question about citizens' concerns in which at most two answers can be chosen from a closed list:

What do you think are the two most important issues facing your country at the moment? (Max. 2 answers)

✉ Stefano Benati
stefano.benati@unitn.it

[1] Dipartimento di Sociologia e Ricerca Sociale, Università di Trento, Via Verdi 26, 38122 Trento, Italy

[2] IMUS. Universidad de Sevilla, Avda. Reina Mercedes s/n, 41012 Sevilla, Spain

- Crime
- Rising prices, inflation
- Taxation
- Unemployment
- Terrorism
- Housing
- Government debt
- Immigration
- Health and social security
- The education system
- Pensions
- The environment, climate and energy issues
- Economic situation
- Other
- None
- Don't know

There is a subtle but hidden problem when this kind of questions are to be analyzed. Answers are usually coded and then reported as the frequency by which *one* single item has been mentioned or not, see Rouet (2016); Bevan et al. (2016); Traber et al. (2022), or a Eurobarometer report, such as Brussels (2018), but actually respondents gave *two* answers. If pairs are broken, then reports could loose analytical detail as a respondent answering the pair *Immigration, Crime* could be substantially different from a respondent answering *Immigration, Unemployment*. In a sense, answers are *budget-constrained* because one cannot elicit all the issues by which he or she is worried, but only a subset of them. People can be worried by immigration, but if immigration is ranked third, then immigration is not recorded. Fixing the maximum number of answers introduces a bias in the statistics, too. As a result, due to the large number of zeros, e.g. *not mentioned*, with respect to ones, e.g. *mentioned*, correlations between items are mostly negative and close to zero. One may argue that the best way to analyze these data is to record the answers in pair, but it can be problematic. There is a combinatorial explosion of all the possible answers an individual can give as they are all possible pairs from the set of 14 items (excluding None and Don't know). They are a total of 91 pairs: one must find some other convenient and parsimonious way to represent these data to continue with the statistical analysis.

In this contribution, we will propose a network model to represent survey data coming from multiple selection survey questions. We will call this network the Items Graph: it is composed by nodes representing question choices/items and (multiple) arcs between nodes representing the actual answers by respondents. The graph contains as many arcs as the survey respondents, with possible loops corresponding to answers in which just one item has been elicited. Representing data as a network allows the possibility of using all the tools developed for complex network analysis, such as the use of centrality measures, see Das et al. (2018), the core-periphery segmentation, see Tang et al. (2019), or the community detection, see Fortunato and Hric (2016).

Taking inspiration from Bevan et al. (2016) and their work on items aggregation, we will show how to apply community detection and modularity maximization to the items graph. In that paper, authors empirically aggregated citizens' concerns in few classes that are used to detect whether there is a correspondence between issues that are considered important at the national, personal, and European level. Concerns were aggregated following a simple rule-of-thumb, without using any quantitative analysis. Indeed, the use of

community detection models and more specifically modularity maximization can be useful as, in the items graph, communities are subsets of concerns that respondents may deem as homogeneous. In this way a qualitative way of aggregating data is replaced by a quantitative one. Moreover, a peculiar advantage of modularity maximization is that the detected clusters can be validated by statistical inference. Indeed, a feature of modularity clustering is that it uses an implicit null hypothesis that assumes that there is no preferential pairing between nodes/items and therefore there are no significant clusters. Its formal definition is delayed to the following section, however, preferential pairing appears when some nodes/items pairs are mentioned more often than what is expected by independent probabilities. Applying the methodology proposed in Zhang and Chen (2017), we will show how to use this null hypothesis to statistical test and to determine if the communities resulting from modularity optimization are significantly different from random communities. The methodology that we will describe is how to calculate test p-values from modularity optimization.

Finally, we will use the Eurobarometer question about the most important national issue to make an exercise with the proposed methodology and to show how modularity maximization can be used in an applied research workflow. It will be seen that clusters are composed of concerns that are logically consistent, that clusters are statistically significant, and then they can be used to determine how different population segments are characterized by different concerns. Finally, we will compare modularity with other standard statistical techniques that can be used to similar purposes, namely, clustering and data reduction and we will show that modularity optimization achieves the most consistent results.

The paper is organized as follows. Section 2 is devoted to introduce the items graph and to prove some of its properties. Section 3 presents the clique partition model for modularity maximization and how to interpret it as an inferential model for hypothesis testing. Section 4 is devoted to describe the application of the newly developed methodology to the Eurobarometer data. In Sect. 5, the results of modularity optimization are compared to what obtained by other statistical techniques. Finally, Sect. 6 draws our conclusions and future research lines on the topic.

2 The items graph

The network model to represent a multi response question is defined as follows. Let $l_1, \ldots, l_n$ be the labels assigned to the answers of the multiple response question. The Items Graph $G = (V, E)$ is composed of the node set $V = \{1, \ldots, n\}$, corresponding to labels let $l_1, \ldots, l_n$, and there is an arc $(i, j) \in E$ for every respondent that answered the l_i, l_j pair. If the answer is a single item l_i, then the arc is a loop $(i, i) \in E$. The degree δ_i of a node i is the number of arcs incident to i, with loops counted twice. Note that G contains multiple arcs and multiple loops. Moreover, let m_{ij} be the number of the i, j-pair answers, and let m_i be the number of i-single answers, then $\delta_i = 2\, m_i + \sum_{j:j\neq i} m_{ij}$. Let m be the number of respondents, then $|E| = m$.

It can be conjectured that the graph G has a structure that can be revealed by clustering. That is, items $l_1, \ldots, l_n$ could be interpreted as specific expressions of latent variables, expressing preoccupation about some main and general issue, for example the Economy, the Security, the Social Welfare and so on. Therefore, from the operational point of view, items $l_1, \ldots, l_n$ can be clustered into homogenous groups using an appropriate clustering model. Here, we propose a clique partitioning model with a modularity objective function.

2.1 Modularity as independence in multiple response questions

Define $X_i \cup X_j$ as the event that a respondent elicited the l_i, l_j pair, with the notation $X_i \cup X_i$ denoting the event that the respondent elicited l_i as a single answer, then $\Pr[X_i \cup X_j]$ is the probability the one respondent elicited the l_i, l_j pair. Interpreted in the items graph, $\Pr[X_i \cup X_j]$, with possibly $i = j$, is the probability that an arc $(i, j) \in E$. Define $\Pr[X_i]$ as the probability that item l_i is one of the elicited item by a respondent and define $\Pr[X_i|X_j]$ as the conditional probability that a respondent has chosen l_i, giving that he or she elicited l_j. If $\Pr[X_i|X_j] \neq \Pr[X_i]$, then we will say that there is a *preferential pairing* between items l_i and l_j. Depending on the difference between the probabilities, we can asses whether the pair l_i, l_j has been elicited more or less frequently than expected.

Now we consider the following problem: What is the expected number r_{ij} of respondents that selected the l_i, l_j pair under the hypothesis that there is no preferential pairing? Or, interpreted in the items graph, what is the expected number of arcs between the pair i, j under the condition of independence?

Consider the survey graph $G = (V, E)$ and an auxiliary oriented graph $G' = (V, E')$, in which for every non-oriented edge $ij \in E$ there are two oriented arcs (i, j) and (j, i) in E' and, if the arc is a loop, for every non-oriented loop (i, i) there are two oriented loops, say $(i, i)^+$ and $(i, i)^-$. So, $|E'| = 2m$. Between G and G' there is the following connection: Let $\mathcal{A}$ be the event of selecting one edge at random from E, let $\mathcal{E}$ be the event of selecting one arc at random from E'. We have:

$$\Pr[\mathcal{A} = ij] = \Pr[\mathcal{E} = (j, i)] + \Pr[\mathcal{E} = (i, j)] = 2\Pr[\mathcal{E} = (i, j)] \tag{1}$$

Observe that the number of arcs of E' leaving a node i is δ_i, exactly as the number of entering arcs. If we draw at random on arc e from E', then we have (calculated as the ratio between favorable and possible cases):

$$\Pr[e \text{ leaves } i] = \frac{\delta_i}{2m} \tag{2}$$

$$\Pr[e \text{ enters } i] = \frac{\delta_i}{2m} \tag{3}$$

If there is no preferential pairing, e.g. independence, then:

$$\begin{aligned} \Pr[\mathcal{E} = (i, j)] &= \Pr[\text{ the arc leaves } i]\Pr[\text{ the arc enter } j|\text{ the arc leaves i}]] \\ &= \Pr[\text{ the arc leaves } i]\Pr[\text{ the arc enter } j] \\ &= \frac{\delta_i}{2m}\frac{\delta_j}{2m} = \frac{\delta_i \delta_j}{4m^2} \end{aligned}$$

Then we can state:

Theorem 2.1 : *Under the hypothesis of no preferential matching, the number r_{ij} of expected respondents of pair l_i, l_j is $r_{ij} = \frac{\delta_i \delta_j}{2m}$.*

Proof From equation (1), $\Pr[\mathcal{A} = ij] = \frac{\delta_i \delta_j}{2m^2}$ and then, as there are m edges in G, the expected number of edges is $r_{ij} = \frac{\delta_i \delta_j}{2m}$.

As a consequence of the Theorem, we can establish the difference $m_{ij} - r_{ij}$ as the measure of dissimilarity between the actual number of ij-respondents with the theoretical one,

under the assumption that there is no preferential pairing between items. Actually, this measure is at the core of the modularity index, see Newman and Girvan (2004), that is used for community detection in social networks. The only difference is that there is no loops neither multiple arcs in social networks, so that values r_{ij} are an approximation of a null hypothesis, while in our case they are an exact value. Modularity maximization coincides with the clique partition problem, see Agarwal and Kempe (2008), and it is revised in the following subsection.

3 The clique partitioning/modularity maximization model

The Clique Partitioning (CP) problem is one of the cornerstone of combinatorial optimization, see Grötschel and Wakabayashi (1989, 1990). It can be formulated as follows: Let $G = (V, E)$ be a complete graph. Let c_{ij} be the similarity measure between node i and node j, with c_{ij} being possibly positive, denoting similarity, and negative, denoting dissimilarity. Let $P = \{C_1, C_2, \ldots, C_q\}$ be a feasible partition of V and let $\sum_{(i,j)\in C_k} c_{ij}$ be the sum of the similarity and dissimilarity between vertices of group k. The CP problem consists of finding the node partition $P = \{C_1, C_2, \ldots, C_q\}$ to maximize the objective function $\sum_{C_k \in P} \sum_{(i,j)\in C_k} c_{ij}$. Its Integer Linear Programming (ILP) formulation is: Let x_{ij} be binary variables such that $x_{ij} = 1$ if node i and j are in the same cluster, 0 otherwise. Then the ILP formulation is:

$$z(G) = \max \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij} x_{ij} \tag{4}$$

$$x_{ij} + x_{jk} - x_{ik} \leq 1, \text{ for all } i, j, k \in V, i < j < k, \tag{5}$$

$$x_{ij} - x_{jk} + x_{ik} \leq 1, \text{ for all } i, j, k \in V, i < j < k; \tag{6}$$

$$-x_{ij} + x_{jk} + x_{ik} \leq 1, \text{ for all } i, j, k \in V, i < j < k; \tag{7}$$

$$x_{ij} \in \{0, 1\} \text{ for all } i, j \in V, i < j. \tag{8}$$

The objective function 4 selects clusters with high internal similarity and calculate the optimal modularity function $z(G)$ of the graph G. The triangle constraints 5, 6, 7 represent the property that, if i and j are in the same cluster and so are j and k, then i and k must also be in the same cluster. Finally, constraints 8 restrict variables to be binary.

From the problem formulation, note that it is important that similarities c_{ij} take positive and negative values, otherwise there is no incentive to discard negative arcs and the best partition would be $P = \{V\}$. Moreover, the optimal partition $P = \{C_1, C_2, \ldots, C_q\}$ can contain clusters with positive and negative internal arcs, as long as the sum of the positive similarities overpasses the sum of the negative dissimilarities. Finally, note that the number of clusters q of the partition is not fixed in advance, but it is a problem outcome.

To determine whether the items graph $G = (V, E)$ has a clustered structure, we compare the actual graph with an hypothetical items graph $G' = (V, E')$ having the property of no preferential pairing. Let m_{ij} be the number of arcs of E connecting two nodes i and j, which

corresponds to the number of respondents that actually answered the l_i, l_j pair, let $r_{ij} = \frac{\delta_i \delta_j}{2m}$ be the expected number of arcs if there is no matching preferences, then the difference $c_{ij} = m_{ij} - r_{ij}$ is an indicator of the discrepancy between a structured and an unstructured graph: the highest the difference, the most the actual graph G departs from the theoretical G' having no pairing patterns. Therefore, we can detect cluster of paired items finding groups of nodes with high internal cohesion. That is, the actual graph has a cluster structure $P = \{C_1, C_2, \ldots, C_q\}$ if the number of arcs connecting the nodes within a group C_k are more than what is expected, that is, if $\sum_{(i,j)\in C_k} c_{ij} > 0$, and this holds true for every $k = 1, \ldots, q$. Therefore, to determine the optimal node clustering the natural choice is to use the CP problem. That is, finding the node partition $P = \{C_1, C_2, \ldots, C_q\}$ such that the objective function $\sum_{C_k \in P} \sum_{(i,j)\in C_k} c_{ij}$ is maximized.

3.1 Inference with the CP model

The graph G' can be considered as the benchmark for a null hypothesis, see Zhang and Chen (2017). There it can be seen how to compare the actual modularity value $z(G)$ (from the objective function (4)) with the theoretical values of $z(G')$ of a graph G' in which there are no preferential pairings. Under the null hypothesis, $z(G')$ is characterized by a probability distribution that can be used to calculate the p-value of the test. Unfortunately the analytical distribution of $z(G')$ is unknown, but it can be simulated empirically by making a large number of artificial graphs G', characterized by no preferential pairing. Let $G_i, i = 1, \ldots, N$ be a i.i.d. sequence of simulated random graphs, $I\{\omega\}$ the indicator function of event ω, then the test p-value is approximated by the formula:

$$\hat{p}\text{-value} = \frac{1}{N} \sum_{i=1}^{N} I\{z(G) \leq z(G_i)\} \tag{9}$$

It remains to describe how G_i's are simulated, that is, what is the formal definition of the null hypothesis. We are using the configuration model, see Newman (2010): given a graph $G(V, E)$ with n nodes and degree sequence $\delta = (\delta_1, \ldots, \delta_n)$, the null model for the modularity measure is a random graph model having the same degree sequence but without preferential pairings. It can be simuleted by the following operations. Every edge $e = ij$ of the empirical graph $G = (V, E)$ is cut into two parts, say l_1 and l_2, with l_1 incident to i and l_2 incident to j, called *stubs*. Next, two different stubs are selected randomly and paired and the process repeated for a large number of iterations. The way in which G' is built implies that the degree sequence δ remains unvaried, but eventual preferential pairings are broken by the random reassignment of stubs. It is worth noting that the typical flaws of the procedure when applied to community detection, namely, the appearance of loops and multiple edges, see Cafieri et al. (2010), does not apply to items graph, as in the latter case multiple edges and loops are allowed.

To summarize, the p-value of the test is calculated through the following procedure:

- Step 1: Calculate $z(G)$, the value of the best CP of objective function 4
- Step 2: Repeat $i = 1, \ldots, N$ times:
  - Generate a random uniform G_i graph with fixed degree sequence (it can be done with the rewiring method described in Newman (2010)).
  - Calculate $z(G_i)$, the objective function 4 of CP applied to G_i.

- Use the sample $z(G_i)$, $i = 1, \ldots, N$, to determine the empiric distribution of $z(G)$ under the null hypothesis.
- Calculate p-value using formula (9).

In the next experiments, we have used the software GuRoBi, Gurobi Optimization, LLC (2022) to calculate $z(G)$ by solving the CP problem. Even though the CP problem is NP-complete, the instance size is small as the number of items is 16 at most. So, computational times are negligible even though they must repeated N times, less than 30 seconds when N = 1000, as in the following experiments.

4 An application: what is salient for public opinions?

The salience of a political issue is important to political analysis, as it affects both voters' behavior and governments' priorities. Moreover, salience could depend by different social characteristics, such as social classes, political position and so on. In the next application it can be seen how the methodology described so far can be applied to discover:

- whether citizens' concerns can be clustered into homogeneous classes, grouping together concerns having the same latent source;
- whether citizens worried for the same problem class can be characterized by any social feature, such as age, job, subjective social class or political position.

For illustrative purposes, we use the Eurobarometer ZA6928 surveyed available in the Gesis database European Commission (2018) and we compare three national audiences from Italy, Spain and West Germany. We consider the two selections question: *What do you think are the two most important issues facing your country at the moment?*. Here respondents can elicit up to two items among the list reported in the introduction, with the possibility of selecting no or just one item. Formally, the answer to the question is a variable X whose outcomes are every pair of issues, for example $x_i = Immigration,Crime$, or the outcome is a singleton, such as $x_i = Health$. The X domain is composed of as many answers as the item pairs, that is, the 16 items can be combined in more than 90 ways to obtain the faithful representation of the survey data. However, this variable X has never been analyzed in its full complexity. Rather, the data frame containing the survey responses X proceeds with a simplification. X is splitted into 16 variables/columns, say Y_i, $i = 1, \ldots, 16$, reporting the dichotomy *mentioned/not mentioned* for every single issue. After that, the statistical analysis is usually carried out using the simplified variables Y_i, see for example Rouet (2016); Bevan et al. (2016); Traber et al. (2022). and one of same Eurobarometer reports such as Brussels (2018).

One may argue that the variables Y_i are not the faithful translation of the original answers. Giving the possibility of two answers, *What are the most important issues* is translated into *Is [issue name] among the two most important issues?*, one may claim that passing from X to Y_i, $i = 1, \ldots, 16$ results with information lost and it could flaw the following statistic analysis. For example, in the original survey the *Immigration* issue can be mainly combined with *Unemployment* or, alternatively, with *Crime*. The two possibilities may lead to a different social interpretation of the choice: in the former case, one may claim that immigration is a concern because it worsen the job market, while in the latter case, because it worsen the public security. Next, suppose that two basic exploratory

techniques such as two-way tables and correlations are used to analyze X. One cannot apply two-way tables directly to X as it would imply a table with at least 90 lines, therefore one could use correlation on the simplified data Y_i to reveal whether some issues have been consistently mentioned in pairs. Unfortunately, this is not a viable methodology: due to the constraint on choosing at most two item, a mention (standing for 1 value) is always most often combined with a no mention, (standing for 0 value) and then the correlation is a negative number for most pairs Y_i, Y_j.

We will see how the items graph can be used to overcome the difficulty of the aforementioned procedures and then to retrieve the full information available in X. In Fig. 1a we report the histogram and the items graph of the three nations. From the histogram, it can be seen that the two most important issues aggregated by countries are Unemployment and Economy for the Spanish, Unemployment and Immigration for the Italians, Immigration and Education for the Germans. Among the less mentioned issues it is remarkable that Pensions is evenly mentioned in the three countries, Environment and Housing are an issue for the Germans only, Taxation for the Italians only, and a not trivial frequency of Spanish reported Other, perhaps referring to the Catalunya dispute that was concurrent to the survey. In the items graphs of Fig. 1b, c, d, data about the item combinations are reported in the form of the number of respondents that elicited an issues pair. The visualization is provided in such a way that nodes and arcs are larger if more respondents answered that issue or pair of issues and for graphical purposes arcs smaller than a given threshold are canceled. Regarding the Italian concerns, see Fig. 1b, it can be seen that most answers lie in the triangle Immigration-Unemployment-Economy, but Pensions seems a well connected issue too. Actually, it is questionable if the weights we are observing are significant, relying on the fact that they results from what we called preferential pairing, or they are just a visual effect resulting from the large number of choices. Note also that some loops are visible: for some respondents there is only one national problem. Regarding the other nations, in Fig. 1c the Spanish items graph reveals the connection between Unemployment and Economy, but with Health and Pensions as well, with which Unemployment may form some preferential pairing. Finally, in Fig. 1d, the German graph reveals the triangle formed by Immigration-Terrorism-Crime.

To check whether the most cited issues pair emerged as cases of preferential pairings, we applied modularity optimization. In the case of Italy, we found the optimal modularity value $z(G) = 0.051$ corresponding to the partition reported in Fig. 2a.

- Group 1: Unemployment, Economy, Immigration, Debt, Taxation, Pensions.
- Group 2: Crime, Terrorism, Prices.
- Group 3: Housing, Health, Education, Environment.

As can be seen, group 1 is mainly composed by Economic issues, group 2 by Security issues, group 3 by Welfare issues. Note that Immigration is an issue combined with Economic rather than Security issues and Pensions are combined with the Economic rather than with the Welfare. As described in Sect. 3.1, hypothesis testing is done by network rewiring and simulation, Items graphs $G_i, i = 1, \ldots, 1000$ are simulated under the null hypothesis of the configuration model and then optimal modularity $z(G_i^0)$ is calculated. Next, we compare the modularity $z(G)$ with the $z(G_i)$ histogram to calculate the experimental p-value. For the Italian case, in Fig. 2d we reported the histogram of the simulated values $z(G_i)$ where it can be seen that the null hypothesis modularity ranges from 0.015 to 0.040. Indeed, given the empiric modularity $z(G) = 0.051$, the experimental p-value is 0, rejecting the hypothesis that the actual items graph resulted from the configuration model.

To corroborate this claim, note that the histogram is reminiscent of a bell curve approximately normal from which we can calculate the test z-score and we obtain the value of 7.78, way larger than any typical hypothesis testing threshold.

The same analysis is repeated for Spain and Germany. As can be seen in Fig. 2b, Spanish concerns are divided in two main groups:

- Group 1, Security issues: Crime, Prices, Taxation, Terrorism, Immigration, Environment.
- Group 2, Economic and Welfare issues: Economy, Unemployment, Housing, Debt, Health, Education, Pensions.

Note that Economic issues are merged with the Welfare, while Immigration is included among the Security issues. The empiric modularity is $z(G) = 0.072$, while the H_0-modularity ranges from 0.015 to 0.030, see the histogram (2b), the empirical p-value is 0.0 and the test z-score is 15.59. Therefore we can reject the null hypothesis of the configuration model.

The German concerns are divided into three main groups:

- Group 1, Security issues: Crime, Unemployment, Terrorism, Immigration.
- Group 2, Economic issues: Economy, Prices, Taxation, Debt.
- Group 3, Welfare issues: Housing, Health, Education, Pensions, Environment.

Note that Immigration and Unemployment are among the Security issues, while the Welfare and the Economic issues are clearly specified by the expected terms. The empiric modularity is $z(G) = 0.072$, while the H_0-modularity ranges from 0.010 to 0.040, the empirical p-value is 0.0 and the z-score is 16.11. Therefore we can reject the null hypothesis of the configuration model.

In the next analysis, we will show how issue clustering can be applied as a technique of dimensionality reduction, projecting the whole list of issues into two or three classes. Then, we can analyze if respondents whose concerns are within one of the classes can be described in terms of some social characteristics, such as age, job, and so on. The features that we will consider are:

- Age: We have used the Eurobarometer recoded age into 4 classes: 15-24, 25-39, 40-54, 55 years old and more.
- Social class: Respondents can locate themselves on a 5-tired subjective social class level that we recoded into Lower and Middle class (Upper class respondents are aggregated as they are never more than a few units).
- Job position: We have used the Eurobarometer recoding into Employed, Self-Employed, and Not-Working.
- Political position: Respondents can locate themselves on a 10-tired political space that we recoded into Left (from 1 to 3), Middle (from 4 to 6), Right (from 7 to 10).

In Fig. 3a, b, c, we reported the histogram with the size of the clusters detected by modularity community detection. In the case of Italy, see Fig. 3a, it can be seen that the greatest cluster is composed by respondents whose concerns revolve around Economic issues, while people worried about Security or Welfare appear as residual. Conversely, there are many respondents that were unclassified, as they are represented by arcs belonging to two different groups. So, we simplify the analysis considering respondents as they belong to

just two groups: the economic group or the other, composed of the unclassified, the welfare, and the security. Next, we calculate all the two-ways table crossing the recoded concerns with social characteristics and we calculate their significance by the p-values. We found that the most important social characteristic that determines an economic concern is represented by the subjective social class, whose p-value is 0.017. In Fig. 3d conditional frequencies are reported and it can be seen how the lower class is more worried by Economic issues than the middle class, even though Economic issues are at the core of the overall Italian concerns.

In Fig. 3b and c histograms about the Spanish and the German clusters are reported. In the case of Spain, the greatest cluster is represented by the Economic/Welfare concerns, still the Security cluster is not negligible, as it contains 7.6% of the respondents. Therefore we continue the analysis keeping the two groups and the unclassified. The most significant social characteristic describing concerns is the political position, for which the p-value is 0.064. In Fig. 3e, conditional frequencies are reported: it can be seen that concerns about security increase from left to right wing voters, while concerns about economy or welfare decrease from left to right wing voters. In the case of Germany, the Economy is the smallest cluster having only 1.6% of respondents, therefore we aggregate this cluster to the unclassified to remain with Security and Welfare clusters only. The most significant social characteristic describing concerns is political position, for which the p-value is $1.2(10)^{-9}$. In Fig. 3f conditional frequencies are reported and it can be seen that Security concern increases from left to right wing voters, while Welfare concern decreases from left to right wing voters, as was the case of Spain.

In conclusion, the exercise shows how information of the items graph can be retrieved and used to determine what are the issues that are at the core of national public opinion's concern. We have seen that clusters can be interpreted as expressions of broad and general latent variablesthat can be used to reduce data dimensionality. From the exponential possible items pairs we have remained with a parsimonious data representation composed of two or three items classes to which standard statistical analysis can be effectively applied. In conclusion, we think that the items graphs is a convenient model to represent data coming from multiple issues questions and a useful tool to complement or improve the statistical techniques.

5 Comparison with alternative methods of analysis

In the previous section we have shown that data from multi-items questions can be interpreted as networks so that all the methodologies developed for network models can be readily applied to survey analysis too. Most peculiarly, among all the methodologies available, we focused on modularity optimization. Using modularity items can be clustered into homogeneous groups and in this way a considerable dimensional reduction can be obtained: from all the pairs of 14 items to just two or three groups. We have shown the viability of the method simulating a research workflow composed of various steps:

- Step 1: Cluster items into homogeneous groups using modularity optimization;
- Step 2: Verify their statistical significance;
- Step 3: Interpret groups as expression of latent or more general meanings;
- Step 4: Use clusters to reduce data dimension;
- Step 5: Use reduced data to discover association between variables.

It can be seen that the core of the network approach is represented by the Steps from 1 to 4, while Step 5 is contingent to other research purposes. In our example, we have used two-way tables to detect whether there were significant associations between issues and social features, but other techniques such as ordinary or generalized regression could be more appropriate in other applications.

The network modeling of multi-items questions is worth of investigation as it allows the application of many techniques developed for network analysis. However, it could be questioned whether other specialized and more standard statistical techniques could be applied as well, more precisely to the steps from 1 to 4, and whether they are as effective as network clustering to discover the hidden structure of the answers. So, we consider Step 1 and Step 4 of the workflow above, consisting in clustering and data reduction, and we will replace modularity optimization with two alternative standard techniques, namely, the *k*-means and the principal component analysis. We will see that the results of the two standard methods are less consistent, less interpretable and less reliable than modularity optimization.

As reported in the previous section, multi-item responses are coded in data frames as binary variables Y_i, $i = 1, \ldots, n$, with n the number of items, using the dichotomy *mentioned/not mentioned*. Instead of interpreting them as network data, they could be used as they are coded and without any transformation. In Step 1 of the research workflow, clustering items can be obtained by methods other than modularity optimization, for example using the *k*-means. The *k*-means algorithm takes the number of classes k as an input and then calculates the optimal data partition in k groups. Next, after running the *k*-means with different values of k, the correct value of k and therefore the optimal clustering are determined with some rule-of-thumb. The *k*-means can be applied to survey data using the following input. Data Y_i are reported in the matrix $A = [a_{ij}; i = 1, \ldots, m; i = 1, \ldots, n]$ with m the number of respondents, then $a_{ij} = 1$ if respondent i elicited issue j, $a_{ij} = 0$ otherwise. The distance between item i and item j is $d_{ij} = \sum_{k=1}^{m}(a_{ki} - a_{kj})^2$ and can be readily used by the *k*-means algorithm.

When the *k*-means is run on the the same data sets of the previous section, completely different clusters are obtained. Clusters are reported in Table 1. It can be seen that clusters are very unbalanced as they are composed of one or two items, and then all the remaining items in another cluster. For example, considering Italy, for $k = 2$, one cluster is composed

Table 1 Items clusters detected by *k*-means for varying values *k*

	Clusters		
k	Italy	Spain	Germany
2	1: Immigration, Unemployment	1: Unemployment, Economy	1: Immigration
	2: All the others	2: All the others	2: All others
3	1: Immigration	1: Unemployment	1: Immigration
	2: Unemployment	2: Economy	2: Terrorism, Crime
	3: All the others	3: All the others	3: All the others
4	1: Immigration	1: Unemployment	1: Immigration
	2: Unemployment	2: Economy	2: Terrorism, Crime
	3: Economy	3: Terrorism	3: Education, Environment
	4: All the others	4: All the others	4: All the others

of the two Italian most cited issues, Immigration and Unemployment, separated from all the other issues forming the second cluster. For $k = 3$, Immigration and Unemployment are splitted into two separate clusters against a third cluster composed of all the other issues. Next, for $k = 4$ Immigration and Unemployment remain separatedt and a new cluster is composed of the Economy only. Similar patterns are found in the Spanish and German data in which the most cited issues form a class by themselves (or in pair). As a result, the application of the *k*-means does not seem to have a great value as it only reproduces the ordering of the most cited issues.

From the workflow described above, it can be seen that clustering has been used to reduce the variables to two or three aggregated classes in Step 4. Nevertheless, another standard way to obtain data reduction does not pass from clustering, but is obtained through principal component analysis (PCA). In PCA, variables are grouped together by the weights of the linear functions, namely the principal components, that best summarize the covariance matrix. Indeed, the approach can be applied to multi-item questions as well and the PCA could reveal the same items structure found by modularity maximization. As we will see, this is not the case: the PCA cannot group together items, but rather, and similarly to the *k*-means algorithm, PCA recovers just the items that were selected most without providing additional information on the data structure.

We applied PCA to the national data sets of our previous experiments represented by the matrix A above. From a qualitative point of view it can be seen in Fig. 4 that PCA cannot reduce the total variability significantly. In that figure, the explained cumulative variance is reported as the number of components increases. As can be seen, using the first component only reduces the total variability to a small extent, as the first component explains less than 13% of the total variability for all the three nations considered. Moreover, to reach the threshold of explaining at least 50% of the total variance, up to five components must be considered by the analysis. Next, from a quantitative point of view, PCA cannot group together items in a significant way. In Tables 2, 3 and 4, the linear weights of the first three components are reported for the three nations considered. To interpret the weights, we impose a threshold of 0.45 to report significant weights (that are reported in bold figures), the other figures (less than 0.45)are reported as 0.0. In social sciences and

Table 2 Principal Components Analysis, Italy: the first three components

	PC1	PC2	PC3
Crime	0.000	0.000	0.000
Economy	0.000	0.000	**0.814**
Prices	0.000	0.000	0.000
Taxation	0.000	0.000	0.000
Unemployment	**−0.882**	0.000	0.000
Terrorism	0.000	0.000	0.000
Housing	0.000	0.000	0.000
Debt	0.000	0.000	0.000
Immigration	0.000	**−0.793**	0.000
Health	0.000	0.000	0.000
Education	0.000	0.000	0.000
Pensions	0.000	0.000	0.000
Environment	0.000	0.000	0.000
Other	0.000	0.000	0.000

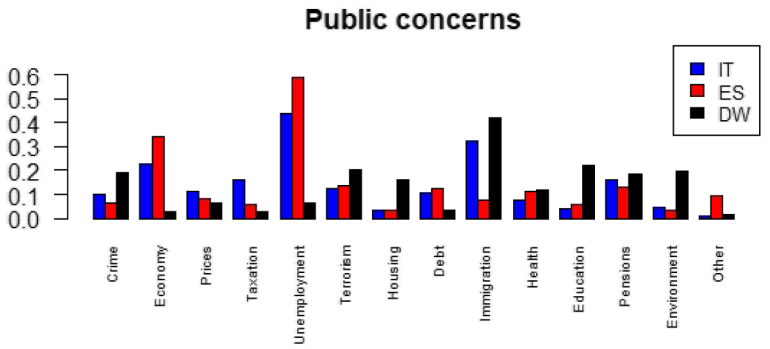
Table 3 Principal Components Analysis, Spain: the first three components

	PC1	PC2	PC3
Crime	0.000	0.000	0.000
Economy	0.000	**0.933**	0.000
Prices	0.000	0.000	0.000
Taxation	0.000	0.000	0.000
Unemployment	**−0.911**	0.000	0.000
Terrorism	0.000	0.000	**0.756**
Housing	0.000	0.000	0.000
Debt	0.000	0.000	0.000
Immigration	0.000	0.000	0.000
Health	0.000	0.000	0.000
Education	0.000	0.000	0.000
Pensions	0.000	0.000	0.000
Environment	0.000	0.000	0.000
Other	0.000	0.000	0.000

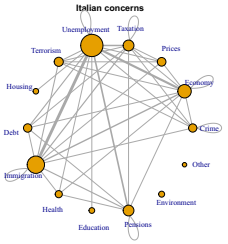
Table 4 Principal Components Analysis, Germany: the first three components

	PC1	PC2	PC3
Crime	0.000	0.000	0.000
Economy	0.000	0.000	0.000
Prices	0.000	0.000	0.000
Taxation	0.000	0.000	0.000
Unemployment	0.000	0.000	0.000
Terrorism	0.000	**0.573**	0.000
Housing	0.000	0.000	0.000
Government debt	0.000	0.000	0.000
Immigration	**0.837**	0.000	0.000
Health	0.000	0.000	0.000
Education	0.000	**−0.543**	**−0.551**
Pensions	0.000	0.000	**0.676**
Environment	0.000	0.000	0.000
Other	0.000	0.000	0.000

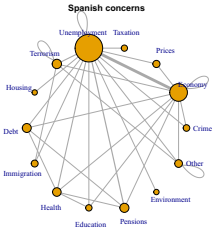
when applied to survey data this a standard method of composing an interpretable index out of the component weights, see for an example Akkerman et al. (2014). Regarding Italy, it can be seen that the first three components put an heavy weight on one item only, this item being Unemployment, Immigration, and the Economy for each component. More precisely, the first component separates those who answered *Unemployment* from the rest of the data, the second component those who answered *Immigration*, the third those who answered *Economy*. Looking at the preoccupations histogram of Fig. 1b, it can be seen that these are exactly the items that were selected most by the Italian respondents and therefore the three components are repeating what was already known by the descriptive statistics. Regarding Spain, the first three components selected one item only and they are Unemployment, Economy and Terrorism. Again, from Fig. 1c it can be seen that components are just repeating the three most selected items by the public opinion. Finally, regarding



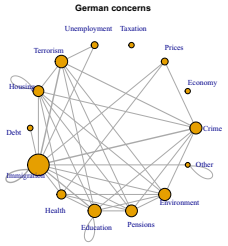
(a) National histogram.



(b) Italian issues.

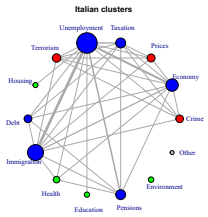


(c) Spanish issues.

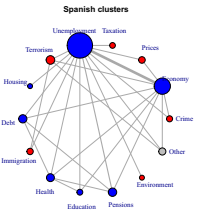


(d) German issues.

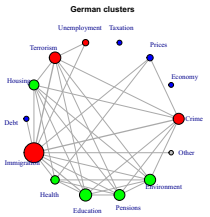
Fig. 1 Issues histogram and national graphs



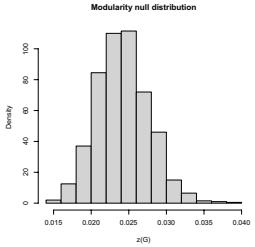
(a) Italian issue clusters.



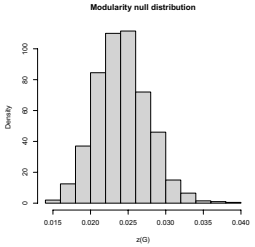
(b) Spanish issue clusters.



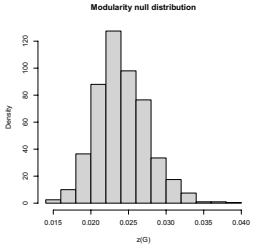
(c) German issue clusters.



(d) Italian null distribution.

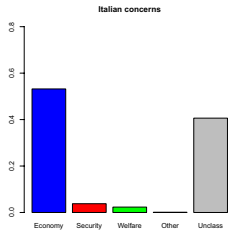


(e) Spanish null distribution.

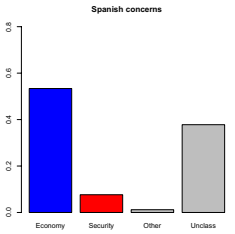


(f) German null distribution.

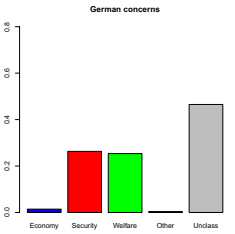
Fig. 2 Issues clusters and relative modularity null distributions



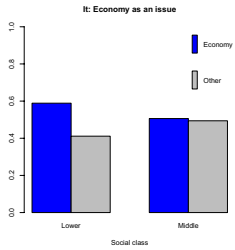
(a) Italian cluster histogram.



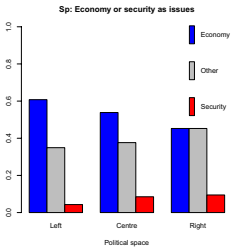
(b) Spanish cluster histogram.



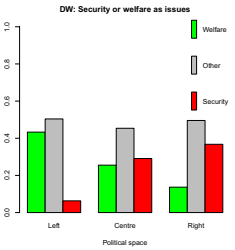
(c) German cluster histogram.



(d) Italian two-ways table.



(e) Spanish two-ways table.



(f) German two-ways table.

Fig. 3 Issues histograms and two-ways tables

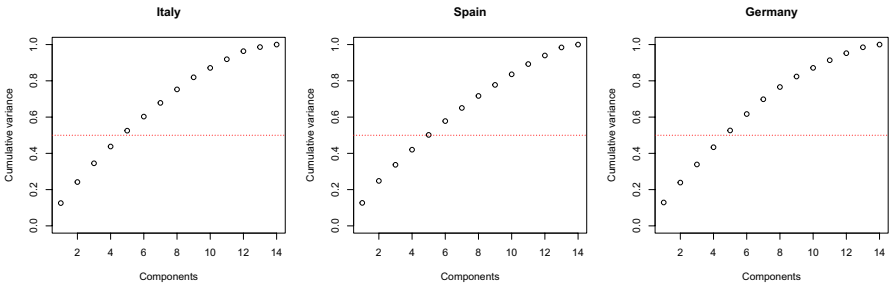


Fig. 4 Cumulative variance explained by principal components

Germany, the first component has a single issue, Immigration, while the second and the third component selected two items. The second component selects who answered Terrorism but not Education (see the opposite sign of the weights), and the third selects who answered Pensions but not Education. While Immigration is the most cited item among the German issues, see Fig. 1d, it is difficult to understand the meaning and relevancy of the other two components and in any case those item groups does not appear peculiarly meaningful.

6 Conclusion and future research

In this contribution, we provided a network model to represent survey questions with multiple selections and we showed how to apply community detection to cluster the question items. Community detection is only one of the many techniques that are used in network analysis, so once that (hopefully) we have demonstrated the utility of the items graph other network technique may be applied to it as well, such as centrality measures, core-periphery decomposition, and so on. One peculiar mention could be deserved to the so-called overlapping community detection model, in which one node can belong to more than one community, see Xie et al. (2013) for a survey, and Benati et al. (2022)for some appropriate mathematical model. In our application, overlapping communities could be useful to reduce the number of respondents that were classified as *other* for the plain fact that that their choices belong to two different groups. If communities/groups can overlap, then more arcs can belong to one of the selected groups and respondents could be better classified. A second possibility is suggested from the Eurobarometer application, as actually there are *three* questions about what are the most important issues of concern. They are most important issues of concern for the *nation*, *personally*, and for the *European union*, that can be modeled as three items graphs. The three graphs can be combined and connected as they contain the same item list. One respondent is represented by three arcs, one for each graph, so what is obtained is a multilayer graph to which specific networks techniques can be applied, see Mucha et al. (2010); Dickison et al. (2016). Finally, there are surveys with questions in which respondents can elicit more than two items. For example in the same Eurobarometer there is a question about what makes a sense of community between European citizens. There, the items are: History, Religion, Values, Geography and many others, among which respondents can elicit up to *three* items. In this case, the representing answers by arcs is not sufficient as an arc can connect only two items. However, an arc can be readily extended to be an hyper-arc, that is, an arc connecting more than two nodes in the so-called hypergraphs, for which modularity optimization can be applied as well, see Kaminski et al. (2019); Kumar et al. (2020).

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Stefano Benati and Justo Puerto. The first draft of the manuscript was written by Stefano Benati and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by Università degli Studi di Trento within the CRUI-CARE Agreement. This work was supported by the financial aid of NetMeetData: Ayudas Fundacion BBVA a equipos de investigacion cientifica 2019.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Agarwal, G., Kempe, D.: Modularity-maximizing graph communities via mathematical programming. Europ. Phys. J. B **66**(3), 409–418 (2008)

Akkerman, A., Mudde, C., Zaslove, A.: How populist are the people? Measuring populist attitudes in voters. Comparat. Polit. Stud. **47**(9), 1324–1353 (2014)

Benati, S., Puerto, J., Rodriguez-Chia, A.M., Temprano, F.: A mathematical programming approach to overlapping community detection. Phys. A Statist. Mech. Appl. **602**, 127628 (2022)

Bevan, S., Jennings, W., Wlezien, C.: An analysis of the public's personal, national and eu issue priorities. J. Europ. Publ. Policy **23**(6), 871–887 (2016)

Brussels, K.P.: Standard eurobarometer 89, first results, public opinions in the european union. Technical report, European Commission, Directorate-General for Communication, Brussels (2018)

Cafieri, S., Hansen, P., Liberti, L.: Loops and multiple edges in modularity maximization of networks. Phys. Rev. E **81**, 046102 (2010)

Das, K., Samanta, S., Pal, M.: Study on centrality measures in social networks: a survey. Soc. Netw. Anal. Min. **8**, 1–11 (2018)

Dickison, M., Magnani, M., Rossi, L.: Mambridge University Press. Cambridge University Press (2016)

European Commission, B.: Eurobarometer 88.3 (2017). GESIS Data Archive, Cologne. ZA6928 Data file Version 1.0.0, https://doi.org/10.4232/1.13007 (2018)

Fortunato, S., Hric, D.: Community detection in networks: a user guide. Phys. Rep. **659**, 1–44 (2016)

Grötschel, M., Wakabayashi, Y.: A cutting plane algorithm for a clustering problem. Math. Progr. **45**(1–3), 59–96 (1989)

Grötschel, M., Wakabayashi, Y.: Facets of the clique partitioning polytope. Math. Progr. **47**(1–3), 367–387 (1990)

Gurobi Optimization, LLC. Gurobi Optimizer Reference Manual. http://www.gurobi.com (2022)

Kaminski, B., Poulin, V., Pralat, P., Szufel, P., Theberge, F.: Clustering via hypergraph modularity. PLoS ONE **14**(11), e0224307 (2019)

Kumar, T., Vaidyanathan, S., Ananthapadmanabhan, H., Parthasarathy, S., Ravindran, B.: Hypergraph clustering by iteratively reweighted modularity maximization. Appl. Netw. Sci. **5**(1), 52 (2020)

Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.-P.: Community structure in time-dependent, multiscale, and multiplex networks. Science **328**(5980), 876–878 (2010)

Newman, M.: Networks: An Introduction. Oxford University Press, Oxford (2010)

Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Phys. Rev. E **69**(2), 026113 (2004)

Rouet, G.: European union: fears and hopes. Eastern J. Europ. Stud. **7**(1), 5–33 (2016)

Tang, W., Zhao, L., Liu, W., Yan, B.: Recent advance on detecting core-periphery structure: a survey. CCF Trans. Pervas. Comput. Interact. **1**, 175–189 (2019)

Traber, D., Hänni, M., Giger, N., Breunig, C.: Social status, political priorities and unequal representation. Europ. J. Polit. Res. **61**(2), 351–373 (2022)

Xie, J., Kelley, S., Szymanski, B.K.: Overlapping community detection in networks: The state-of-the-art and comparative study. ACM Comput. Surv. **45**(43), 1–35 (2013)

Zhang, J., Chen, Y.: A hypothesis testing framework for modularity based network community detection. Statist. Sin. **27**, 437–456 (2017)