Hypothesis testing in the clique/modularity partition problem

Stefano Benati[1]

[1]Department of Sociology and Social Research
School of International Studies
University of Trento

June 25, 2021

The Clique Partition Model

- We have $U = 1, \ldots, n$ statistic units.
- For all pair $i, j \in U \times U$ we have a (dis)similarity measure c_{ij}.
- c_{ij} is positive: the two units are similar.
- c_{ij} is negative: the two units are diverse.
- For a cluster/clique $C_k \subseteq U$, the quality of the clique is:

$$f(C_k) = \sum_{i,j \in C_k} c_{ij}$$

- and for a partition $(C_1, \ldots, C_m)$, the quality of the partition is:

$$f(C_1, \ldots, C_m) = \sum_{k=1}^{m} f(C_k)$$

The Clique Partition Problem

The Clique Partition Problem is:

$$\max_{C_1,\ldots,C_m} \left[\sum_{k=1}^{m} f(C_k) \right]$$

In the figure:

- The black arcs are positive (connecting similar units);
- The red arcs are negative (connecting dissimilar units).

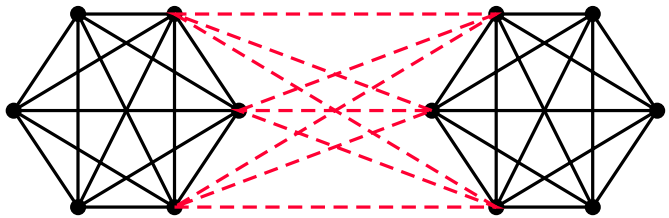


Figure: Clique Partition

Remark: In all relevant applications, the Clique Partition Problem is formulated on *complete* graphs!

Qualitative Survey Data

Application 1: Clustering qualitative data from surveys: A survey is composed of m questions. Take question k and compare units i and j, let x_{ik} the answer of i to question k:

$$d_{ij,k} = \begin{cases} 0 & \text{if } x_{ik} = x_{jk} \\ 1 & \text{otherwise} \end{cases}$$

Clique partition similarity matrix (Regnier, 1965):

$$c_{ij} = 2 * m - \sum_{k=1}^{m} d_{ij,k}.$$

This motivated the first formulations of clique as a combinatorial problem, by Grotchel, Wakabayashi (1989, 1990, Math Prog) and Johnson, Mehrotra, Nemhauser (1993, Math Prog).

Community detection in networks

Application 2: Community detection

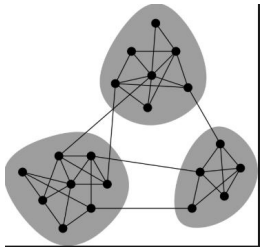


Figure: Network communities

- Let V the vertex set of a graph, $i \in V$ represents individuals, companies, nations, and so on.
- There is an arc $(i, j) \in E$ if i and j are connected, 0 otherwise;

For notation purposes, let m be the number of arcs, let d_i the degree of vertex i, $e_{ij} = 1$ if $(i, j) \in E$, 0 otherwise.
The Modularity maximization problem (Newmann, 2006, PNAS) is formulated as a maximum clique problem (Agarwal and Kemp, 2008, Eu. Physical Journal B) with input data:

$$c_{ij} = e_{ij} - \underbrace{\frac{d_i d_j}{2 * m}}_{\text{Null Hypothesis of random connections}}$$

Clique Partitions and k-means

The main competitor of clique partitions is the k-means. If you look at Google, for one mention of clique partition there are 1000 mentions of k-means. But we could advocate the use of clique partitions because:

- It is an **ILP** versus Continuous Non-Convex Problem.
- It works equally well both with both **cardinal and qualitative** data.
- The number of clusters k is an **output**, it does not need to be fixed in advance.
- It detects **outliers**, units that are automatically not assigned to any group.
- It can be used for both **clustering and community** detection.
- It can include the **null-hypothesis**, in the form of expected values of modularity.

Regarding the last points, this property has been recognized by Benati, Puerto, ERC-Sinergy Grant (2017), but also published by Zhang and Cheng, 2017, Statistica Sinica.

Integer Linear Programming formulation

Clique/Modularity partitioning can be formulated as Integer Linear Programming (Agarwal, Kempe, 2008, Europ. J. Physics, B):

$$x_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are in the same cluster} \\ 0 & \text{otherwise} \end{cases}$$

$$\max_{i,j>i} c_{ij} x_{ij}$$

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

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

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

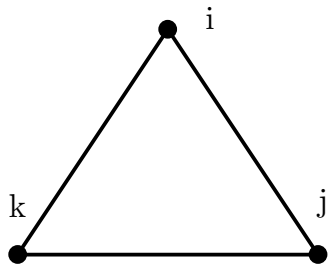


Figure: The Closing Triangle Constraint

Solution approaches to Maximum Modularity

- Matrix Spectral Decomposition, Newmann, 2006.
- Column Generation: Cafieri, Liberti, Hansen, Caporossi, Aloise, Perron, 2010, Physical Review,
- Hierarchical Aggregation based on Optimal Bipartition: Cafieri, Liberti, Hansen. 2011, Physical Review, 2011.
- Strengthening formulation with Cohesion Constraints: Cafieri, Costa, Hansen, 2013, IMA Journal of Complex Networks, 2014.

National problems and Public Opinions

The following question is taken from the Eurobarometer, Autumn 2017:

Question QA3A:

QA3a: What do you think are the two most important issues facing your country at the moment? (MAX. 2 ANSWERS)

- 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 (SPONTANEOUS)
- None (SPONTANEOUS)
- Don't know

European Problems: Single Issues

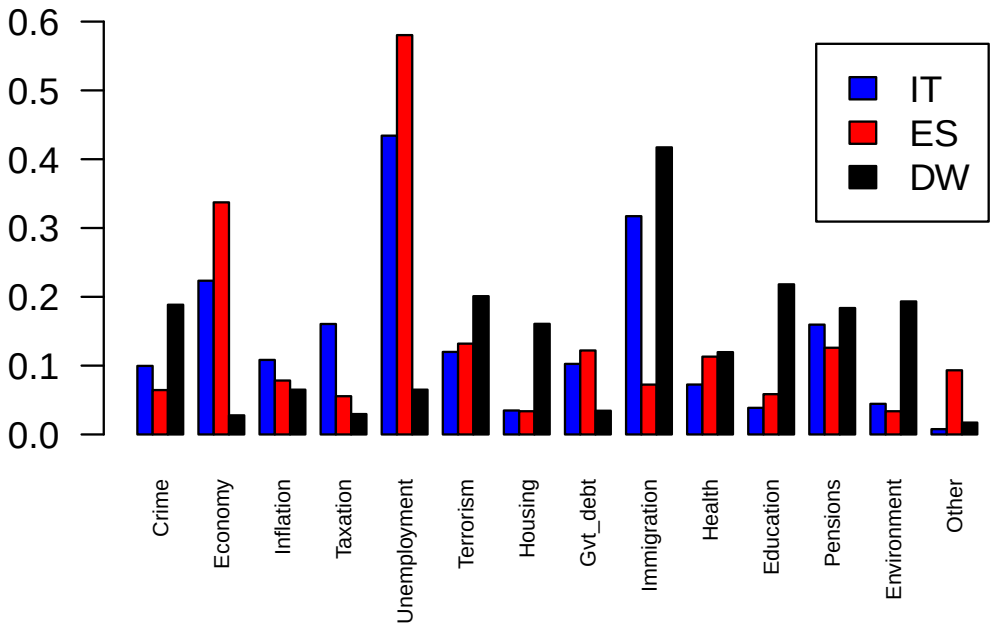


Figure: Country Problems in Italy, Spain, West Germany, in Autumn 2017

Remark: But respondents actually gave *pairs* of issues!

A too simplified framework

A respondent saying *"Immigration"*, *"Crime"* is providing an information that is different from *"Immigration"*, *"Unemployment"*. But, the analysis of this pairs is problematic:

- Respondents can elicit one or two items: weights on answer are not uniform.
- We cannot use correlation: it is negative for any pair (due to the cardinality constraint: only two item can be barred).
- We cannot use two ways tables (e.g. separating men from women), because there are at least $\binom{13}{2}$ possible answers of an individual.

Nevertheless, the scientific literature seldom realized this flaw and all answers like the one above are simplified in the dichotomy *mentioned/not mentioned*.

The Survey Graph

- Let $l_1, \ldots, l_n$ be labels assigned to answers of a multi-item question.
- Let $V = \{1, \ldots, n\}$ the vertex set 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.
- there is a loop $(i, i) \in E$ for every respondent that answered l_i as a singleton.

We call the resulting $G = (V, E)$ *Survey Graph*

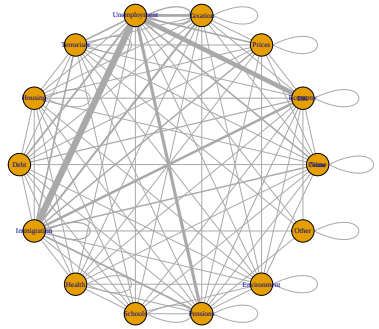


Figure: Country Problems in Italy, Autumn 2017

Remarks on Survey Graph

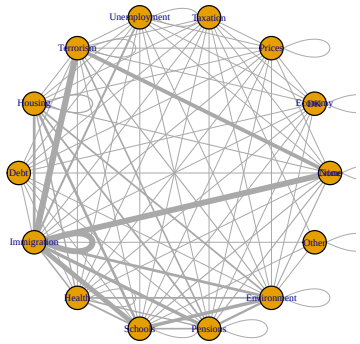


Figure: Country Problems in Germany, Autumn 2017

- There are multiple arcs between the same $i, j \in V$ pair.
- There are multiple loops.
- If there are m respondents, then $|E| = m$;
- Let d_i be the degree of $i \in V$ (approximately, it is the number of respondents that mentioned l_i), as singleton are counted twice).

The Null Hypothesis for Survey Graph

- Let $l_i, i = \ldots, n_l$ the possible answers (the question items).
- Let $\Pr[X_i]$ be the probability that item l_i is elicited by one respondent,
- Let $\Pr[X_i \cap X_j]$ is the probability the one respondent elicited the l_i, l_j pair.
- Interpreted in the survey graph, $\Pr[X_i \cap X_j]$ is the probability that arc $(i, j) \in E$.

The condition of independence

If $\Pr[X_i|X_j] \neq \Pr[X_i]$, then there is a preferential pairing of items l_i and l_j.

Question: What is the expected number r_{ij} of respondents that should have selected the l_i, l_j pair under the hypothesis that there is no preferential pairing?

The Probability of Selecting One Arc/Edge

Figure: The Survey Graph and the Auxiliary Graph

Note that for every non-oriented arc $ij \in E$ there are two oriented edges (i,j) and (j,i) in E'. So $|E'| = 2m$.

- Let $\mathcal{A}$ be the event of selecting one arc at random from E;
- Let $\mathcal{E}$ be the event of selecting one edge at random from E'

.
Then we have:

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

A condition of independence

If we draw at random on arc e from E', then we have (calculated as the ratio between favorable and possible cases):

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

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

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

So after some straightforward passages, under independence the expected number of arcs between i and j:

$$r_{ij} = \frac{d_i d_j}{2m}$$

The Graph Modularity

- Let m_{ij} be the actual number of arcs between $i, j \in V$;
- Let r_{ij} be the hypothetical number of arcs between $i, j \in V$ in condition of independence;
- The difference $c_{ij} = m_{ij} - r_{ij}$ is an indicator of the discrepancy between the empiric and the random graph.

Clique/Modularity Optimization in Survey Graphs

The objective function is:

$$z(G) = \frac{1}{2m} \max_{i,j} c_{ij} x_{ij}$$

subject to: The triangle inequalities of the clique partition.

Function $z(G)$ is called the *graph modularity*.

Hypotesis testing on modularity

Observe that the Survey Graph is characterized by self-loops and multiple arcs, therefore the expected number of arcs between nodes is an *exact* formula, not an approximation. So we can formulate a null hypothesis:

The Null Hypothesis for Survey Graphs

H_0 : The Survey Graph does not contain preferential pairings.

We use the modularity of the graph as the reference to distinguish random from structured graph, so that the p-value of the test can be calculated by simulation, using as the reference the *configuration model* for random graphs.

Calculation of the p-value

The *Configuration Model* for a random graph is a graph in which arcs between nodes are random, but the degree distribution is the same as the empiric graph. In other words, the problem marginal distribution remains the same.
Let $G_S = (V, E_S)$ the empiric survey graph. Let $G_R = (V, E_R)$ the random survey graph from the configuration model.
We are expecting that the graph modularity values are:

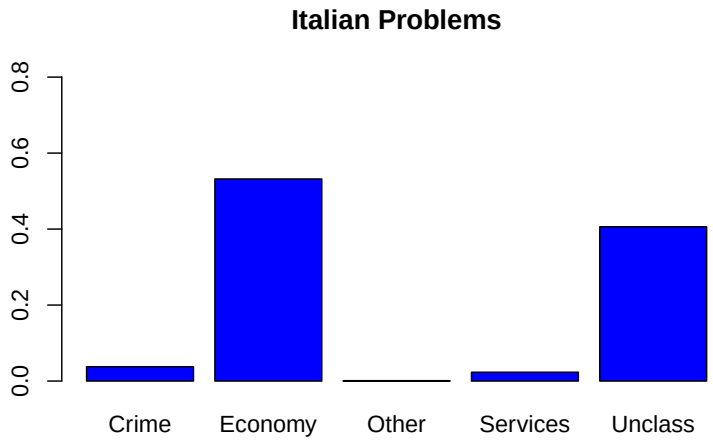
$$z(G_S) >> z(G_R)$$

- Step 1: Generate $i = 1, \ldots, n$ random graph from the same configuration model $G_{R,i}$.
- Step 2: Calculate modularity $z(G_{R,i}), i = 1, \ldots, n$.
- Step 3: p-value $= \dfrac{I_{\{z(G_{R,i}) > z(G_S)\}}}{n}$

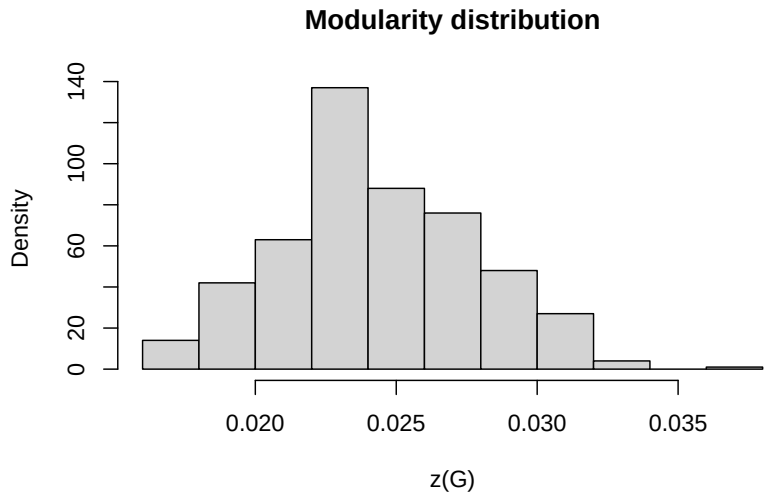
Italian National Problems

For the case of Italy, Autumn 2017, we found three clusters:

- **Group 1, Crime:** Crime, Prices, Terrorism
- **Group 2, Economy:** Economy, Taxation, Unemployment, Debt, Immigration, Pensions
- **Group 3, Services:** Housing, Health, Schools, Environment
- **Group 4, Other:** Other



Modularity Hypothesis Testing



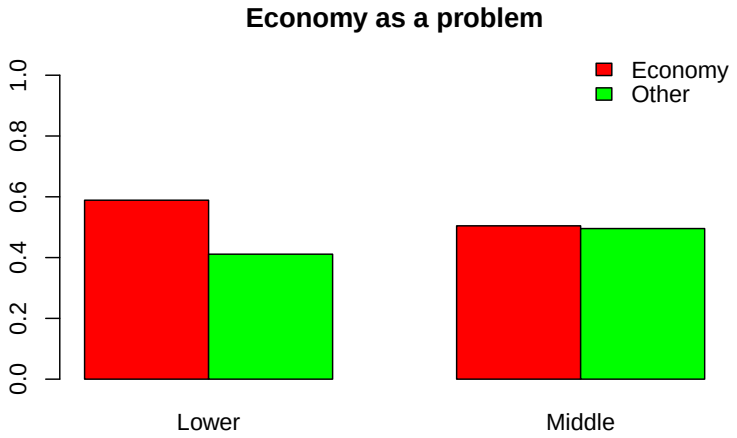
The empiric modularity value $z(G_S) = 0.051$ is outside the range of the null distribution. Therefore we can reject the hypothesis that the survey graph does not contain preferential pairings.

Bivariate Analysis

Respondents were recoded into two classes:

- Those whose answers are both in the Economic clusters;
- All other respondents.

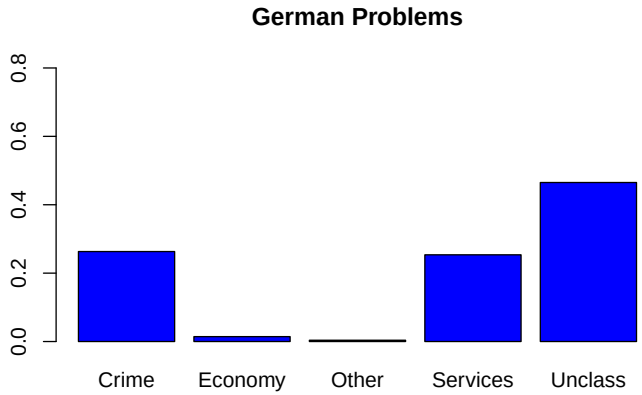
Next, I controlled for the individual features (sex, age, political opinions) that could identify respondents worried for the Economic problems. The most significant (using the Chi-square test) is social class.

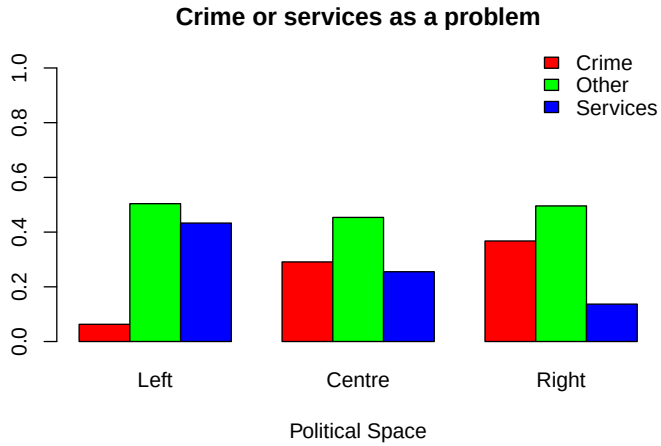


German Clustered Problems

For the case of Germany, we found three clusters:

- **Group 1, Crime:** Crime, Unemployment, Terrorism, Immigration.
- **Group 2, Economy:** Economy, Inflation, Taxation, Government debt.
- **Group 3, Services:** Housing, Health, Education, Pensions, Environment.

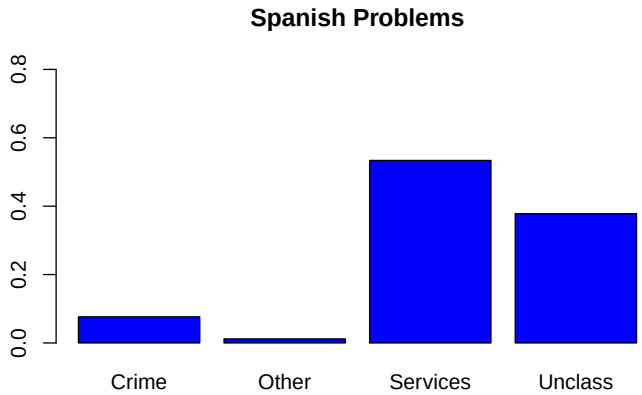


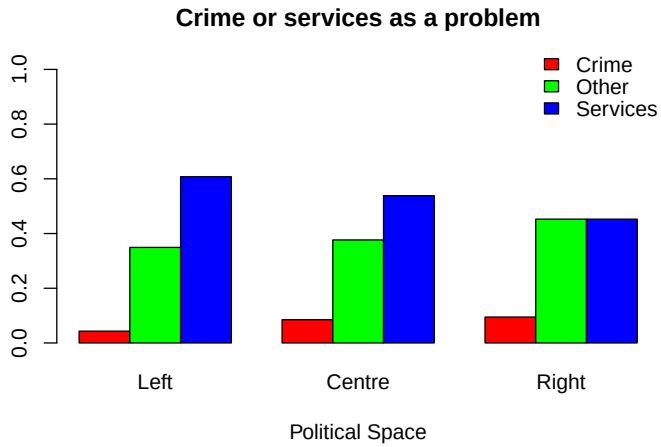


Spanish Clustered Fears

For the case of Spain, we found two clusters:

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





Summary

Now a brief summary:

- Modularity can be used to analyze survey questions with multi-items response.
- The outcome are problem clusters that could identify political/social cleavages.
- Clique partition models are the core of the methodology.
- The methodology opens up substantive analysis of some interest and originality.

Model extension

Some possible methodological development:

- The p-value is calculated after solving n (very large number) ILPs. Here, this optimization can be replaced by the continuous relaxations.
- Some questions are organized in groups:
  - What are the national problems?
  - What are the personal problems?
  - What are the European problems?

  Solutions: Multi-layer clique partition.
- Some survey questions let respondents elicit at most THREE or more items. Solution: Hypergraph modularity/clique clustering.
- The Null-Hypothesis model can be extended to other applications, for example voting in the United Nation Assembly.

Controversy in the United Nations

Voting in the United Nations General Assembly are: $1 = Yes$, $2 = Abstention$, $3 = No$, so the diffence between nations i and j on bill k is $\Delta_{ijk} = |x_{ik} - x_{jk}|$.

If there are n votes, the S-score (Agreement Index) between i and j, proposed by Signorino and Ritter, 1999, International Studies Quartely, is:

$$s_{ij} = 1 - \frac{\sum_{k=1}^{n} \Delta_{ijk}}{n}$$

Note that it is a value between -1 and 1, exactly as values that are used by clique partitioning when applied to survey data. But we can do more, assuming a null hypothesis of random voting versus preferential joint voting.

Voting in UNGA are often characterized by large numbers of YES votes.

- Suppose nation A: $\Pr[A = yes] = 5/6$;
- Suppose nation B: $\Pr[B = yes] = 2/3$.

The probability by which they both vote YES (under independence hypothesis) is 10/18. I So, S-scores can be modified to control for this null hypothesis.

$$s_{ij}^t = \text{Expected agreement index under random voting}$$

Next, we can use the difference:

$$c_{ij} = s_{ij} - s_{ij}^t$$

to find nation community clusters.

Remark: In this application we must calculate p-values, solving the Clique problem on a graph of 193 nations.