POLITECNICO DI TORINO

Master's Degree in Mathematical Engineering

Master's Degree Thesis

Introducing Fractionalization Index in Balance Theory

Supervisors

Prof. Claudio ALTAFINI

Prof. Giacomo COMO

Candidate

Marco RATTA

October 2022

Summary

In a world which is every day more interconnected and permeated with social relationships (both physical and virtual) modelling such interactions has become an interesting issue, in order to achieve a better comprehension of phenomena such as the development of communities.

From a mathematical perspective, the general theory of unsigned networks - where nodes are linked one to each others by arcs with positive weights, representing friendly and collaborative interaction between individuals can only in part describe real situations, and signed networks are commonly used in order to take into account also negative interactions (representing unfriendly relationships).

Referring to Cartwright and Harary's generalization of Heider's theory, the two theories of structural balance and weak balance have been developed in order to investigate the cause of conflicts in networks of individuals whose mutual relationships are characterizable in terms of friendship and hostility. In this regard structural balance theory postulates that the origin of such tensions is related to the presence of negative cycles (i.e. cycles with an odd number of negative edges), meaning that a network is exactly balanced if it has no negative cycles, while it is increasingly unbalanced the more the negative cycles are present in the network.

Relaxing the hypothesis allowing the network to have "weakly balanced cycles" (i.e. negative cycles with more than one negative link) weak balance theory instead generalizes the previous one postulating that a network is weakly balanced if it can be exactly partitioned into communities where every within-community link is positive and every between-communities link is negative, while it is increasingly unbalanced the more is the amount of unbalanced cycles.

As it is hard to expect that real networks are exactly balanced, particularly interesting is to quantify how far a network is from balance; in this regard, among the various methods introduced in literature, one of interest is the frustration index (also known as line index of imbalance) which represents the minimum amount of edges that must be reversed in order to achieve a balanced network. Despite the simple derivation (that refers to the Ising spin glass model), its computation turns out to be an NP hard problem (which does not allow it to be calculated for large networks), and heuristics have been proposed to estimate it in a feasible time.

For the case of weakly balanced fully connected networks, and supposing that the related partition of the network is known, we prove that the frustration index is related to the fractionalization - a dispersion index well known in literature and vastly used in economy and ecology - by an exact formula. We show that on empirical data the two indexes are well correlated in the vast majority of the cases. Two consequences follow: first knowing the network partition, we can interpret fractionalization as an easy to compute estimation of the frustration; second we can use frustration to estimate the number of expected communities in community detection algorithms.

The aim of this work is to provide a formal analysis of these concepts in the case of fully connected networks, Erdos Renyi networks, and quasi weakly balanced networks (i.e. allowing the network to present violations from the weakly balanced state). For all these cases the exact formula for the frustration index is obtained and shown numerically using artificial networks. Moreover for the fully connected case some examples of real networks belonging to economy, politics and etno-linguistic are considered in order to validate the formula. Lastly a novel approach for community detection is proposed, and compared with the state of the art both for unsigned and signed networks in the case of planted partition graphs.

Acknowledgements

Non nascondo che una parte di me credeva - e forse in parte sperava - che il giorno che mi sarei trovato a scrivere queste poche righe sarebbe stato anche il giorno del mio addio all'università e ora che so che non è così le mie emozioni sono un po' contrastanti.

Credevo che mi sarei sentito sollevato all'idea di lasciare il Politecnico, lo studio, le ansie da esame, gli appelli ravvicinati e tutto ciò che orbita attorno a un sistema a mio avviso ancora inadeguato che guarda troppo poco alle necessità degli studenti e allaloro salute fisica e mentale.

E in effetti è così, sono contento - ed è davvero la cosa più importante di questo giorno per me - di lasciarmi alle spalle questo capitolo della mia vita, ma non voglio concludere il mio percorso senza riconoscere ciò che di buono mi ha portato questa esperienza e ringraziare chi mi è stato accanto nei modi più diversi durante qeuesti anni.

Il mio primo pensiero va alla mia famiglia che sin dal momento della mia scelta mi ha messo nelle condizioni di seguire ciò che più mi appassionava sostenendomi economicamente e soprattutto emotivamente. Grazie a papà per avermi spronato a puntare in alto, grazie a mamma per aver cercato di sollevarmi dalle pressioni che ho sempre riposto su me stesso, grazie a Vale ed Emiliano per le dolci parole di incoraggiamento che mi hanno sempre riservato (e per l'esempio di amore reciproco che mi danno) e grazie a Samu per riuscire a regalare un po' di leggerezza ai miei rientri a casa.

Grazie ad Alice, mia compagna di vita, per essere stata e continuare pazientemente ad essere la mia ancora di salvezza da un mondo che mi esalta, mi stimola, ma che spesso mi investe e di cui non ho imparato a tenere il passo.

Grazie a chi ha condiviso con me anche solo una parte del percorso universitario: a Francesca, Mariachiara, Anna, Emanuele, Francesco, Giulia, Carlotta, Serena, Federico e tutti coloro che hanno reso questi anni un po' più leggeri.

Grazie al gruppo Arcobaleno, a Gianmarco per essermi stato sempre vicino, a Gennaro per aver portato un po' di sollievo anche nei momenti più tesi, a Edoardo e Italo; a Elisa per essere capace di strappare sempre un sorriso, a Fabrizio per la sua genuinità e tutti i ragazzi che sono passati dal collegio in questi ormai quattro

anni, facendomi apprezzare la diversità e la vita comunitaria.
Un enorme grazie a Federico per essere da sempre e costantemente amico e punto di riferimento, a Dordo, Ventu, Giulia, Meggy, Marco, Ari, Anna, Luca, Elia e tutti i miei amici di Bovezzo per continuare a esserci nonostante la mia distanza.
In ultimo ci tengo a ringraziare Claudio e la Linkoping University per la possibilità che mi è stata data con questo lavoro di tesi e il Politecnico di Torino per avermi dato la possibilità, aldilà di tutto, di costruire il mio futuro.
Questo percorso non sarebbe stato lo stesso senza di voi, vi devo tutto e vi ringrazio dal profondo.

Table of Contents

List of Figures IX

1 Effective Number of Parties 4

1.1 Generalities 4

1.2 Laakso-Taagepera ENP 5

2 Networks Balance 8

2.1 Generalities 8

2.2 Definitions 9

2.3 Measuring structural balance 12

2.4 Frustration 15

2.4.1 Ordinary formulation 16

2.4.2 Another formulation 17

3 Frustration in WB unweighted networks 21

3.1 Fully connected networks 21

3.1.1 Numerical results 23

3.2 Sparse networks 27

3.2.1 Numerical Results 29

3.3 Quasi WB fully connected networks 29

3.3.1 Numerical Results 33

4 Clusterization of signed networks 37

4.1 Modularity 37

4.2 Constant Potts Model 39

4.3 SQ2 model 41

4.3.1 General Idea 41

4.3.2 The algorithm 45

4.4 Numerical Results 47

4.4.1 Test 1 47

4.4.2 Test 2 48

4.4.3 Test 3 49
4.4.4 Test 4 50

A Politic graphs 53

B Etno-Linguistic graphs 55

C Smartphones market share graphs 61

Bibliography 66

List of Figures

2.1 Balanced triad (a) . . . 9
2.2 Balanced triad (b) . . . 9
2.3 Imbalanced triad (c) . . . 10
2.4 Imbalanced triad (d) . . . 10

3.1 Comparison between average frustration and average fractionalization for fully connected weakly balanced networks with increasing number of groups . . . 24
3.2 Average $\frac{E_{best}}{n/2}$ for weakly balanced fully connected networks with increasing number of groups . . . 24
3.3 Correlation between frustration and fractionalization index in weakly balanced fully connected networks for increasing number of groups . 24
3.4 Frustration index VS fractionalization index in weakly balanced fully connected networks . . . 24
3.5 Frustration index VS effective number of parties in weakly balanced fully connected networks . . . 25
3.6 Frustration index VS $\frac{E_{best}}{n/2}$ in weakly balanced fully connected networks 25
3.7 Frustration index VS Fractionalization index for Swedish and Albanian parliamentary networks. Over the inset: distance from 50 % in the best case, over all possible coalitions. . . . 26
3.8 Frustration index VS Fractionalization index for Tanzanian and Portuguese etno-linguistic networks, over the inset: distance from a 50 % partition . . . 27
3.9 Frustration index VS Fractionalization index for Italian and Japanese smartphones market share networks . . . 28
3.10 Comparison of average frustration and average fractionalization for sparse weakly balanced networks with increasing number of groups . 29
3.11 Average ξ ratio for networks with increasing number of violations σ 29
3.12 Frustration index VS $\frac{E_{best}}{n/2}$ for weakly balanced sparse networks . . . 30

3.13 Frustration index VS effective number of parties for weakly balanced sparse networks . . . 30
3.14 Correlation between frustration index and fractionalization index for sparse weakly balanced networks . . . 30
3.15 Frustration index VS fractionalization index for sparse weakly balanced networks . . . 30
3.16 Comparison of average frustration and average fractionalization for QWB networks with increasing number of groups . . . 34
3.17 Average ξ ratio for networks with increasing number of violations σ 34
3.18 Frustration index VS $\frac{E_{best}}{n/2}$ for quasi weakly balanced networks . . . 35
3.19 Frustration index VS effective number of parties for quasi weakly balanced networks . . . 35
3.20 Correlation between frustration index and fractionalization index in quasi weakly balanced networks for increasing number of parties . . 35
3.21 Frustration index VS fractionalization index for quasi weakly balanced networks . . . 35

4.1 Adjacency matrix example 2 . . . 44
4.2 Performance Test 1 . . . 48
4.3 Performance Test 2 . . . 49
4.4 Performance Test 3 . . . 50
4.5 Performance Test 4 . . . 51

A.1 Frustration VS time for European multiparty parliaments . . . 54

B.1 Frustration VS fractionalization for etno-linguistic dataset . . . 60

C.1 Frustration VS fractionalization for smartphones market share dataset 62

Network Basic Definitions

In this chapter we introduce some basic notation about networks which will be used in the following sections.

Definition 1. *Denote $V = v_1, v_2 \dots v_n$ a set of nodes, denote as "undirected signed network" the set $G = (V, E^-, E^+)$ where the set of couples $E^+ \subseteq V \times V$ contains all the positive links between nodes and similarly $E^- \subseteq V \times V$ contains all the negative links between nodes; such that $E^+ \cap E^- = \emptyset$.*

Definition 2. *Define the "positive adjacency matrix" as follows:*

$$A^+_{ij} = \begin{cases} 1 & \textit{if } (i,j) \in E^+ \\ 0 & \textit{otherwise} \end{cases} \tag{1}$$

Definition 3. *Define the "negative adjacency matrix" as follows:*

$$A^-_{ij} = \begin{cases} 1 & \textit{if } (i,j) \in E^- \\ 0 & \textit{otherwise} \end{cases} \tag{2}$$

Definition 4. *Define the "signed adjacency matrix" as $A = A^+ - A^-$, which can be expressed as:*

$$\mathcal{A}_{ij} = \begin{cases} 1 & \textit{if } (i,j) \in E^+ \\ -1 & \textit{if } (i,j) \in E^- \\ 0 & \textit{otherwise} \end{cases} \tag{3}$$

Definition 5. *Given a signed undirected graph $G = (V, E^-, E^+)$ we define the "Laplacian of G" as the matrix $\mathcal{L} = \Delta - A$, where $\Delta = diag\{\delta_1 \dots \delta_n\}$ and $\delta_i = \sum_{j=1}^{n} |a_{ij}|$*

Definition 6. *Given a signed undirected graph $G = (V, E^-, E^+)$ we define the "normalized Laplacian of G" as the matrix $L = I - \Delta^{-1}A$, where $\Delta = diag\{\delta_1 \dots \delta_n\}$ and $\delta_i = \sum_{j=1}^{n} |a_{ij}|$*

Definition 7. *We call the "positive neighbors" of the node v all the nodes which are connected with v by a positive edge, which means $\mathcal{N}_v^+ = \{u \mid (v, u) \in E^+\}$*

Definition 8. *We call the "negative neighbors" of the node v all the nodes which are connected with v by a negative edge, which means $\mathcal{N}_v^- = \{u \mid (v, u) \in E^-\}$*

Definition 9. *We call the "neighbors" of a node v all the nodes which are connected with v by an edge (whether positive or negative), which means $\mathcal{N}_v^+ = \mathcal{N}_v^+ \cup \mathcal{N}_v^-$*

Definition 10. *We define the "positive (negative) degree" of a node v as the number of nodes which are connected to v by a positive (negative) edge, which means $d_v^+ = |N_v^+|$, $d_v^- = |N_v^-|$.*

Definition 11. *We define the "total degree" of a node v as the total number of nodes linked to it by a positive or a negative edge, which means $d_v = |N_v^+| + |N_v^-|$*

Definition 12. *A sequence of nodes $v_1v_2v_3 \dots v_n$ is called a "n-length path" if $\forall k \in 1 \dots n$ it holds that $A_{k,k+1} \neq 0$.*

Definition 13. *Given a n-path $C = v_1v_2v_3 \dots v_{n+1}$, we call "sign" of the path the following:*

$$sgn(C) = \prod_{i=1}^{n} A_{v_i,v_{i+1}}$$

Definition 14. *We define a "n-cycle" a closed path, which means a sequence of nodes $v_1v_2v_3 \dots v_nv_1$ such that $\forall k \in 1 \dots n$ it holds that $A_{v_kv_{k+1}} \neq 0$ and $A_{1,n} \neq 0$.*

Definition 15. *Given a n-cycle $C = v_1v_2v_3 \dots v_1$, we call "sign" of the cycle the following:*

$$sgn(C) = \prod_{i=1}^{n} A_{v_i,v_{i+1}}$$

Definition 16. *Given a n-cycle $v_1v_2v_3 \dots v_1$, we call "chord" an edge connecting any two non consecutive nodes of the cycle.*

Theorem 1. *Let $C = v_1v_2v_3 \dots v_1$ be a cycle with a chord between nodes v_1 and v_r. Then let $C_1 = v_1v_2 \dots v_rv_1$ and $C_2 = v_rv_{r+1} \dots v_nv_r$ be the induced subcycles. Then $sgn(C) = sgn(C_1)sgn(C_2)$.*

Chapter 1

Effective Number of Parties

1.1 Generalities

The “effective number of parties ” (ENP) is a simple and intuitive concept used in population analysis to represent the fragmentation of a population.

Let’s consider a political scenario in which a parliament is divided into different parties each one with its own size (but the same concepts hold for every situation in which a population is formed by different groups competing with each others).

We can generally say that the effective number of parties is the number of groups of equal size corresponding to a given fragmentation of a population.

Intuitively it’s clear that this number is not strictly correlated to the true number of parties. In fact if we consider a political parliament composed by 4 parties with fractional sizes $(0.45, 0.35, 0.1, 0.1)$, it’s clear that even if namely the number of parties represented is 4, the number of important parties is 2, hence the situation in terms of balance of power is much more similar from the one where - for instance - the party vote-shares is $(0.6, 0.4)$ (which is effectively a 2-party system) while it’s sensibly different from another in which the party vote-shares is $(0.25, 0.25, 0.25, 0.25)$ (which on the contrary is a true multiparty system).

There should be a way to weight each party depending on its size, and to discount the smaller ones.

One way to operate intuitively is to choose a threshold of exclusion, discounting parties whose share is under that threshold; even though it’s unclear how to determine this value as of course it depends strongly on the party constellation itself (a party with a 5% for instance is almost irrelevant in a 3-party political scenario $(0.5, 0.45, 0.5)$, but it’s way more important in a 20-party equally shared political scenario).

These observations suggest us that we need a mathematical formula able to weight every party depending on their relative size.

1.2 Laakso-Taagepera ENP

The Laasko-Tasgepera effective number of parties, first introduced in [**1**] and further investigated in [**2**], is the most used formula for our scope and it relies on the Herfindal Hirshman index (HH), developed in 1940s and widely used in economics to measure the market concentration and monopoly detection [**3**].
One of the main features of the HH index is that it's connected to mean and standard deviation of the groups distribution; in fact suppose we have a market of size N divided in n groups $\tilde{s}_1, \dots, \tilde{s}_n$ such that $\sum_i^n \tilde{s}_i = N$ and let's call $s_1, \dots, s_n$ the relative shares which means $s_i = \frac{\tilde{s}}{N}$, then the Herfindal-Hirschman index is defined as follows:

$$HH = \frac{1}{n} + nV$$

where $\frac{1}{n}$ represents the mean and the variance V is defined as

$$V = \sigma^2 = \frac{\sum_1^n (s_i - (1/n))^2}{n}$$

where the summation runs over all the groups, i.e. every component gives is contribution to the index.
The formula can be rewritten as follows:

$$HH = \frac{1}{n} + \sum_1^n (s_i - (1/n))^2$$

and exploiting the fact that s_i are relative sizes, which means they sum to 1, it can be further reduced to:

$$\begin{aligned} HH &= \frac{1}{n} + \sum_1^n (s_i - (1/n))^2 \\ &= \frac{1}{n} + \sum_1^n \left(s_i^2 - \frac{2s_i}{n} + \frac{1}{n^2} \right) \\ &= \frac{1}{n} + \sum_1^n s_i^2 - \frac{2}{n} + \frac{1}{n} \\ &= \sum_1^n s_i^2 \end{aligned} \tag{1.1}$$

The Herfindal Hirshman index has 2 main properties which are worth to be mantioned:

1. The range of values it can assume is fixed and in particular the fact that it is bounded between 0 and 1 means that it can be used to compare different scenarios regardless of the number of groups involved.

2. It encodes the probability of two random selected individuals to belong to the same group.

Once defined the Herfindal-Hirshman index we can define the Laakso-Taagepera effective number of parties as:

$$N_{LT} = \frac{1}{\sum_1^n s_i^2} = \frac{1}{HH} \tag{1.2}$$

Another related index which encodes the same information of the previous ones is the so called "Fractionalization Index ".
It can be interpreted as an indicator of how fragmented is a multi-group population (or equivalently how well distributed are the resources among the groups, if we think of the groups as competing with each others for a good). The Fractionalization index is defined as:

$$F = 1 - HH = 1 - \frac{1}{N_{LT}} \tag{1.3}$$

Similarly to the Herfindal-Hirshman index, also the fractionalization index is bounded between 0 and 1, however in order to better understand the analysis in the next sections it has to be noticed that given a multi-group population composed by n groups, while the upper bound remains 1, the lower bound is $1-1/n$, consequence of the fact that the lower bound of the effective number of parties is n.

Chapter 2

Networks Balance

2.1 Generalities

In different types of networks, especially when they represent social interactions between individuals, it becomes necessary to allow edges to be either positive or negative in a way that we can both encode friendship (intended as a positive interaction between nodes) and animosity (intended as a negative interaction between nodes).
To be complete we could define several (even a continuum) degrees of friendship letting the weight corresponding to edges vary in the real field, but let us focus to the simplest case where edges are in one of these two values: positive (like) or negative (dislike); these networks are called "signed networks".
In this setup it has to be clear that the absence of an edge between nodes is not the same as having a negative interaction; differently from unsigned networks, where an interaction is forced a priori to be positive and the absence of interaction meant two nodes are as far as possible to be related, now the absence of edge represents two nodes that do not interact, which is a different situation with respect to two people disliking each others.

Let us consider now all the possible configurations of three edges in a triangle of possible nodes where + indicates a positive interaction and - indicates a negative interaction. It is reasonable to imagine that some of these configurations lead to some social problems if brought to real life; let us analyze all of them:

(a) ” +++ triangle ” , which means three people all liking each other.

(b) ” + - - triangle”, which means two of the three people are friends and they have a common enemy. It’s not as trivial as the situation (a) but it seems stable: the 3 people can be partitioned into 2 groups with two friends on one side and a common enemy on the other.

(c) " ++- triangle ", which represents one of the three people is liked by the other two but these two are enemies. Our intuition suggests that this situation could be unstable; in fact if we assume that two friends tend to behave in the same way it is clear that the person liked by the other two is in an ambiguous position as he's friend with two people hating each other. We can imagine that in this situation there's an implicit strength from the two people competing to convince the loved guy to hate the other one, hence to make the configuration become + - - which we have seen to be stable.

(d) The last triangle is the - - - and also this configuration is somewhat ambiguous. On one hand it consists of people who all dislike each others, which seems not to be strange; on the other hand it's reasonable to think that two of the three individuals may like to form an alliance against the third common enemy. Because of this we can assume that this configuration is unstable.

Starting from this qualitative perspective let's now point out some definitions and theorems which will lead us to define the concept of "structural balance".

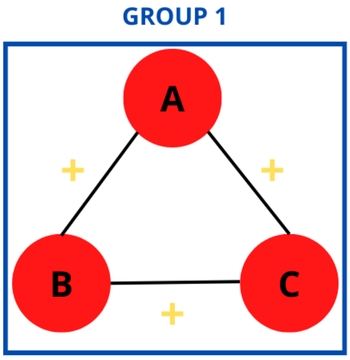


Figure 2.1: Balanced triad (a)

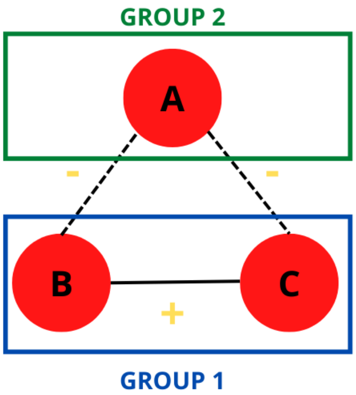


Figure 2.2: Balanced triad (b)

2.2 Definitions

Definition 17. *Given a signed undirected graph $G = (V, E^-, E^+)$, a triad of nodes (i,j,k) is considered "balanced" if it holds that*

$$A_{ij}A_{jk}A_{ik} = 1$$

Definition 18. *A complete signed graph $G = (V, E^-, E^+)$ is considered "balanced" if all triads (i, j, k) with $i \neq j \neq k$ are balanced.*

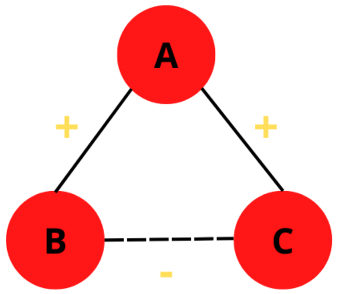


Figure 2.3: Imbalanced triad (c)

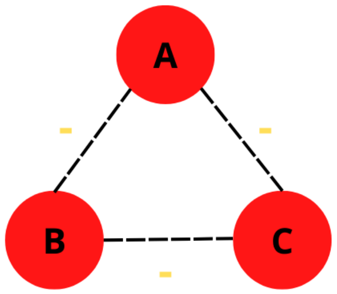


Figure 2.4: Imbalanced triad (d)

Theorem 2. *A complete signed graph $G = (V, E^-, E^+)$ is balanced if and only if the set of nodes V can be partitioned into two subsets V^+ and V^- such that $V^+ \cap V^- = \emptyset$ and $\forall e \in E^+$ it holds either $e \in V^+ \times V^+$ or $e \in V^- \times V^-$, while $\forall e \in E^-$ it holds that $e \in V^+ \times V^-$.*

Proof.
$\boxed{\rightarrow}$ Assume G is balanced. Consider some node $v \in V$ and the two sets $V_1 = v \cup \mathcal{N}^+(v)$ and $V_2 = V \setminus V_1$. Consider an edge $(u, w) \in V_2 \times V_2$; then $(u, v) \in E^-$ and $(u, w) \in E^+$ by definition of structural balance, this means all edges in V_2 are positive and similarly any adge $(u, w) \in V_1 \times V_1$ is positive. Hence we can partition V in two disjoint subsets V_1 and V_2.

$\boxed{\leftarrow}$ By hypothesis V can be partitioned in 2 disjoint subsets V_1 and V_2 such that every in-group edge is positive and every group-group edge is negative. Up to simmetry only 2 triads are possible: the first v_i, v_j, v_k all in V_1 and the second v_i, v_j, v_k such that $v_i \in V_1$, $v_j \in V_1$ and $V_k \in V_2$. Considering the edges between the nodes, the first is a $(+++)$ tryad, while the second is a $(+--)$ tryad. It's easy to see that both tryads have positive sign, which means they are both balanced. □

Definition 19. *Let $G = (V, E^-, E^+)$ be a signed graph, A the signed adjacency matrix and $C = v_1 v_2 v_3 \dots v_1$ a cycle. Then the cycle C is called balanced whenever $sgn(C) = 1$*

Definition 20. *A signed graph $G = (V, E^-, E^+)$ is considered "balanced" whenever all cycles $v_1 v_2 v_3 \dots v_1$ are balanced.*

Theorem 3. *A connected signed graph $G = (V, E^-, E^+)$ with adjacency matrix A is balanced if and only if the set of nodes V can be partitioned into two subsets V^+ and V^- such that $V^+ \cap V^- = \emptyset$ and $\forall e \in E^+$ it holds either $e \in V^+ \times V^+$ or $e \in V^- \times V^-$, while $\forall e \in E^-$ it holds that $e \in V^+ \times V^-$.*

Proof.
$\boxed{\leftarrow}$ Assume G in balanced. Then select any $v \in V$ and set V_1 as the set of nodes which can be reached by v through a positive path. Define $V_2 = V \setminus V_1$. Let $e = (u, v) \in E^-$ and suppose $e \in V_1 \times V_1$. By construction of V_1 then both u and w have a positive path to v, so that the path $u - v$ through v is also positive. But if (u, w) is negative, it would be contained in a negative cycle, which contraddicts the balance hypotesis; hence $e \notin V_1 \times V_1$. Similarly, suppose that $e \in V_2 \times V_2$. Then both the paths connecting $u - v$ and the one connecting $w - v$ are negative (otherwise u and w would be in V_1). The $u - w$ path through v is then positive since it's the product of two negative paths; and again - since $(u, w) \in E^-$ it contraddicts the balance hypothesis. As a consequence, all negative edges lie between V_1 and V_2. Then there is a positive $u - v$ path and a negative $w - v$ path, so that the $u - w$ path throw v is negative, which combined with the positive edge (u, w) leads to a negative cycle, contraddicting again the balance; hence positive edges lie either in $V_1 \times V_1$ or $V_2 \times V_2$. As a consequence we conclude that G is balanced.

$\boxed{\rightarrow}$ Suppose G can be partitioned in two disjoint sets V_1 and V_2 such that every positive edge lies either in $V_1 \times V_1$ or $V_2 \times V_2$. Let C be a cycle, there are two possible cases: if all the nodes of the cycle are in the same subset, them all edges within the cycle will be positive and the cycle will be balanced; if C has some node $u \in V_1$ and some node $v \in V_2$, then any $u - v$ path contains an odd number of negative edges (hence is negative), it follows that the cycle - which is product of negative paths - will be positive. As all the cycle belong to one of this two cases the theorem is proven.

□

Definition 21. *A cycle $C = v_1 v_2 v_3 \dots v_1$ is considered "weakly balance" if it does not contain a single negative edge, which means*

$$\sum_{i=1}^{n-1} A_{v_i,v_i+1} \neq n-1 \quad \textit{or equivalently} \quad \sum_{i=1}^{n-1} A^-_{v_i,v_i+1} \neq 1$$

Lemma 1. *Let $C = v_1 \dots v_k v_1$ be a cycle with a chord between the nodes v_1 and v_r in C. Then let $C_1 = v_1 \dots v_r v_1$ and $C_2 = v_1 v_k \dots v_r v_1$ be the induced subcycles. Then C is weakly balanced if C_1 and C_2 are weakly balanced.*

Proof. We denote by $m_1^- \neq 1$ and $m_2^- \neq 1$ the number of negative edges of respectively C_1 and C_2 and m^- the number of negative edges for C. Then if the link (v_1, v_r) is positive, it means that $m^- = m_1^- + m_2^- \neq 1$ which implies C is balanced. If the link (v_1, v_r) is negative then we have $m_1^- \geq 2$ and $m_2^- \geq 2$, which leads to $m^- = (m_1^- - 1) + (m_1^- - 1) \geq 2$, so that C is weakly balanced.

□

Theorem 4. *A signed network $G = (V, E^-, E^+)$ is considered "weakly structurally balanced" if all chordless cycles are weakly balanced.*

Proof.

$\boxed{\rightarrow}$ If G is weakly balanced all cycles are balanced; as every chordless cycle is a cycle the all chordless cycles are also balanced.
$\boxed{\leftarrow}$ Assume every chordless cycle is weakly balanced. Let's procede by induction on $|C|$. All chordless cycles are balanced so we have an inductive base for $|C| = 3$ (because all tryads are chordless). Assume now that every cycle such that $|C|< r$ is balanced, then consider a cycle of length r: if it contains a chord then it can be split into two cycles C_1 and C_2 such that $|C_1| < r$ and $|C_2| < r$, which are both balanced. Then by the previous lemma cycle C is also balanced as product of two balanced cycles.

□

Theorem 5. *Let $G = (V, E^-, E^+)$ be a connected signed graph. Then G is weakly structurally balanced if and only if it can be partitioned into subsets $V_1 \ldots V_n$ such that $\forall i \neq j$, $V_i \cap V_j = \emptyset$ and $\forall e \in E^+$ it holds $e \in V_i \times V_i$, while $\forall e \in E^-$ it holds that $e \in V_i \times V_j$ with $i \neq j$.*

Proof.
$\boxed{\rightarrow}$ Suppose G is weakly balanced. Let $G^+ = (V, E^+)$ be the positive part of the signed graph, and let the clusters be defined by the connected components of G^+. Any positive edge then clearly cannot fall between clusters, because different connected components cannot be connected through a positive link. Consider then some negative link $(u, v) \in E^-$. Suppose that u and v are both in the same cluster V_c. Then there exists a positive $u - v$ path because they are in the same component , thus yielding a cycle with exactly a single negative link, contradicting weak balance. Hence we proved that any negative link fall between clusters.
$\boxed{\leftarrow}$ Let's suppose that G is split into clusters as stated in the theorem. Any cycle completely contained within a cluster has only positive links. Consider a cycle through $u \in V_c$ and $v \in V_d$; then any path between u and v must contain at least a single negative link, so that any cycle must contain at least two negative links.

□

2.3 Measuring structural balance

In the previous section we introduced the main ideas behind structural balance, however the necessary conditions for balance to hold are often very strict and in practice it's not common to deal with network without cycles with odd number of negative edges (or with a single negative link).

Because of that several ways to measure the extent to which a graph is balanced have been introduced by researchers and resumed in [**6**]; we briefly discuss the main ones in this section and will analyze in detail one of them in the next one.

1. *Measures based on cycles*

   The simplest of such measures is the so called "degree of balance" suggested by Cartwright and Harary which is the fraction of balanced cycles of a network

   $$D(G) = \frac{\sum\limits_{k=3}^{n} O_k^+}{\sum\limits_{k=3}^{n} O_k}$$

   where O_k^+ is the total number of positive cycles in the network (close paths with even number of positive edges) and O_k is the total number of cycles of length k.

   Another measure strictly related to the previous mentioned is the "relative k-balance" which is again a cycle based measure where the two sums in the numerator and the denominator are restricted to a single term of fixed index k, that is the fraction of balanced cycles of fixed length k.

   $$D_k(G) = \frac{O_k^+}{O_k}$$

   A generalization of this two measures is the "weighted degree of balance" obtained by weighting cycles based on length using a non negative decreasing function $f(k)$ which can be chosen arbitrarily (for instance $1/k$, $1/k^n \dots$).

   $$D(G) = \frac{\sum\limits_{k=3}^{n} f(k) O_k^+}{\sum\limits_{k=3}^{n} f(k) O_k}$$

   Among all the "relative k-balance" indexes the most commonly used is the so called "triangle index" (with $k = 3$); as the network increases in fact the counting of cycles becomes hard to compute, however for the triangle index we have a closed function of the adjacency matrix A

   $$T(G) = D_3(G) = \frac{O_3^+}{O_3} = \frac{Tr(A^3) + Tr(|A|^3)}{2Tr(|A|^3)}$$

2. *Spectral measures*

   Beside checking cycles - which can easily become an unfeasible approach as the cardinality of the network grows - there are easier approaches to measure the distance from structural balance related to the eigenvalues.
   If in the previously discussed measures based on cycles we place more weight on shorter walks using a specific weight function as proposed in [] we come out with a form of the weighted ratio of balanced to total closed walks based on the eigenvalues of the adjacency matrix; in particular:

   $$W(G) = \frac{K(G)+1}{2}$$

   with

   $$K(G) = \frac{\sum_k \frac{Q_k^+ - Q_k^-}{k!}}{\sum_k \frac{Q_k^+ + Q_k^-}{k!}} = \frac{Tr(e^A)}{Tr(e^{|A|})}$$

   where for the calculation of $Tr(e^A)$ it can be used the fact that A is a symmetric matrix for undirected graphs and it holds that

   $$Tr(e^A) = \sum_i e^{\lambda_i}$$

   where λ_i are the eigenvalues of the adjacency matrix.
   Another eigenvalue based measure comes from spectral graph theory and comes from the fact that smallest eigenvalue of the signed Laplacian matrix - called. "algebraic conflict" and indicated with $\lambda(G)$ is equal to zero if and only if the graph is balanced and it increases the far the graph is from monotonicity.

3. *Measures based on frustration*

   A quite different measure which we only introduce here and we elaborate in the next chapter is the "frustration index", also referred as "line index of imbalance".
   A set E^* of edges is called "deletion-minimal" if deleting all edges in E^* results in a balanced graph but no proper subset of E^* has this property. Each edge in E^* lies on an unbalanced cycle. The graph resulted from deleting all edges in E^* is called "balanced transformation" of a signed graph and the frustration index equals the minimum cardinality among all deletion-minimal sets

$$L(G) = \min_{E^*} |E^*|$$

Similarly, in a setting where each vertex is given a black or white colour, calling "frustrated" all the positive vertices whose endpoints have different colours and all negative edges whose endpoints have same colour, the frustration index is therefore defined as the smallest number of frustrated edges over all possible 2-colourings of the nodes.

2.4 Frustration

Let's consider a network in which the nodes represent entities and edges represent the binary symmetric relationship between them (friend or enemy) and let's suppose moreover that either friendship or enmity holds between every couple of nodes.
This network can be described by a complete undirected signed graph $G = (V, E^-, E^+)$ where $E \in \{-1,1\}$ and $A \in \mathrm{Sym}(\mathbf{B}_2)$.
As suggested by spin glass theory and according to the above definitions, computing the global balance is equivalent to split the nodes set V into V^+ and V^- as to minimize the total inconsistencies among all possible splitting cuts.
To achieve this let's write a function which assigns each node a value of +1 (if it belongs to faction X) or -1 (if it belongs to faction Y).

$$\sigma : V \to \mathbf{B}_2 \quad \sigma(v_i) = s_i \in \{+1, -1\} \tag{2.1}$$

In this scenario there are two possible types of inconsistency between two nodes v_i and v_j (up to simmetry):

- The two nodes are mapped through σ to the same faction but they are enemies, which means $\sigma(v_i) = \sigma(v_j)$ and $A_{ij} = -1$
- The two nodes are mapped through σ to different factions but they are friends, which means $\sigma(v_i) \neq \sigma(v_j)$, and $A_{ij} = 1$.

Minimizing the total number of inconsistencies all over the possible partitions of the V set is equivalent to minimize the following energy functional

$$h(s) = \frac{1}{2} \sum_{(i,j)} (1 - A_{ij} s_i s_j)$$

In fact, as in both the above cases it holds that $A_{ij} s_i s_j = -1$ while otherwise $A_{ij} s_i s_j = +1$, only inconsistent edges will contribute to the summation.
The summation runs over all the couples of adjacent nodes, however in our case

the graph is fully connected hence the adjacency matrix is full and the energy functional can be written as

$$h(s) = \frac{n(n-1)}{2} - \frac{1}{2}\sum_{(i,j)} A_{ij} s_i s_j$$

Recalling that the number of edges m in a fully connected graph is equal to all the possible pairs of nodes, which is

$$m = \binom{n}{2} = \frac{n(n-1)}{2}$$

then the energy functional takes the form

$$h(s) = m - \frac{1}{2}\sum_{(i,j)} A_{ij} s_i s_j$$

2.4.1 Ordinary formulation

Definition 22. *A diagonal matrix S is called "signature matrix" if $S_{ii} \in \{+1, -1\}$ $\forall i \in 1 \ldots n$*

Using the above definition and defining $\mathcal{S}^n$ as the set containing all the possible 2^n $n \times n$ signature matrices we can define

Definition 23. *Let $G = (V, E^-, E^+)$ be a signed undirected graph and let S be a signature matrix as defined above, we define the "frustration" of the graph as:*

$$\min_{S \in \mathcal{S}^n} h(s)$$

Theorem 6. *Minimizing $h(S)$ among all $S \in \mathcal{S}^n$ is equivalent to minimize the following energy:*

$$e(S) = \frac{1}{2}\sum_{i,j \neq i} (|\mathcal{L}| + S\mathcal{L}S)_{ij}$$

Proof.

$$\begin{aligned}
e(S) &= \frac{1}{2}\sum_{i,j\neq i}\left(|\mathcal{L}| + S\mathcal{L}S\right)_{ij} \\
&= \frac{1}{2}\sum_{i,j\neq i}\left(|\Delta - A| + S\left(\Delta - A\right)S\right)_{ij} \\
&= \frac{1}{2}\sum_{i,j\neq i}\left(|\Delta - A| + S\Delta S - SAS\right)_{ij} \\
&= \frac{1}{2}\sum_{i,j\neq i}\left(\Delta + |A| + S\Delta S - SAS\right)_{ij} \qquad (2.2) \\
&= \frac{1}{2}\sum_{i,j\neq i}\left(2\Delta + |A| - SAS\right)_{ij} \\
&= \frac{1}{2}\sum_{i,j}\left(|A| - SAS\right)_{ij} \\
&= m - \frac{1}{2}\sum_{i,j}\left(SAS\right)_{ij}
\end{aligned}$$

□

2.4.2 Another formulation

Even though this formulation of the energy is natural to use as it comes directly from the "deleting edges" approach, another possible formulation for the energy functional substitutes the ordinary Laplacian matrix with the normalized Laplacian matrix:

$$e_n(S) = \left(\frac{1}{2}\sum_{i,j\neq i}|L| + SLS\right)_{ij}$$

Even if the two formulations reach the minimum at the same S_{best}, they are not equivalent, in fact

$$
\begin{aligned}
e_n(S) &= \left(\frac{1}{2}\sum_{i,j\neq i}|L| + SLS\right)_{ij} \\
&= \left(\frac{1}{2}\sum_{i,j\neq i}|I - \Delta^{-1}A| + S\left(I - \Delta^{-1}A\right)S\right)_{ij} \\
&= \left(\frac{1}{2}\sum_{i,j\neq i}|I - \Delta^{-1}A| + SIS - S\Delta^{-1}AS\right)_{ij} \\
&= \left(\frac{1}{2}\sum_{i,j\neq i} I + |\Delta^{-1}A| + SIS - S\Delta^{-1}AS\right)_{ij} \qquad (2.3)\\
&= \left(\frac{1}{2}\sum_{i,j\neq i} 2I + |\Delta^{-1}A| - S\Delta^{-1}AS\right)_{ij} \\
&= \left(\frac{1}{2}\sum_{i,j} \Delta^{-1}|A| - S\Delta^{-1}AS\right)_{ij} \\
&= \left(n - \frac{1}{2}\sum_{i,j} \Delta^{-1}SAS\right)_{ij}
\end{aligned}
$$

Comparing the two expressions we notice that while the first one is bounded between 0 and $|E|$ (number of edges), the second one is bounded between 0 and n (number of nodes).
While the first boundaries seem intuitive from the reasoning done above, this second ones may seem hard to interpret as the frustration is not strictly related to the number of nodes but more to the number of edges to delete to reach the balance. To have some more insight about what this normalized energy represents let's divide the expression by n, it follows that:

$$\epsilon(S) = \frac{e_n(S)}{n} = 1 - \frac{1}{2}\sum_{i,j}\frac{1}{n}\Delta^{-1}SAS \qquad (2.4)$$

Let's now define a new matrix $\tilde{A}$ such that

$$[\tilde{A}]_{ij} = \frac{A_{ij}}{n\delta_i}$$

We can see that the denominator of the above expression varies depending on the row index of the element we are considering; anyway like in the previous cases if we consider a fully connected network $\delta_1 \dots \delta_n = n - 1$, which means

$$[\tilde{A}]_{ij} = \frac{A_{ij}}{n(n-1)}$$

As $n(n-1)$ is (twice) the number of edges in the network, every element of this modified adjacency matrix no longer represents an edge, but a fraction of it.
The energy functional hence becomes

$$\epsilon(S) = 1 - \frac{1}{2}\sum_{i,j} S\tilde{A}S \tag{2.5}$$

In the same fashion as the previous reasoning we can now claim that minimizing $\epsilon(S)$ among all $S \in \mathcal{S}$ means finding an S such that the induced partition of the nodes minimizes the fraction of frustrated edges (with respect to the total number of edges).

Let's now go back to $e_n(S)$; in order to understand what this energy is trying to mimize let's introduce for a generic network the following parameter:

$$\gamma = \frac{m}{n}$$

where m is the number of edges of the network and n is the number of nodes.
It can be seen as a "density parameter" which underlines how "full of edges" is the network (hence it encodes the sparsity of the network).
Using this new parameter it follows that:

$$e(S) = m - \frac{1}{2}\sum_{i,j} SAS = m \cdot \left(1 - \frac{1}{2}\sum_{i,j} S\tilde{A}S\right) = \gamma \cdot e_n(S) \tag{2.6}$$

Chapter 3

Frustration in WB unweighted networks

3.1 Fully connected networks

Let's consider a weak balanced and fully connected network. As previously stated for this class of networks the property of clusterability holds, which means V can be split into $n \geq 2$ subsets $V_1, \dots V_{n_p}$ of size $c_1 \dots c_{n_p}$ such that every within set edge is positive and every between sets edge is negative.

In this scenario both the adjacency matrix A (which is full off diagonal) and the matrix Δ can be represented as block matrices and in particular $\Delta = diag\{\delta_i I_{c_i} \dots \delta_{c_n} I_{c_n}\}$ and $S = diag\{s_i I_{c_i} \dots s_{n_p} I_{c_{n_p}}\}$. Moreover each block of Δ satisfies the following

$$(\delta_i I_{c_i})\, \mathbf{1}_{c_j} = \sum_{j \in \mathcal{I}} |A_{ij}| \mathbf{1}_{c_j} = \sum_{j \in \mathcal{I}} |w_{ij}| E_{c_i c_j} \mathbf{1}_{c_j} + (E_{c_i} - I_{c_i})\, \mathbf{1}_{c_i} = \left(\sum_{j \in \mathcal{I}} |w_{ij}| c_j - 1 \right) \mathbf{1}_{c_i}$$

where $\mathcal{I} = \{1, \dots, n_p\}$ and $E_{c_i c_j}$ is the $c_i \times c_j$ ones matrix. Then, the the weighted energy functional - considering w_{ij} as the weights associated to the edge connecting v_i and v_j - can be rewritten as follows:

$$e(S) = \frac{1}{2} \mathbf{1}_n^T \Delta^{-1} (|A| - SAS)\, \mathbf{1}_n = \frac{1}{2} \sum_{i,j \in \mathcal{I}} \frac{c_i c_j}{\delta_i} (|w_{ij}| - s_i w_{ij} s_j)$$

Let's now assume that every between-group weight is negative and equal to -1 while every within-group weight is positive and equal to +1, then it holds that $\forall i \in \mathcal{I}$, $\delta_i = n - 1$ and the above expression becomes

$$e(S) = \frac{1}{2(n-1)} \sum_{i,j \neq i} c_i c_j \left(1 + s_i s_j\right) = \frac{1}{n-1} \sum_{\substack{i,j \neq i \\ s_i s_j > 0}} c_i c_j$$

Let's now consider the partition of V induced by the signature matrix S and let's call $\mathcal{C}_+ = \{i \in \mathcal{I} : s_i = +1\}$ and $\mathcal{C}_- = \{i \in \mathcal{I} : s_i = -1\}$. Let moreover be $n_{c_+} = \sum_{i \in C_+} c_i$ and $n_{c_-} = \sum_{i \in C_-} c_-$, then the frustration index can be computed as follows:

$$\begin{aligned}
\zeta &= \frac{1}{n-1} \min_{\substack{diag\{s_1, \dots s_{n_p}\} \\ s_i = \pm 1}} \sum_{\substack{i,j \in \mathcal{I}, j \neq i \\ s_i s_j > 0}} c_i c_j \\
&= \frac{1}{n-1} \min_{\mathcal{C}_+ \subseteq \mathcal{I}} \left(\sum_{\substack{i,j \in \mathcal{C}_+ \\ j \neq i}} c_i c_j + \sum_{\substack{i,j \in \mathcal{C}_- \\ j \neq i}} c_i c_j \right) \\
&= \frac{1}{n-1} \left(\min_{\mathcal{C}_+ \subseteq \mathcal{I}} \left(\sum_{i,j \in \mathcal{C}_+} c_i c_j + \sum_{i,j \in \mathcal{C}_-} c_i c_j \right) - \sum_{i \in \mathcal{I}} c_i^2 \right) \\
&= \frac{1}{n-1} \left(2 \min_{\mathcal{C}_+ \subseteq \mathcal{I}} \left(n_{c_+}^2 - n \cdot n_{c_+} \right) + n^2 - \sum_{i \in \mathcal{I}} c_i^2 \right) \\
&= \frac{n^2}{n-1} \left(-2 \max_{\mathcal{C}_+ \subseteq \mathcal{I}} \left(\frac{n_{c_+}}{n} - \frac{n_{c_+}^2}{n^2} \right) + 1 - \frac{\sum_{i \in \mathcal{I}} c_i^2}{n^2} \right) \\
&= \frac{n^2}{n-1} \left(-2 \max_{\mathcal{C}_+ \subseteq \mathcal{I}} \left(\frac{n_{c_+}}{n} - \frac{n_{c_+}^2}{n^2} \right) + F \right)
\end{aligned}$$

We see that in this particular scenario the frustration index is proportional (up to a constant term which depends on the size of the network) to a difference between the fractionalization index and a term which is the result of the maximization of a parabolic function, whose maximum is reached when n_{c_+} is as close as possible to $n/2$.

Consider now $\mathcal{C}_+$ as the greatest cardinality group, then to have some more insight about the first term of the formula we can rewrite $n_{c_+} = \frac{n}{2} + E_{best}$, where E_{best} is the minimum distance from $\frac{n}{2}$; then we have

$$
\begin{aligned}
\zeta \cdot \frac{n-1}{n^2} &= F - 2 \max_{\mathcal{C}_+ \subseteq \mathcal{I}} \left(\frac{n_{c_+}}{n} - \frac{n_{c_+}^2}{n^2} \right) \\
&= F - 2 \left(\frac{1}{2} + \frac{E_{best}}{n} - \frac{1}{4} - \frac{E_{best}^2}{n} - \frac{E}{n} \right) \\
&= F - 2 \left(\frac{1}{4} - \frac{E_{best}^2}{n} \right) \\
&= F - \frac{1}{2} + \frac{1}{2} \left(\frac{E_{best}}{n/2} \right)^2
\end{aligned} \tag{3.1}
$$

From the above equation is clear that the frustration is proportional to a sum of the fractionalization index, a constant term and a new term which encodes the distance of the cardinality of the optimal partition induced by S from $n/2$.

3.1.1 Numerical results

In order to prove the formula above four different numerical experiments are made: first using randomly generated artificial networks, second building a network from a dataset containing the political fragmentation of different European countries, then using an etno linguistic dataset and finally using a dataset containing information about the market share of smartphones.

1. *Randomly generated networks*

   In MATLAB we generate random networks of 1000 nodes divided into n groups. Every couple of nodes belonging to the same group is given a +1 weight while for every couple of nodes belonging to different groups a -1 weight is assigned; the network is then weakly balanced and fully connected. n is varied between 3 and 15 and 100 simulations for each step are carried on. Frustration and fractionalization are computed for every simulation, then for every n the mean of the two indexes over the 100 simulations is computed and plotted in the graph over n.

   It can be observed in figure 3.1 that fractionalization and frustration index follow the same trend and in particular for number of groups greater than 5 the two curves representing the two indexes overlap, meaning that on average they coincide. This is further confirmed by plots 3.3, 3.4 and 3.5 which show that the two indexes are very well correlated (close to 1) for values of fractionalization greater than 0.5 (which leads to low values of ENP). As found in the previous section (formula 3.1) we can observe that the gap between the two curves for small values of n is due to the term $\left(\frac{E_{best}}{n/2}\right)^2$ which quickly falls to zero for growing values of number of parties (see plots 3.2 and 3.6).

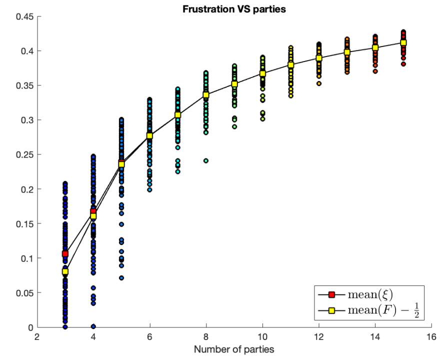


Figure 3.1: Comparison between average frustration and average fractionalization for fully connected weakly balanced networks with increasing number of groups

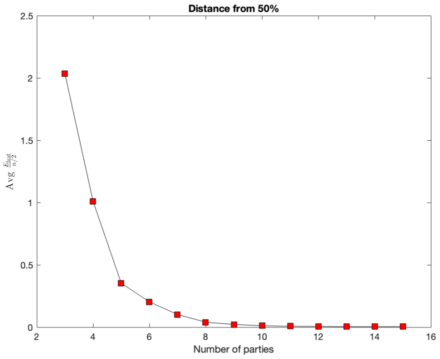


Figure 3.2: Average $\frac{E_{best}}{n/2}$ for weakly balanced fully connected networks with increasing number of groups

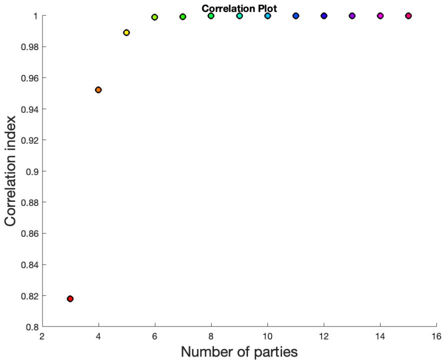


Figure 3.3: Correlation between frustration and fractionalization index in weakly balanced fully connected networks for increasing number of groups

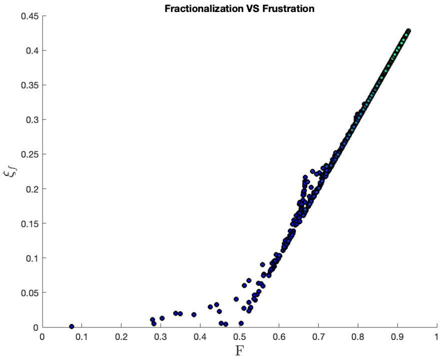


Figure 3.4: Frustration index VS fractionalization index in weakly balanced fully connected networks

2. *Politics Networks*

For these simulations data belonging to European parliaments are used. In particular for each election year from 1997 a fully connected network is build for every country in the dataset: each node represents a member of the parliament and each edge - which is given +1 or -1 weight for sake of simplicity - represents the relationship between couples of nodes. Accounting that every couple of member of parliament belonging to the same party behaves the same way (i.e. "friend" with every other member of the same party and rival of every member

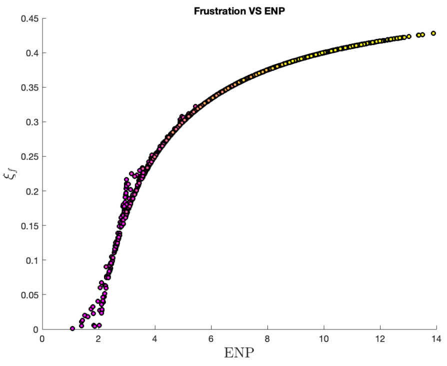


Figure 3.5: Frustration index VS effective number of parties in weakly balanced fully connected networks

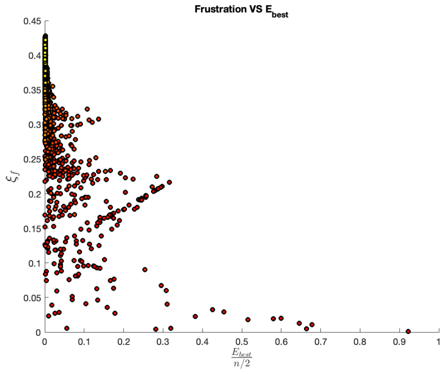


Figure 3.6: Frustration index VS $\frac{E_{best}}{n/2}$ in weakly balanced fully connected networks

belonging to a different party) the network turns out to be weakly balanced, so this scenario can be used to prove our formula (see for more details [**4**]. For each country, frustration is computed and plotted versus fractionalization in figure 3.7 . It can be shown that frustration ξ and fractionalization F have high correlation in the vast majority of the cases analyzed (red line in the figure) and this is due to the fact that the number of competing parties is generally greater than 4. The cases where fractionalization is not a fair approximation of the frustration index in fact are the ones where there are few parties dominated parliament, meaning the term of the formula describing the distance of the best coalition from 50% is generally high (see subplots). See Appendix A for analogous plots for all countries investigated.

3. *Etno-Linguistic networks*

   For these simulations data belonging to the etno linguistic dataset ([**15**]) are employed. This dataset - collecting for 160 countries worldwide the fraction of people belonging to different etno and/or linguistic groups - aims to find connections between the etno-linguistic fractionalization (ELF) and some existing cultural indexes; as an example it turns out that there is a correlation between ELF and the economic progress of countries [**16**] or the political implications [**17**].
   Our goal is to compare the ELF index, computed with the formula, with the frustration index of a network designed taking the data from the dataset. The networks are designed as follows: for each of the 160 countries addressed a fully connected weakly balanced network is built taking all the individuals of the country population as nodes and assigning a positive edge +1 to every couple of nodes belonging to the same etno-linguistic group and a negative

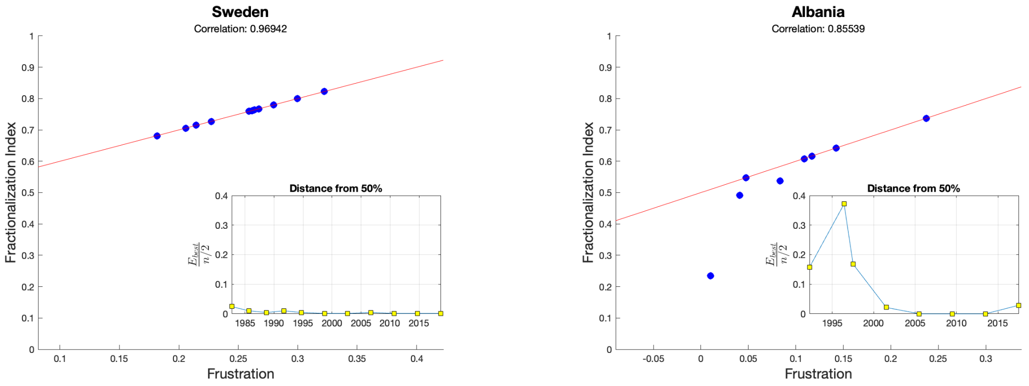


Figure 3.7: Frustration index VS Fractionalization index for Swedish and Albanian parliamentary networks. Over the inset: distance from 50 % in the best case, over all possible coalitions.

edge -1 to every couple of nodes belonging to different etno-linguistic groups. In this scenario it has to be mentioned that negative connections between nodes do not stand necessarily for rivalry or competition, but they must be understood more like cultural distance.

It can be shown also in this scenario that etno-linguistic fractionalization and the frustration index of the related network are positively correlated, in particular for countries with high cultural diversity as the middle African countries. Low correlation between fractionalization and frustration instead is shown in countries culturally dominated by a single etno-linguistic group; as in the previous case the reason of the low correlation is explained by the high value of the term of the formula describing the distance of the best coalition from 50% (see subplots in 3.8).

4. *Economical networks*

For these simulations data employed belong to economic field; in particular to the worldwide market share of smartphones (data are available at https://gs.statcounter.com/vendor-market-share). As in the previous cases the aim of the analysis is comparing fractionalization and the frustration index of the networks created from the available data. In particular for each of the 22 countries analyzed a weakly balanced fully connected network is built considering each device sold as a node of the network and connecting every couple of devices sold by the same brand with a positive edge +1 and linking every couple of devices sold by different brands with a negative edge -1. This choice may seem not to have a logical explanation as smartphones do not compete each others, for this reason another possible choice for creating the network is letting each node of the network be a fraction of the market,

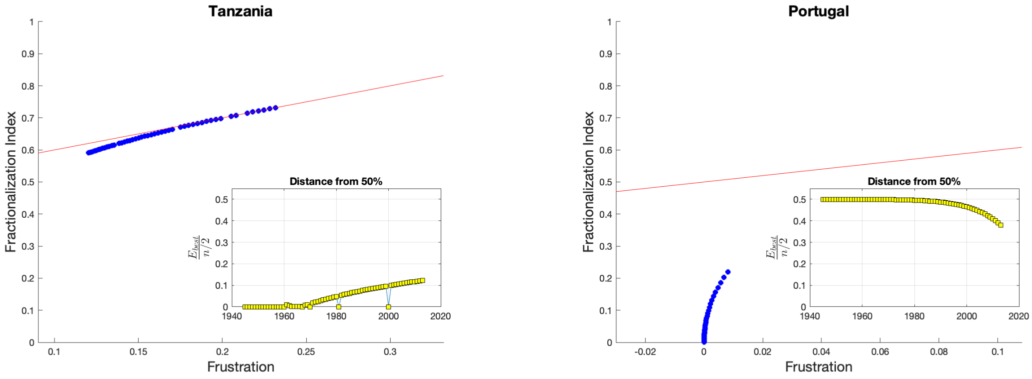


Figure 3.8: Frustration index VS Fractionalization index for Tanzanian and Portuguese etno-linguistic networks, over the inset: distance from a 50 % partition

letting each group be a brand and connecting each node to any other node of the same group with a positive weight (meaning the fraction of the market belong to the same brand) and connecting each couple of nodes belonging to competitor brands with a negative arc. in this way greatest groups will correspond to greatest brands.
Networks for the 22 countries are built for every year between 2015 and 2022 and frustration and fractionalization are computed. It can be noticed that in most of the countries the blue points corresponding to red teorethical line, meaning the fractionalization is a good approximation of the frustration index; this is because of the highly fragmentation of the market (which is contended between many different companies) which leads the term of the equation representing the distance of the best coalition from 50% to be small. The only exception in the analyzed cases is Japan, where in fact the market is dominated by Apple with more than 65% of sold devices every year.

3.2 Sparse networks

Let's now consider a slightly different situations in which the network is still weakly balanced (hence clusterable), but are sparse which means every edge has a certain probability p to exist.

Definition 24. *Let σ be a random variable such that $\sigma \sim \mathcal{U}(0,1)$.Then $G = (V, E^-, E^+)$ is a Erdos-Renyi network if its adjacency matrix is defined as follows:*

$$a_{ij} = \begin{cases} \neq 0 & \textit{if } \sigma > p \\ 0 & \textit{otherwise} \end{cases}$$

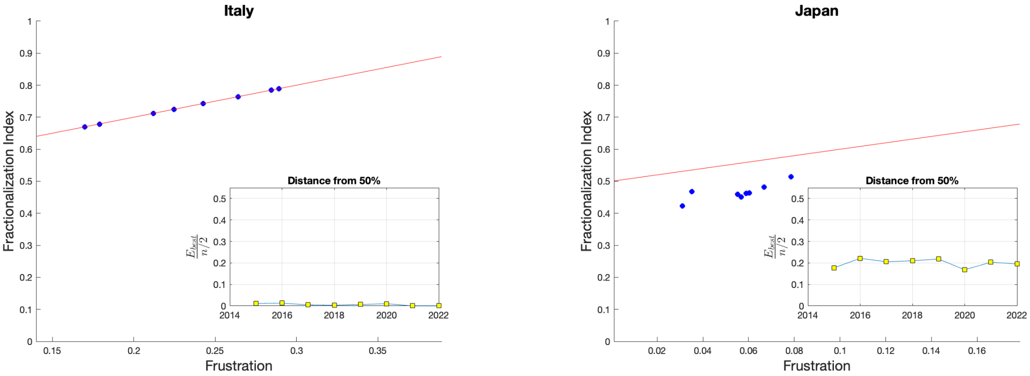


Figure 3.9: Frustration index VS Fractionalization index for Italian and Japanese smartphones market share networks

As now a_{ij} is a uniform random variable, both Δ and the energy functional will be a random variable too.
In particular the following hold:

$$\mathbf{E}[a_{ij}] = \sigma \qquad \mathbf{E}[\delta_i] = \sigma \cdot (n-1)$$

As a consequence the expected values of the energy function takes the form:

$$\begin{aligned}\mathbf{E}[e_n(S)] &= -\frac{1}{2}\sum_{i,j}\mathbf{E}\left[\Delta^{-1}\right]\mathbf{E}\left[|A| - SAS\right] \\ &= -\frac{1}{2}\sum_{i,j}\mathbf{E}\left[\Delta^{-1}\right](S\mathbf{E}[|A| - A]S) \\ &= -\frac{1}{2}\frac{\sigma}{\sigma(n-1)}\sum_{i,j}\left(|\hat{A}| - S\hat{A}S\right)\end{aligned} \tag{3.2}$$

where $\hat{A}$ is the full matrix associated to the sparse one A.
The parameter σ is canceled and we are back to the fully connected case; we can then conclude that - according to what we have discussed in the previous section - choosing the energy function with the normalized Laplacian, doesn't take in account the sparsity of the network.

If instead we choose the energy with the ordinary Laplacian matrix we have:

$$\mathbf{E}\left[e(S)\right] = \gamma\mathbf{E}\left[e_n(S)\right] = \sigma n\mathbf{E}\left[e_n(S)\right]$$

3.2.1 Numerical Results

In order to prove the previous result we carried some simulations in MATLAB. We took at first a fully connected weakly balanced network composed by 1000 nodes splitted into 5 groups randomly and computed its frustration index ξ_{full}. We then introduced a sparsity parameter σ representing the percentage of arcs deleted randomly from the fully connected graph, we made σ vary between a lower bound of 0 and an upper bound of 0.6 and for each value of the sparsity parameter we run 100 simulations computing the frustration index of the new sparse weakly balanced graph ξ_{sparse}. As we expect it turns out that the ratio between the frustration index of the sparse graph and the frustration index of the fully connected one is on average equal to σ. In figure 3.11 the yellow points represent the mean values computed on the 100 simulations of the ratio between ξ_{sparse} and ξ_{full}; the points fit well the red line representing $y = \sigma x$.
Moreover same analysis as in Section 3.1 are made to show the high correlation between the fractionalization and the frustration index (plot 3.14) and its link with the term $\frac{E_{best}}{n/2}$ which drops to zero for high values of the frustration index (see plot 3.12) which correspond to high values of fractionalization index (see plot 3.15) and low values of ENP (see plot 3.13).

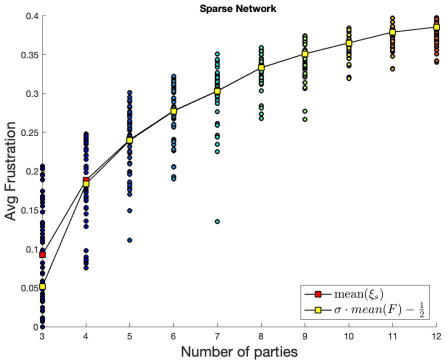


Figure 3.10: Comparison of average frustration and average fractionalization for sparse weakly balanced networks with increasing number of groups

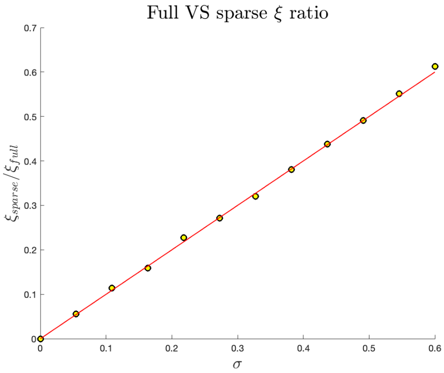


Figure 3.11: Average ξ ratio for networks with increasing number of violations σ

3.3 Quasi WB fully connected networks

In order to make our analysis more interesting it's worth to explore what happens if we consider more generic networks.

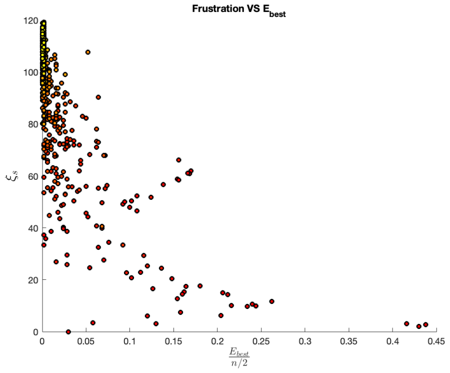


Figure 3.12: Frustration index VS $\frac{E_{best}}{n/2}$ for weakly balanced sparse networks

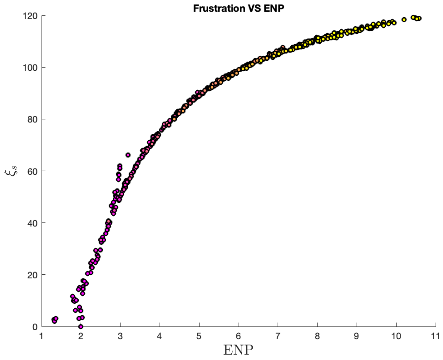


Figure 3.13: Frustration index VS effective number of parties for weakly balanced sparse networks

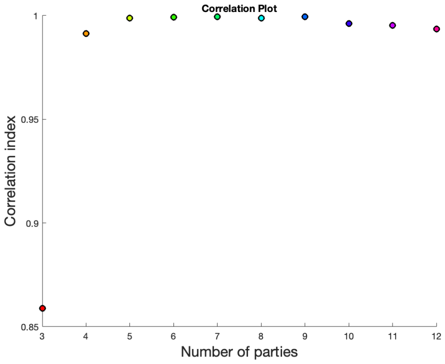


Figure 3.14: Correlation between frustration index and fractionalization index for sparse weakly balanced networks

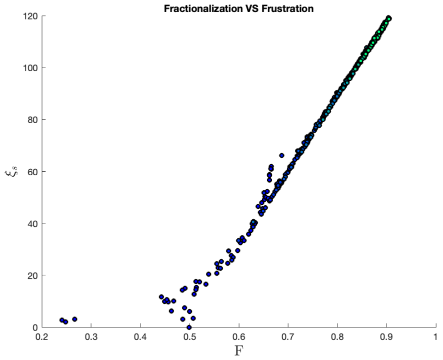


Figure 3.15: Frustration index VS fractionalization index for sparse weakly balanced networks

We already found an exact formula for the frustration in the case of weakly balanced networks, in this section we are going to extend the reasoning to a more general class of networks which we'll call "quasi weakly balanced", defined as follows:

Definition 25. *A complete signed graph $G = (V, E^-, E^+)$ is considered "quasi weakly balanced" whenever among all triads (i, j, k) with $i \neq j \neq k$ at least $(1 - \epsilon)\%$ are balanced.*

where $\epsilon \in (0{,}1)$ is a parameter (ϵ small means the network is close to balance). Let's consider as hypothesis that the network is fully connected and quasi weakly balanced. Now it is no longer true that all the in-cluster edges are positive and the cluster-cluster edges are negative, the statement though is valid just for the

$(1-\epsilon)\%$ of them.
As a consequence the inconsistencies of the energy functional will no longer be present only in the in the off diagonal inter coalition blocks, but also in the diagonal blocks and in the off diagonal off coalition blocks. This is because allowing the network to be not exactly weakly balanced leads the adjacency matrix to be not exactly a block matrix, and consequently the form of the energy function of the fully connected case is no longer true.

Let's consider an Erdos-Renyi quasi weakly balance network whose clusters have dimensions $C = c_1 \dots c_n$. Let's call C^+ and C^- respectively the bi-partition of the set C, and $n_{c_1} \dots n_{c_n}$ the cardinality of the elements of the set C. Then - given a certain bi-partition of the clusters - we can split the energy functional into 3 contributions:

- The first contribution counts the number of negative edges connecting nodes belonging to different clusters in the same faction.

$$e_1 = \sum_{\substack{C_i, C_j \in \mathcal{C}_- \\ i \neq j}} (1-\sigma) n_{c_i} n_{c_j} + \sum_{\substack{C_i, C_j \in \mathcal{C}_+ \\ i \neq j}} (1-\sigma) n_{c_i} n_{c_j}$$

- The second contribution counts the number of negative edges connecting nodes belonging to the same clusters:

$$e_2 = \sum_{C_i} \sigma n_{c_i}^2$$

- The third contribution counts the number of positive edges connecting nodes belonging to different factions:

$$e_3 = \sum_{\substack{C_i \in \mathcal{C}_- \\ C_j \in \mathcal{C}_+}} \sigma n_{c_i} n_{c_j}$$

The energy functional takes the form

$$e = e_1 + e_2 + e_3$$

The frustration is the minimum of the energy functional among all possible partitions, however the term e_2 is independent from the partition as it takes into account only the in-cluster edges, moreover it can be expressed in terms of fractionalization F as

$$e_2 = n^2(1-F)\sigma$$

As a consequence we can call

$$e = e_2 + \tilde{e} = n^2(1-F)\sigma + \tilde{e}$$

where

$$\tilde{e} = e_1 + e_3$$

and the frustration takes the form

$$\xi = \min_s e(s) = n^2(1-F)\sigma + \min -s\tilde{e}$$

Let's now work with the term $\tilde{e}$ trying to exploit as much as possible the block structure of the adjacency matrix.
It holds that:

$$\begin{aligned}
\tilde{e} = & \sum_{\substack{C_i, C_j \in \mathcal{C}_- \\ i \neq j}} (1-\sigma) n_{c_i} n_{c_j} + \sum_{\substack{C_i, C_j \in \mathcal{C}_+ \\ i \neq j}} (1-\sigma) n_{c_i} n_{c_j} + \sum_{\substack{C_i \in \mathcal{C}_- \\ C_j \in \mathcal{C}_+}} \sigma n_{c_i} n_{c_j} \\
= & (1-\sigma)\left(n^2 - \sum c_i^2 - 2n_{c_+} n_{c_-}\right) + 2\sigma n_{c_+} n_{c_-} \\
= & n^2(1-\sigma)\left(1 - \frac{\sum n_{c_i}^2}{n^2} - \frac{2n_{c_+} n_{c_-}}{n^2}\right) + 2\sigma n_{c_+} n_{c_-} \\
= & n^2(1-\sigma)\left(F - \frac{2n_{c_+} n_{c_-}}{n^2}\right) + 2\sigma n_{c_+} n_{c_-} \\
= & n^2(1-\sigma)F + 2(2\sigma - 1) n_{c_+} n_{c_-}
\end{aligned}$$

We can see that the first term of the expression does not depend from the partition either, hence it can be taken out from the minimization and the frustration takes the form:

$$\begin{aligned}
\xi = \min_s e(s) &= n^2 \left[\sigma\left(1-F\right) + F\left(1-\sigma\right)\right] + 2\left(2\sigma - 1\right) \min_s n_{c_+} n_{c_-} \\
&= n^2\left(\sigma + F - 2\sigma F\right) + \left(2\sigma - 1\right) \min_s n_{c_+} m_{c_-}
\end{aligned}$$

We observe that the minimization depends again only on the dimension of the factions. As the term $c_+ c_-$ can be written as $c_+(1-c_+)$ the problem is the minimization of a quadratic form; as a consequence what we expect is that the minimum is achieved when the coalitions have more or less the same dimension. Calling E_{best} the best distance from 50% then we have

$$C_+ = \frac{n}{2} + E_{best} \qquad C_- = \frac{n}{2} - E_{best}$$

and we the frustration takes the form

$$\begin{aligned}
\xi &= n^2\left(\sigma + F - 2\sigma F\right) + \left(2\sigma - 1\right)\min n_{c_+} n_{c_-} \\
&= n^2\left(\sigma + F - 2\sigma F\right) + \left(2\sigma - 1\right)\left(\frac{n^2}{4} - E^2_{best}\right) \\
&= n^2\left[\sigma + F - 2\sigma F + \left(2\sigma - 1\right)\left(\frac{1}{4} - \frac{E^2_{best}}{n^2}\right)\right] \\
&= n^2\left[\sigma + F - 2\sigma F + \frac{2\sigma - 1}{2} + \left(\frac{2\sigma - 1}{2}\right)\left(\frac{E}{(n/2)}\right)^2\right] \\
&= n^2\left[\left(1 - 2\sigma\right)F + 2\sigma - \frac{1}{2} + \left(\frac{2\sigma - 1}{2}\right)\left(\frac{E}{(n/2)}\right)^2\right]
\end{aligned} \tag{3.3}$$

The expression is consistent, in fact if we put $\sigma = 0$, meaning the network is exactly weakly balanced we obtain the formula investigated in the previous sections. Recalling the formula of the frustration index for fully connected weakly balanced networks then we have

$$\frac{\xi_{QWB}}{\xi_{WB}} = 1 + \frac{2\sigma(1-F) + \sigma\left(\frac{E}{(n/2)}\right)^2}{F - \frac{1}{2} - \frac{1}{2}\left(\frac{E}{(n/2)}\right)^2}$$

Similarly to the sparse case we have a relation between the fully connected weakly balanced frustration index ξ_{WB} and the fully connected quasi weakly balanced frustration index ξ_{QWB}, however - while in the sparse case the ratio between the two frustration indexes was constant - in this case the ratio between the two is no longer function of the only variable σ (violations parameter) but also of the the fractionalization index F. In particular given a network with a fixed fractionalization index its quasi weakly balanced frustration varies linearly with σ.

3.3.1 Numerical Results

In order to prove the above formula we built artificial networks in the same fashion as for sparse graphs. In particular we fixed the total number of nodes (1000 for these simulations) and split them into n parts, then we built at first the weakly balanced fully connected network connecting with a positive arc of weight +1 each pair of nodes belonging to the same group and with a negative -1 arc each pair of nodes belonging to different groups, then we created the quasi weakly balanced transformation of the first network by flipping a percentage σ of edges (it's worth to be mentioned that σ must be strictly lower than 0.5; a value of 0.5 in fact would make the adjacency matrix lose its block structure). We performed two difference experiment:

- Fixed a certain value of σ we made the number of groups vary between 3 and 12 and for each of these values we generate 100 random quasi weakly balanced networks, then we computed the average frustration and average fractionalization. From the graph we can notice that even in this case frustration and fractionalization are well correlated, in particular $(2\sigma - 1)\,F + 2\sigma - \frac{1}{2}$ is a good approximation of ξ. Also in this case the correlation between fractionalization and frustration index turns out to be high and inversely proportional to the term $\frac{E_{best}}{n/2}$ as shown in plot 3.18. Moreover as in the previous cases this happens for high values of frustration (see plot 3.18) and low values of fractionalization (see 3.21).
- We fixed a random weakly balanced fully connected network and we made σ vary between 0.05 and 0.4, then we performed for each iteration of σ 100 simulations computing both the quasi weakly balanced frustration. In the second graph it is shown that the ratio between the fully connected weakly balanced network and the quasi weakly balanced network varies linearly with σ, confirming the relation.

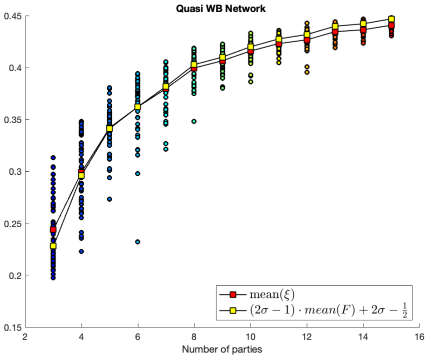


Figure 3.16: Comparison of average frustration and average fractionalization for QWB networks with increasing number of groups

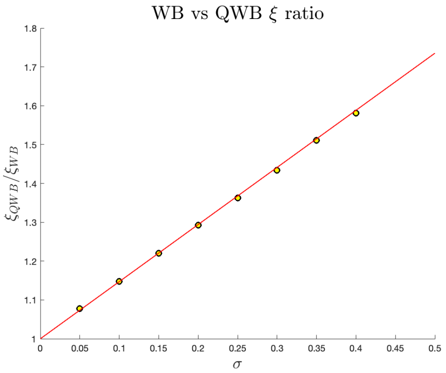


Figure 3.17: Average ξ ratio for networks with increasing number of violations σ

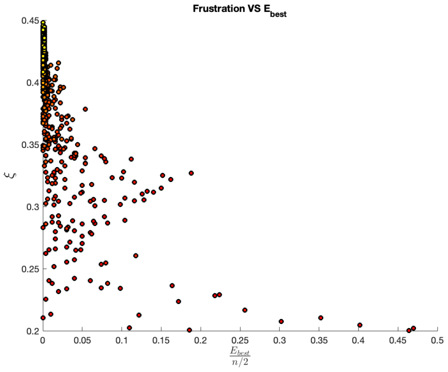


Figure 3.18: Frustration index VS $\frac{E_{best}}{n/2}$ for quasi weakly balanced networks

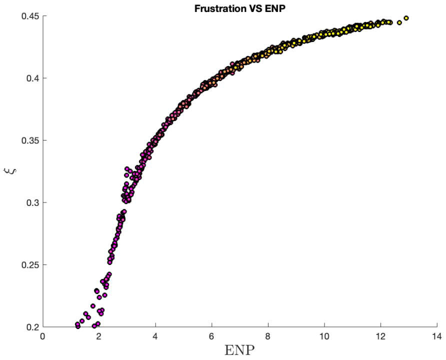


Figure 3.19: Frustration index VS effective number of parties for quasi weakly balanced networks

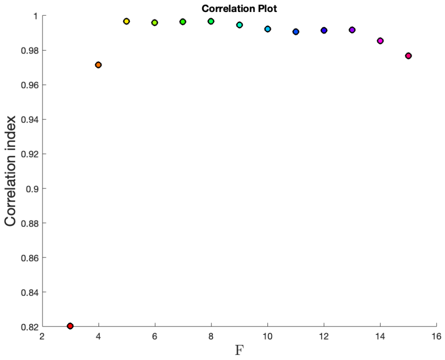


Figure 3.20: Correlation between frustration index and fractionalization index in quasi weakly balanced networks for increasing number of parties

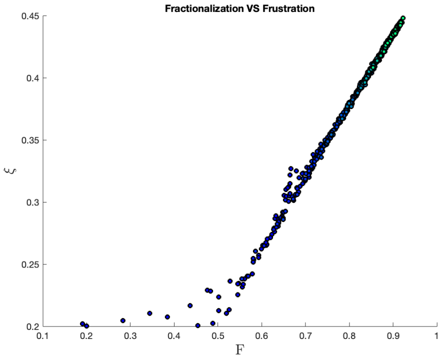


Figure 3.21: Frustration index VS fractionalization index for quasi weakly balanced networks

Chapter 4

Clusterization of signed networks

The goal of the general problem of clusterization is - given a population of individuals or entities (nodes of the network) - to group them into parts called clusters such that all the individuals belonging to a cluster are united in a way that they have to be more similar to each others than to any other individual belonging to another clusters.

While in the context of unsigned networks the problem of clusterization (also called community detection) is easy to undertake - as it consists in maximizing the total amount of positive edges in the groups - dealing with signed networks is a more complex problem. We now analyze two of the most popular methods for community detection.

4.1 Modularity

Let us consider for the moment a network with only positive edges. Modularity is a measure of the structure of a network that quantifies the strength of division in clusters comparing the community structure to a random null model.
Let us consider a network with n nodes and m positive edges, the modularity function is the difference between the fraction of edges laying within the given groups and the fraction of edges within groups of another network built assigning edges randomly to the nodes of the network keeping the degree distribution of the nodes.
The first term of the expression simply consists in counting the number of edges in the groups, hence the summation over all possible couples of nodes belonging to the same groups of the corresponding element of the adjacency matrix

$$\sum_{ij} A_{ij}\delta(\sigma_i, \sigma_j)$$

where $\delta(\sigma_i, \sigma_j) = 0$ if nodes i and j belong to the same group and 0 otherwise. For the second term let's consider a new network where each node i has got degree distribution d_i and each edge is randomly assigned to a node. The probability that an edge is assigned to the node i is $\frac{d_i}{2m}$, as a consequence the number of expected edges laying inside the clusters is

$$\sum_{ij} \frac{d_i d_j}{2m}\delta(\sigma_i, \sigma_j)$$

Finally the cost function representing the modularity is given by the following

$$Q = \sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right)\delta(\sigma_i, \sigma_j)$$

where in this case A_{ij} contains only positive values, i and j represent the indexes of the nodes and σ_i and σ_j represent the indexes community the nodes i and j belong to. Maximizing this expression over all possible partitions gives the desired community structure. Even though this method suffers from the problem of resolution limit - meaning it fails to find small communities in large networks - it performs well for community detection in unsigned and unweighted networks.
If we want to generalize the cost function to signed networks the above formula has to be corrected; consistently with structural balance in fact it is reasonable to expect that as positive links lay within communities, negative links lay between communities. Then we can define the "positive modularity" as

$$Q^+ = \sum_{ij}\left(A^+_{ij} - \frac{d^+_i d^+_j}{2m}\right)\delta(\sigma_i, \sigma_j)$$

and "negative modularity" as

$$Q^- = \sum_{ij}\left(A^-_{ij} - \frac{d^+_i d^+_j}{2m}\right)\delta(\sigma_i, \sigma_j)$$

In order to find a good partition then we would like to maximize the first quantity while minimizing the second one. As the minimization of Q^- is equivalent to the maximization of $-Q^-$, we can define the "signed modularity" as $Q^+ - Q^-$ which is:

$$Q = Q^+ - Q^- = \sum_{ij}\left[A_{ij} - \left(\frac{d^+_i d^+_j}{2m} - \frac{d^+_i d^+_j}{2m}\right)\delta(\sigma_i, \sigma_j)\right] \tag{4.1}$$

More in general when dealing with signed networks it is always possible to define a cost function related to the positive subnetwork Q^+ and a cost function related to negative subnetwork Q^- and defining the cost function associated to the whole network as $Q = Q^+ - Q^-$.

4.2 Constant Potts Model

Let us consider a connected graph $G = (V, E)$ with n nodes and m edges and its adjacency matrix $A_{ij} = 1$ if the nodes i and j are connected and $A_{ij} = 0$ otherwise; let us consider moreover for weighted graph w_{ij} representing the weight corresponding to the edge connecting the two nodes.
In order to have a good partition in principle we wish links within communities are relatively frequent while those between communities are rare. Starting from this idea we can build a general cost function that aims to reward links within communities and at the same time discourage negative and missing connections. This can be written as follows:

$$H_\gamma = -\sum_{ij} \left(a_{ij}A_{ij} - b_{ij}\left(1 - A_{ij}\right)\right)\delta(\sigma_i, \sigma_j)$$

where the term $\delta(\sigma_i, \sigma_j)$ means that the sum is computed over all couples of nodes belonging to the same community and a_{ij} and b_{ij} are both non negative. The minimum of H over all possible choices of community structure corresponds to the optimal partition, however this minimum could not be unique and even more important it strongly depends on the choice of the parameters a_{ij} and b_{ij}.
Among the several choices analyzed in literature the most promising one is the one introduced by Traag and Van Dooren in [**10**] as an improvement of the previous Potts Model of Reichardt and Bornholdt in [**18**]. By defining $a_{ij} = w_{ij} - b_{ij}$ and $b_{ij} = \gamma$ we obtain the following cost function

$$H_\gamma = -\sum_{ij} \left(w_{ij}A_{ij} - \gamma\right)\delta(\sigma_i, \sigma_j)$$

called Constant Potts Model (CPM), meaning that the adjacency matrix is compared with the constant null model γ.
It is interesting to notice that writing the number of edges inside a community as

$$e_c = \sum_{ij} A_{ij}w_{ij}\delta(\sigma_i, c)\delta(\sigma_j, c)$$

and the number of nodes inside a community

$$n_c = \sum_{i} \delta(\sigma_i, c)$$

we can rewrite the cost function as

$$H_\gamma = -\sum_c e_c - \gamma n_c^2$$

Having this form of the cost function it is clear that minimizing the expression means maximize the number of positive edges within the communities trying too keep them relatively small; in this sense the parameter γ plays the role of balancing this two factors. Reminding moreover that fractionalization index is defined as

$$F = 1 - \sum_i n_{c_i}^2$$

we can rewrite the cost function as

$$H_\gamma = -\sum_c e_c + \gamma \left(1 - F\right)$$

The parameter γ in particular acts as the inner and outer edge density threshold, this means that supposing there is a community c with e_c links and n_c nodes that it is better to split it into two communities r and s whenever

$$\frac{e_{r-s}}{2n_r n_s} < \gamma$$

where the numerator indicates the number of edges between the communities r and s. The ratio represents exactly the density of the edges between communities hence the expressions means our best partitions will have a between communities link density lower than γ (and consequently a within communities link density) greater than γ.
In general the parameter γ can vary between $\min_{ij} A_{ij}w_{ij}$ and $\max_{ij} A_{ij}w_{ij}$, however taking the extremes of such interval leads to get the trivial partitions of having one big community ($\gamma = min_{ij}A_{ij}w_{ij}$) and having n unitary communities ($\gamma = max_{ij}A_{ij}w_{ij}$).
In conclusion the best partition using the constant potts model is obtained by further minimizing the already found cost function over all possible values of γ

$$\min_\gamma H_\gamma$$

As discussed in Section 4.1, the generalization of the CPM to signed networks can be done by taking the positive and the negative adjacency matrix A^+ and A^- and by considering

$$H_\gamma^+ = -\sum_{ij} \left(w_{ij}A_{ij}^+ - \gamma^+\right) \delta(\sigma_i, \sigma_j)$$

$$H^-_\gamma = -\sum_{ij}\left(w_{ij}A^-_{ij} - \gamma^-\right)\delta(\sigma_i,\sigma_j)$$

The function to minimize then will be:

$$H = H^+ - H^- = \sum_{ij}\left(w_{ij}A_{ij} - \gamma\right)\delta(\sigma_i,\sigma_j) \tag{4.2}$$

where the new parameter γ is given by subtracting γ^- and γ^+.

4.3 SQ2 model

4.3.1 General Idea

In this section a new model for community detection is proposed starting from analyzing the weak points of the Constant Potts Model and a re definition of the notion of "community"; one of the problems related to community detection in fact is the lack of unicity of the definition of community itself. The general idea is that given a network, a group of nodes is considered community if they have more connections (positive connections in signed graph) with each others than with any other node of the network, however from a practical point of view this definition seems poor as it gives no precise quantitative decision rule.

The state of the art in signed networks represented by the Constant Potts Model finds the community structure of the network by maximizing the number of positive edges while at the same time minimizing the number of negative edges within groups over all the possible partitions of the network; this suggests us that the only parameter taken into account in order to decide a community structure is the interaction of the first order between nodes (the paths of length one connecting the nodes). Even though this choice gives excellent results in the vast majority of the cases, in my perspective it has a meaningful weakness and does not reflect correctly the behaviour of a community.

Let us analyze two simple examples:

1. Let us consider a simple political parliament composed by 9 deputies divided in 3 factions: the right side, the left side and the center; assume moreover that 4 members of parliament belong to the left side, 3 to the right side and the remaining 2 to the center. Let us create a network where every member of parliament represents a node and the edges connecting each couples of nodes represents the relationship between the deputies, in particular assume the network is signed and unweighted (meaning the the weight must be +1 or -1); let us assign a +1 edge to every couple connecting nodes belonging to the left faction, a +1 to each couple of nodes connecting nodes belonging to the right

faction and a +1 to every arc connecting nodes belonging to the center. Let us assign a -1 to each arc connecting couples of nodes belonging one to the left side and one to the right side. For the nodes belonging to the center instead let us consider they sympathize both with the left side and with the right side, then let us put a +1 weight to every arc connecting a couple of nodes one of whom belongs to the center.
Recalling the cost function of the Constant Potts Model and adapting it to the unweighted case (meaning the adjacency matrix is the signed adjacency matrix and the weights are all equal to +1) we have

$$H_\gamma = -\sum_{ij} (A_{ij} - \gamma)\, \delta(\sigma_i, \sigma_j)$$

The signed adjacency matrix of the case above - considering the first 4 nodes belonging to the left, the nodes from 5 to 7 belonging to the right faction and the last 2 nodes to the center - takes the following form:

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

As every node belonging to a certain faction behave the same way of any other we can identify 5 ways of clustering the network. Let us analyze each of them and evaluate it with the Constant Potts Model cost function:

- Case 1: every nodes belongs to the same bug cluster. in this case the cost function is:

$$H^{(1)} = 16 + 25 + 16 - 24 = 33$$

- Case 2: There are two groups, the first containing the nodes belonging to the left side and the right side, and the second corresponding to the center side. In this case the cost function is:

$$H^{(2)} = 16 + 9 - 24 + 2 = 3$$

- Case 3: There are two groups, the first containing the nodes belonging to the left side and the center side, and the second corresponding to the right side. In this case the cost function is:

$$H^{(3)} = 16 + 4 + 16 + 9 = 45$$

- Case 4: There are two groups, the first containing the nodes belonging to the right side and the center side, and the second corresponding to the left side. In this case the cost function is:

$$H^{(4)} = 9 + 4 + 12 + 16 = 41$$

- Case 5: There are three groups corresponding rispectively to the left side, the center side and the right side. In this case the cost function is:

$$H^{(5)} = 16 + 9 + 4 = 29$$

The maximization of the cost function leads to choose the case 3 as the best partition, meaning the center side has to be agglomerated with the greatest group (the left side in this case) however intuitively it would seem more likely the case 5; this suggests us that the CPM could not be totally satisfactory.

2. Let us consider a fully connected unweighted signed network defined as follows: 100 nodes are divided in 3 groups (respectively $\frac{1}{2}, \frac{1}{6}$ and $\frac{1}{3}$ such that any couple of nodes belonging to different groups is connected by a negative arc while every couple of nodes belonging to the same group is connected by a positive arc +1 with probability p and by a negative -1 arc with probability $1-p$. The network is not weakly balanced since there are some inconsistencies within groups, however the cluster structure seems to be evident as we can see plotting the block form of the adjacency matrix.

Let's see how the Constant Potts Model behaves in this class of cases.
Of course as the network in analysis is stochastic, meaning it is defined randomly given a certain probability parameter, it is not possible to check all the possible cases, however it is enough to point out an observation in order to question the goodness of the model.
Let us consider the "right" partition, that is the one described in the figure and let us calculate the CPM cost function on average would be

$$H^{(1)} = n^2 \left(\frac{1}{2} \left(p - (1-p) \right) + \frac{1}{6} \left(p - (1-p) \right) + \frac{1}{3} \left(p - (1-p) \right) \right)$$

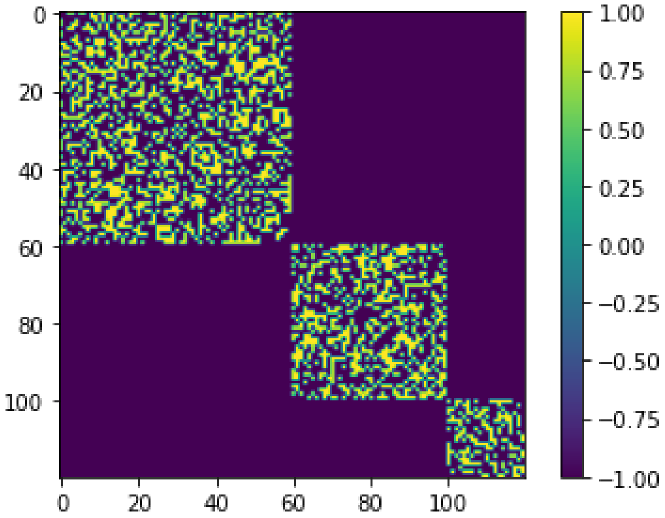


Figure 4.1: Adjacency matrix example 2

which is for a probability parameter p smaller than 0.5 a sum of negative quantities, hence lower than zero.
Let us now consider the trivial partition obtained by assigning every node to its own community, in this case the CPM cost function turns out to be

$$H^{(2)} = \sum_{1}^{n} 1 = n$$

In conclusion with a probability parameter lower than 1 the trivial partition will always be preferred with respect to the right one.

From these 2 examples we have found two classes of networks where the Constant Potts Model does not behave as we wish, in this two cases in fact the simple counting of positive and negative arcs inside the groups is not sufficient to detect the right partition.
The problem is that this method only looks at the within clusters connections which means a group of nodes will be considered as a community if it contains a good amount of positive connections; however in my perspective the only connection of the first order is not enough to detect a community and a key role is played by the second order interactions meaning the path of length 2 connecting nodes.
Let us consider a network and a couple of nodes, let us suppose we want to determine whether they belong to the same community or not; the first order interaction is important since it gives us information about if the two nodes are positively related one to the other, nevertheless other useful informations to take into account

is how many friends nodes they have in common, how many enemies they have in common and how many friends node they don't share (meaning nodes which are positively connected to just one of them). What we want to do is designing a metric which rewards the positive second order interactions and penalizes the negative ones; this is immediate considering that the form of the adjacency matrix in fact let us consider we want to compute the similarity between node i and node j by evaluating the paths of length 2 connecting them. We can sum a +1 for every positive path of length 2 and a -1 for every negative length 2 path. The overall sum will be

$$\sum_k A_{ik} A_{jk} = A^2_{ij}$$

This form is based on the form of the adjacency matrix with the self edge (and consequently the +1 on the diagonal), if we want to consider the adjacency matrix with no self edges, i.e. $\tilde{A}$ such that

$$A = I + \tilde{A}$$

we have

$$\sum_k A_{ik} A_{jk} = \left[I + \tilde{A} \right]^2_{ij}$$

Computing the power and looking at the terms we have

$$= 1 + 2\tilde{A}_{ij} + \tilde{A}^2_{ij}$$

The first term is constant and does not play a role in the maximization, the second term counts the interaction of the first order and the third term counts the interaction of the second order.
In conclusion the cost function we want to maximize over all possible partitions is

$$W_\gamma = - \sum_{ij} \left(\tilde{A}_{ij} - \gamma \right)^2 \delta(\sigma_i, \sigma_j) \tag{4.3}$$

where the parameter γ plays the same role as in the Constant Potts Model.

4.3.2 The algorithm

Given the cost function, the maximization has to be computed over all the possible partitions of the nodes. This is an NP-hard problem and heuristics must be employed to solve the problem in an efficient way. The Louvain algorithm and its subsequent improvement Leiden algorithm are the best known in literature, however only a small number of cost functions are natively implemented in the original code, that is why I designed my own heuristic. It is worth to clarify that it does not pretend to be as good as the existing ones but an alternative to test the goodness and robustness of the new cost function.

Let us suppose we have a network and its adjacency matrix A. The core of the algorithm consists in the function called "split2" which splits the nodes into 2 clusters; this procedure is done by first computing the A^2 matrix and selecting the nodes i and j as the indexes of the entries of the matrix corresponding to the minimum of the elements of the matrix itself; these two nodes are the least similar hence can be considered as centroids of the two subclusters. Once obtained the centroids, a cluster is assigned to each one of the remaining nodes by computing the average distance between the node itself and the two forming clusters. After one single cycle the two subclusters are formed. The second part of the algorithm consist in a decision step; in this phase the mean of the elements of the adjacency matrix of the 2 subclusters obtained are compared with the mean of the elements of the adjacency matrix of the network father (i.e. the network whose split formed the subclusters), the split is accepted if at least one of the subclusters has got an increase of the mean of its adjacency matrix greater than a certain threshold.

The algorithm can be summarized as follows:

Algorithm 1 Split2 algorithm

```
procedure Split2($A, \gamma, \sigma$)
    ▷ $A$ is the adjacency matrix of the network
    ▷ $\gamma$ is the resolution parameter
    ▷ $\sigma$ is the threshold parameter
    ▷ Initialization
    $cl_1 \leftarrow []$        ▷ Vector containing indexes corresponding to subcluster 1
    $cl_2 \leftarrow []$        ▷ Vector containing indexes corresponding to subcluster 2
    ▷ Execution
    $dist \leftarrow (A - \gamma)^2$
    $[c_1, c_2] \leftarrow \arg\max A\_sq$        ▷ Find the indexes of the centroids
    for $v_i \in V$ do
        if $dist(v_i, c_1) < dist(v_i, c_2)$ then
            $cl_1 \leftarrow cl_1 + v_i$        ▷ Add node to cluster 1
            $c_1 \leftarrow mean(c_1, v_i)$        ▷ Update centroid
        else
            $cl_2 \leftarrow cl_2 + v_i$        ▷ Add node to cluster 2
            $c_1 \leftarrow mean(c_2, v_i)$        ▷ Update Centroid
        end if
    end for
end procedure
```

4.4 Numerical Results

In order to validate the algorithm and evaluate the performance of the new cost function four different tests have been carried out.

All the tests are performed on 1000 nodes planted partitioned networks, that are graphs built starting from the block form of the adjacency matrix of a fully connected weak balanced network by adding some sparsity and some violations depending on 4 parameters: *vi_in* (determining the percentage of internal violations), *vi_out* (determining the percentage of external violations), *sp_in* (determining the percentage of internal sparsity) and *sp_out* (determining the percentage of external sparsity).

For each test, a scalar value μ called "mixing parameter" indicating the distance of the network from the weak balance state is defined and numerical experiments are made on increasing values of the mixing parameter in a specified range.

The different community detection algorithms are tested and the goodness of the resulting partitions is evaluated with the metric called Normalized Mutual Information" (NMI), an external index derived from entropy in information theory

$$NMI(X,Y) = \frac{2I(X;Y)}{H(X) + H(Y)}$$

where $H(X)$ and $H(Y)$ represent the entropy of the discrete variables X and Y, that is:

$$H(X) = -\mathbb{E}\left[\log p(x)\right] = -\sum_{x} p(x) \log p(x)$$

$$H(Y) = -\mathbb{E}\left[\log p(y)\right] = -\sum_{y} p(y) \log p(y)$$

and $I(X;Y)$ represents the mutual informations between the two discrete variables X and Y, that is

$$I(X;Y) = H(X) - H(X|Y)$$

where $H(X|Y)$ is the conditional entropy defined as

$$H(X|Y) = -\mathbb{E}\left[\log p(x|y)\right] = -\sum_{x} p(x|y) \log p(x)$$

4.4.1 Test 1

The first test aims to compare the Constant Potts Model cost function implemented in the Leiden algorithm with SQ2.

For this test we choose signed networks with equal internal and external percentage of violations ($vi\_in = vi\_out$) and random sparsity parameters between 0 and 0.2 and set μ as the percentage of violations. The mixing parameter is varied in the range [0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.28, 0.31, 0.33, 0.35, 0.37, 0.39, 0.41, 0.43, 0.45] and for each value 10 simulations are made. The number of communities is randomly chosen between 2 and 7 and the proportion of the clusters is random as well.

From the plot of the NMI we can see that both the algorithms perform superbly for quite low values of the mixing parameter i.e. in networks where communities are well defined. The CPM however keeps an high accuracy even for values of the parameter greater than 0.35, falling only at $\mu = 0.4$, SQ2 instead falls rapidly after $\mu = 0.35$. It is still unclear if the problem is related to the cost function or to the heuristic employed.

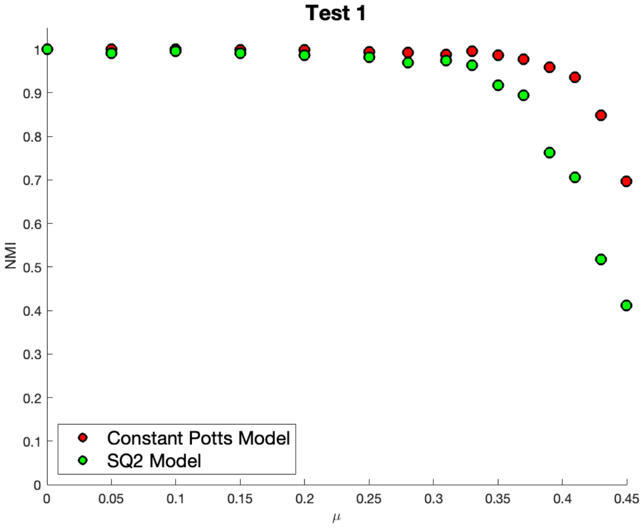


Figure 4.2: Performance Test 1

4.4.2 Test 2

The second test aims to compare CPM and SQ2 for negative dominated networks i.e. networks with a percentage of internal violations $vi\_in \geq 0.5$.

For this test we choose signed networks with fixed $vi\_in = 0.6$ and incremental $vi\_out = \mu$, where the mixing parameter varies in the range [0, 0.03, 0.07, 0.1, 0.13, 0.16, 0.19, 0.22, 0.25, 0.28, 0.31, 0.33, 0.35, 0.37, 0.40]; as in Test 1 the internal and external sparsity sp_in and sp_out is selected randomly between 0 and 0.2 and the number of communities are randomly selected between 2 and 7.

The plot shows - in accordance to what we expect from the analysis of Example 2

-that in this class of networks CPM is not able to replicate the initial clusters for none of the values of μ, meaning it is not able to replicate the initial communities neither where they are well defined. On the other side SQ2 manages to perform quite well for $\mu \leq 0.2$ outperforming CPM which tends to split the network into many small clusters.
It is worth to observe that in this Test the variance in the results over the 10 iteration is significatively high for both the methods, this means that the performances of both the algorithms is influenced by the number of the clusters and their proportions.

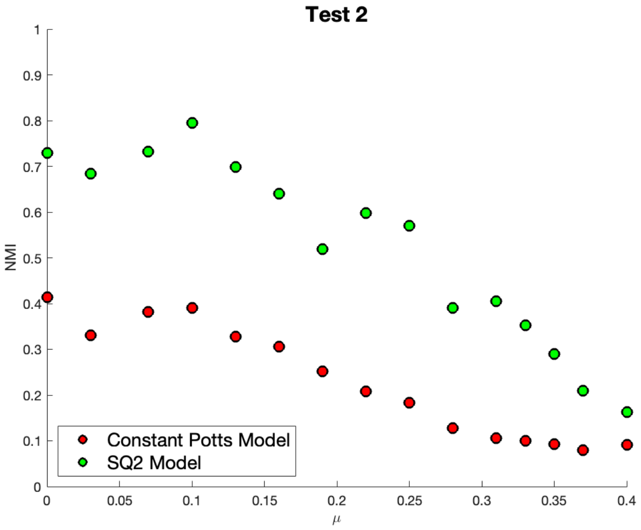


Figure 4.3: Performance Test 2

4.4.3 Test 3

The third test aims to evaluate the performances of CPM and SQ2 for positive dominated networks i.e. networks with high percentage of external violations $vi\_out \geq 0.5$.
For this test, in accordance with the setup of Test 2, we choose signed networks with fixed parameter $vi\_out = 0.6$ and incremental $vi\_in = \mu$, where the mixing parameter varies in the same range. The sparsity parameters, the number of cluster and their proportions are chosen randomly as well.
As in Test 1 the plot shows that both CPM and SQ2 performs well for low values of μ, however the accuracy of SQ2 drops earlier around $\mu = 0.18$ while CPM manages to keep high accuracies until $\mu = 0.28$.

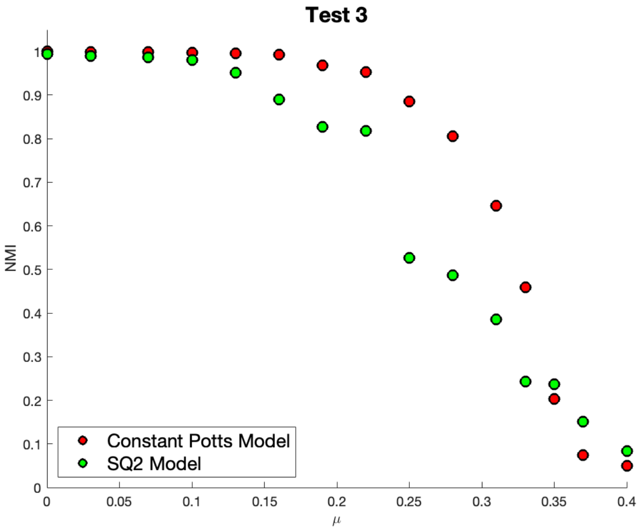


Figure 4.4: Performance Test 3

4.4.4 Test 4

The fourth test aims to compare the performances of CPM, SQ2 and Modularity for unsigned networks (i.e. networks with positive edges only).
In order to obtain a positive network we set $vi\_out = 1$, then we let the mixing parameter coincide with the sparsity parameters $sp\_in = \mu$, $sp\_out = 1 - \mu$ and make μ vary in the range [0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.28, 0.31, 0.34, 0.37, 0.40, 0.43, 0.46]. As in the previous tests the number of clusters is selected randomly between 2 and 7 and the proportion of cluster as well.
The plot shows that modularity cost function suffers to reach good performances even for low values of the mixing parameter; SQ2 and CPM instead performs very well for low values of μ and drop only for $\mu > 0.35$; CPM however seems to keep the accuracy a little higher.

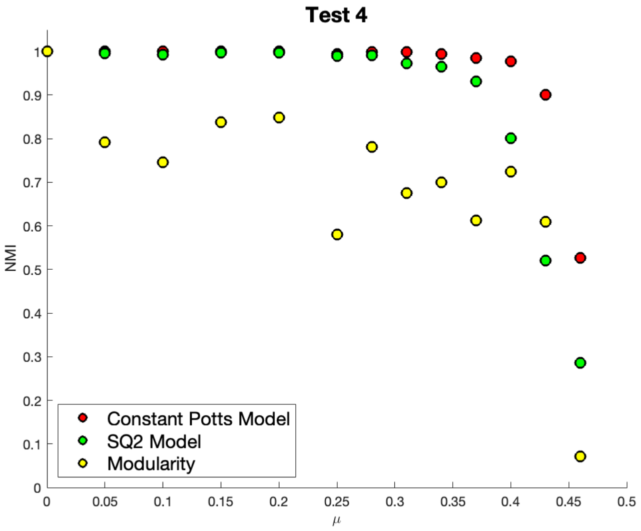


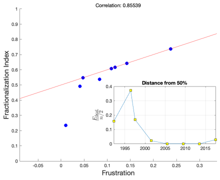
Figure 4.5: Performance Test 4

Appendix A

Politic graphs

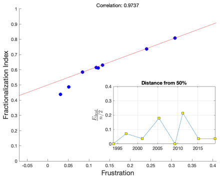
Albania

Correlation: 0.85539



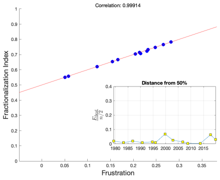
Andorra

Correlation: 0.9737



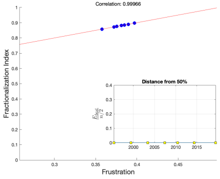
Austria

Correlation: 0.99914



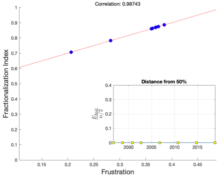
Belgium

Correlation: 0.99966



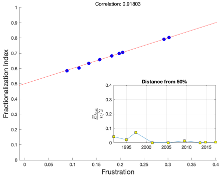
Bosnia Herzegovina

Correlation: 0.98743



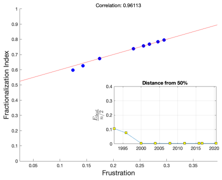
Bulgaria

Correlation: 0.91803



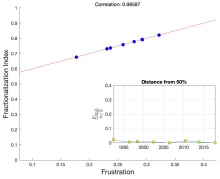
Croatia

Correlation: 0.96113



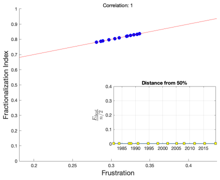
Czech Republic

Correlation: 0.98597



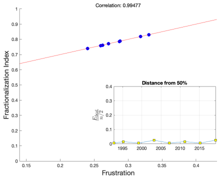
Denmark

Correlation: 1



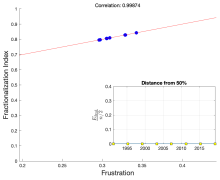
Estonia

Correlation: 0.99477



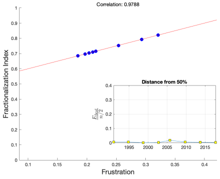
Finland

Correlation: 0.99874



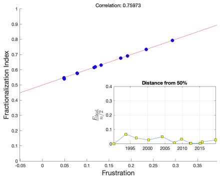
Germany

Correlation: 0.9788



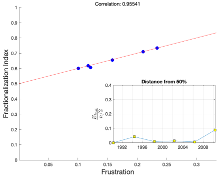
Greece

Correlation: 0.75973



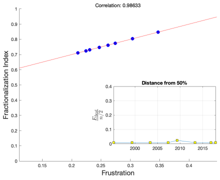
Hungary

Correlation: 0.95541



Iceland

Correlation: 0.98633



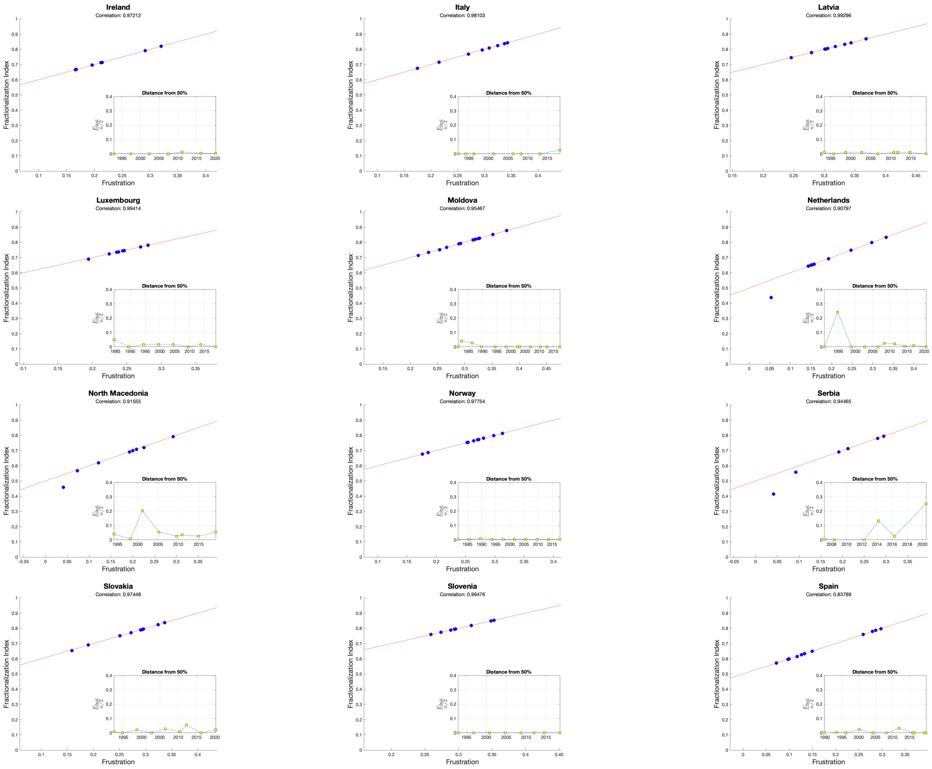
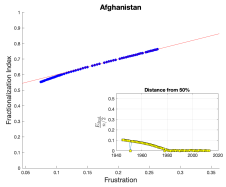
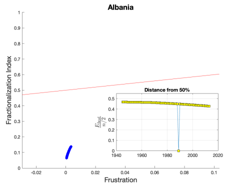


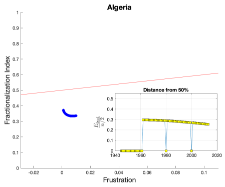
Figure A.1: Frustration VS time for European multiparty parliaments

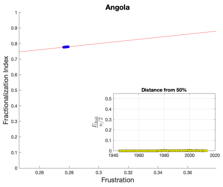
Appendix B

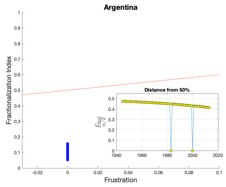
Etno-Linguistic graphs

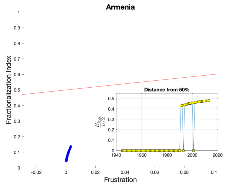


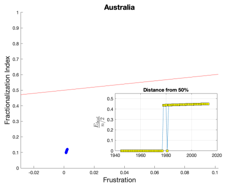


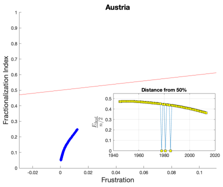


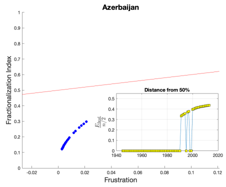


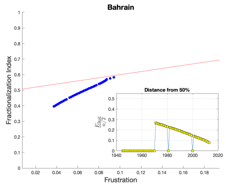


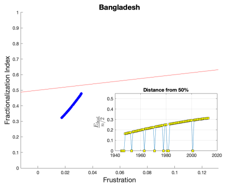


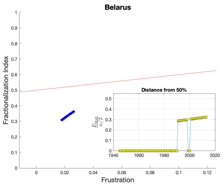


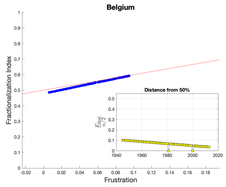


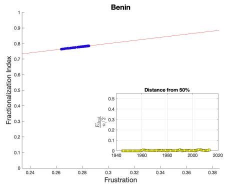


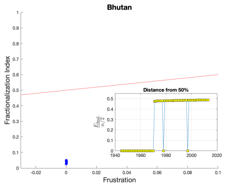


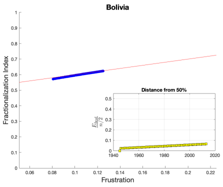


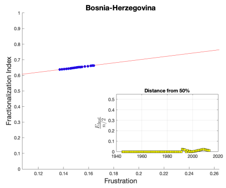


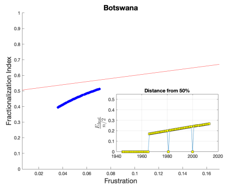


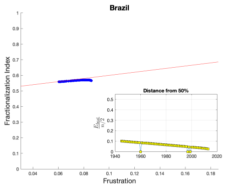


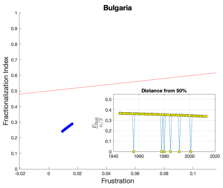


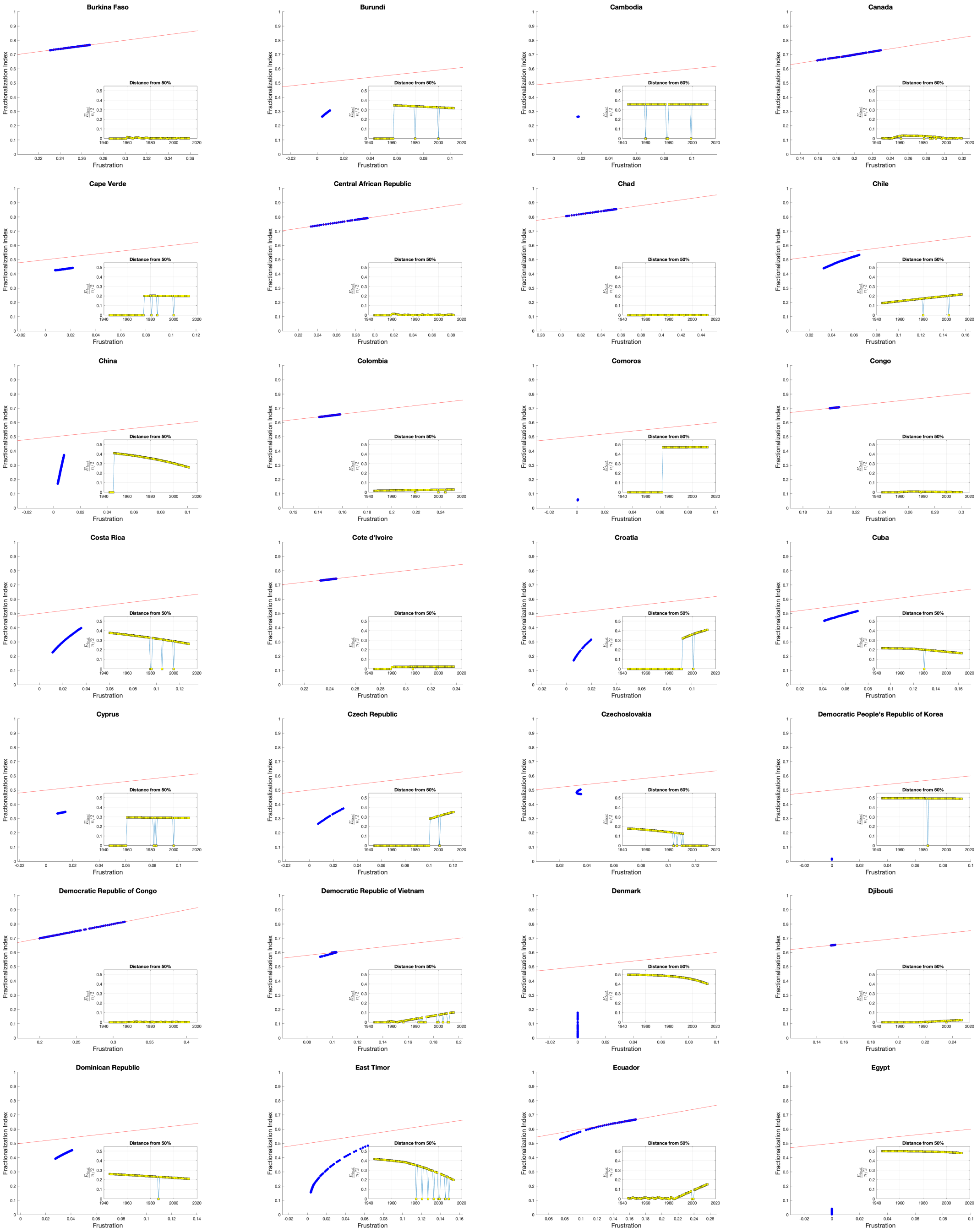


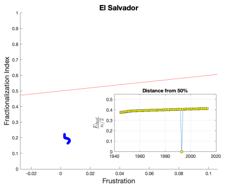


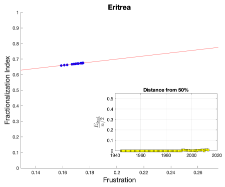


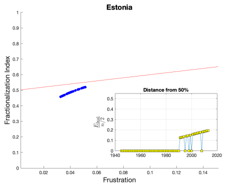


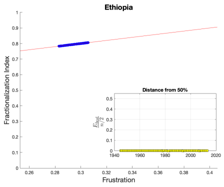


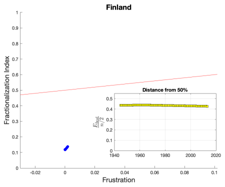

Burkina Faso
Burundi
Cambodia
Canada
Cape Verde
Central African Republic
Chad
Chile
China
Colombia
Comoros
Congo
Costa Rica
Cote d'Ivoire
Croatia
Cuba
Cyprus
Czech Republic
Czechoslovakia
Democratic People's Republic of Korea
Democratic Republic of Congo
Democratic Republic of Vietnam
Denmark
Djibouti
Dominican Republic
East Timor
Ecuador
Egypt
Fractionalization Index
Frustration
Distance from 50%
1940
1960
1980
2000
2020

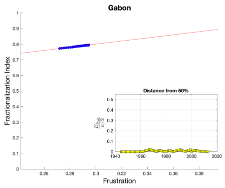

El Salvador
Fractionalization Index
Distance from 50%
Frustration

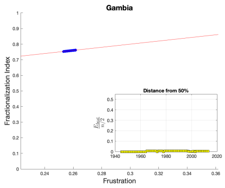

Eritrea
Fractionalization Index
Distance from 50%
Frustration

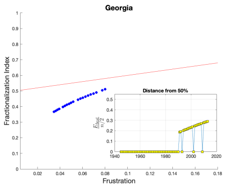

Estonia
Fractionalization Index
Distance from 50%
Frustration

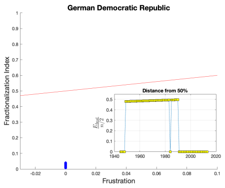

Ethiopia
Fractionalization Index
Distance from 50%
Frustration

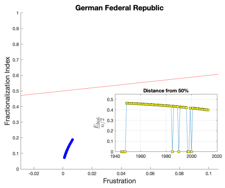

Finland
Fractionalization Index
Distance from 50%
Frustration

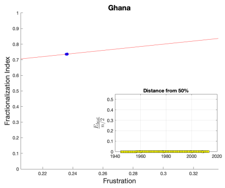

Gabon
Fractionalization Index
Distance from 50%
Frustration

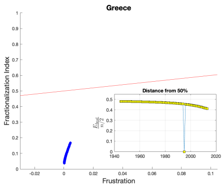

Gambia
Fractionalization Index
Distance from 50%
Frustration

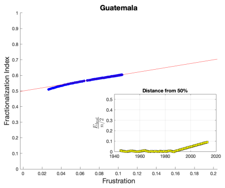

Georgia
Fractionalization Index
Distance from 50%
Frustration

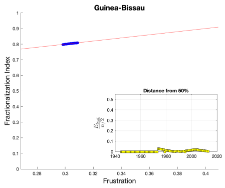

German Democratic Republic
Fractionalization Index
Distance from 50%
Frustration

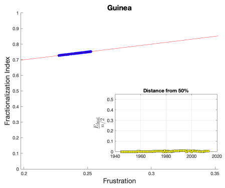

German Federal Republic
Fractionalization Index
Distance from 50%
Frustration

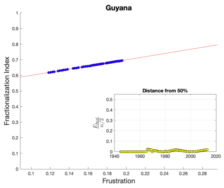

Ghana
Fractionalization Index
Distance from 50%
Frustration

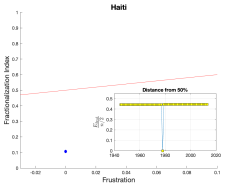

Greece
Fractionalization Index
Distance from 50%
Frustration

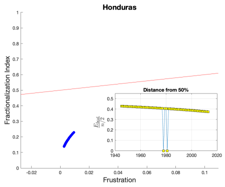

Guatemala
Fractionalization Index
Distance from 50%
Frustration

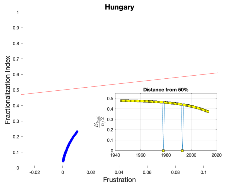

Guinea-Bissau
Fractionalization Index
Distance from 50%
Frustration

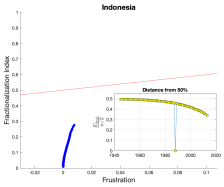

Guinea
Fractionalization Index
Distance from 50%
Frustration

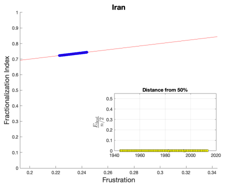

Guyana
Fractionalization Index
Distance from 50%
Frustration

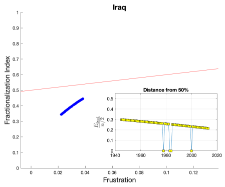

Haiti
Fractionalization Index
Distance from 50%
Frustration

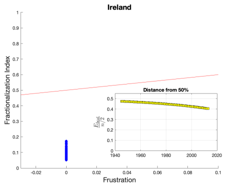

Honduras
Fractionalization Index
Distance from 50%
Frustration

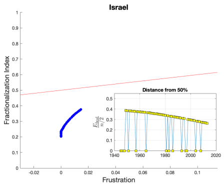

Hungary
Fractionalization Index
Distance from 50%
Frustration

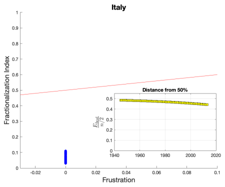

Indonesia
Fractionalization Index
Distance from 50%
Frustration

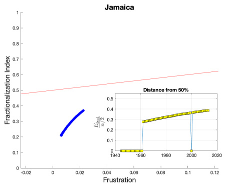

Iran
Fractionalization Index
Distance from 50%
Frustration

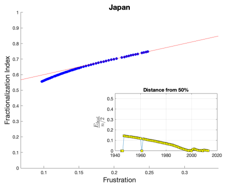

Iraq
Fractionalization Index
Distance from 50%
Frustration

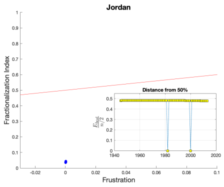

Ireland
Fractionalization Index
Distance from 50%
Frustration

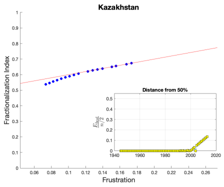

Israel
Fractionalization Index
Distance from 50%
Frustration

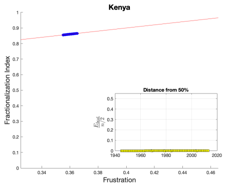

Italy
Fractionalization Index
Distance from 50%
Frustration

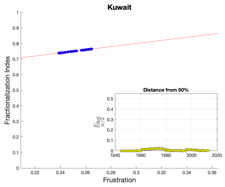

Jamaica
Fractionalization Index
Distance from 50%
Frustration

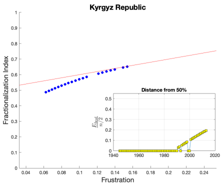

Japan
Fractionalization Index
Distance from 50%
Frustration

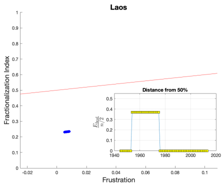

Jordan
Fractionalization Index
Distance from 50%
Frustration

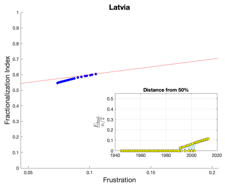

Kazakhstan

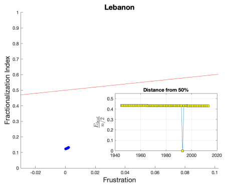

Kenya

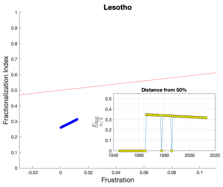

Kuwait

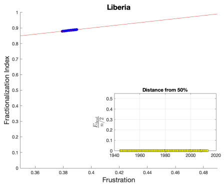

Kyrgyz Republic

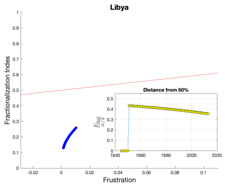

Laos

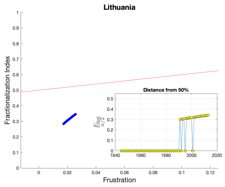

Latvia

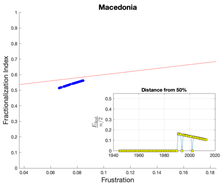

Lebanon

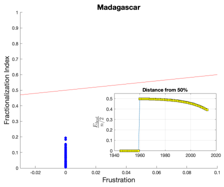

Lesotho

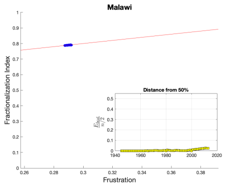

Liberia

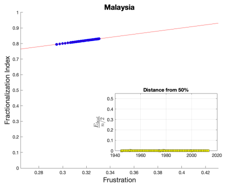

Libya

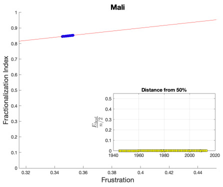

Lithuania

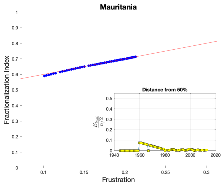

Macedonia

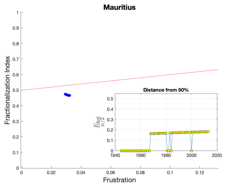

Madagascar

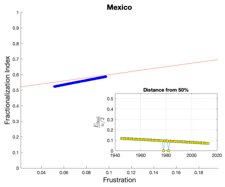

Malawi

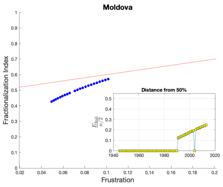

Malaysia

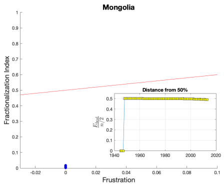

Mali

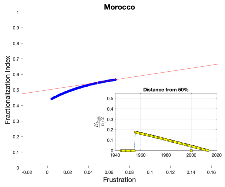

Mauritania

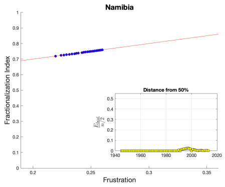

Mauritius

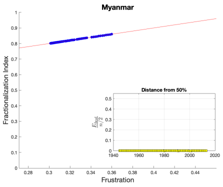

Mexico

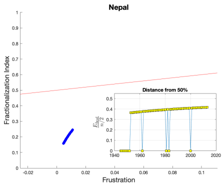

Moldova

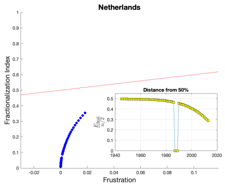

Mongolia

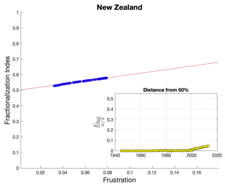

Morocco

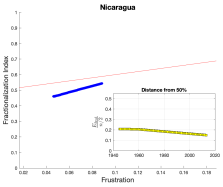

Namibia


Myanmar

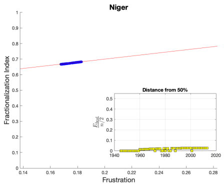

Nepal


Netherlands

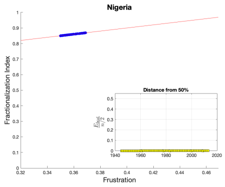

New Zealand


Nicaragua

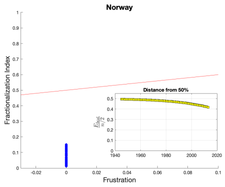
Niger



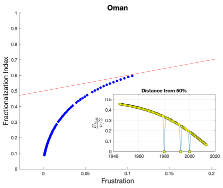
Nigeria



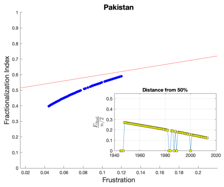
Norway



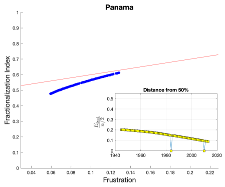
Oman



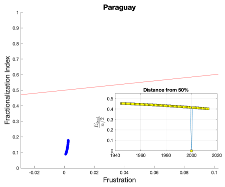
Pakistan



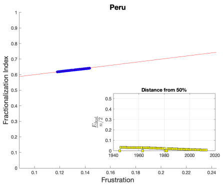
Panama



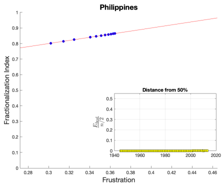
Paraguay



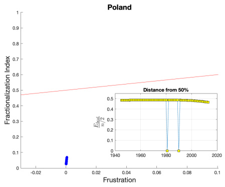
Peru



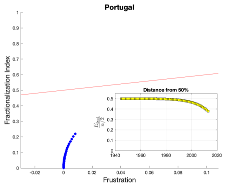
Philippines



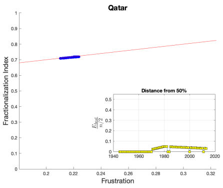
Poland



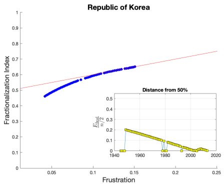
Portugal



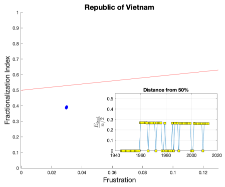
Qatar



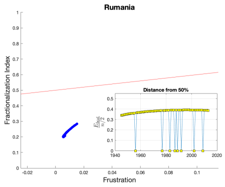
Republic of Korea



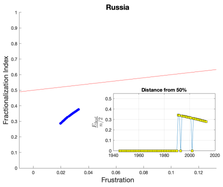
Republic of Vietnam



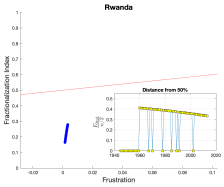
Rumania



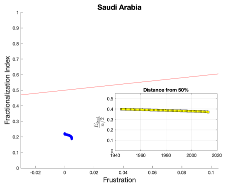
Russia



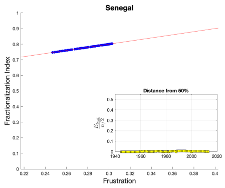
Rwanda



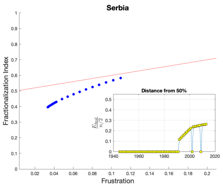
Saudi Arabia



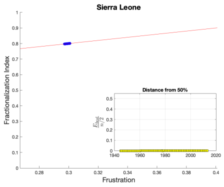
Senegal



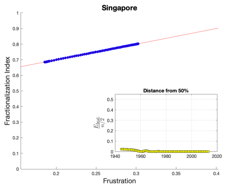
Serbia



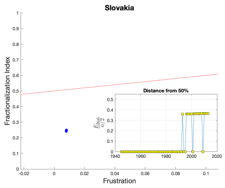
Sierra Leone



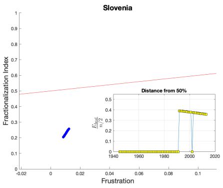
Singapore



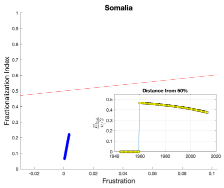
Slovakia



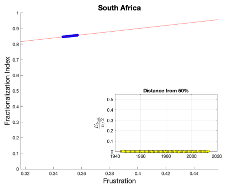
Slovenia



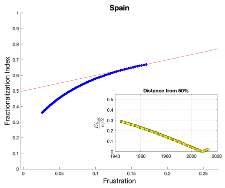
Somalia



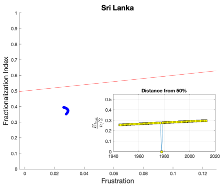
South Africa



Spain



Sri Lanka



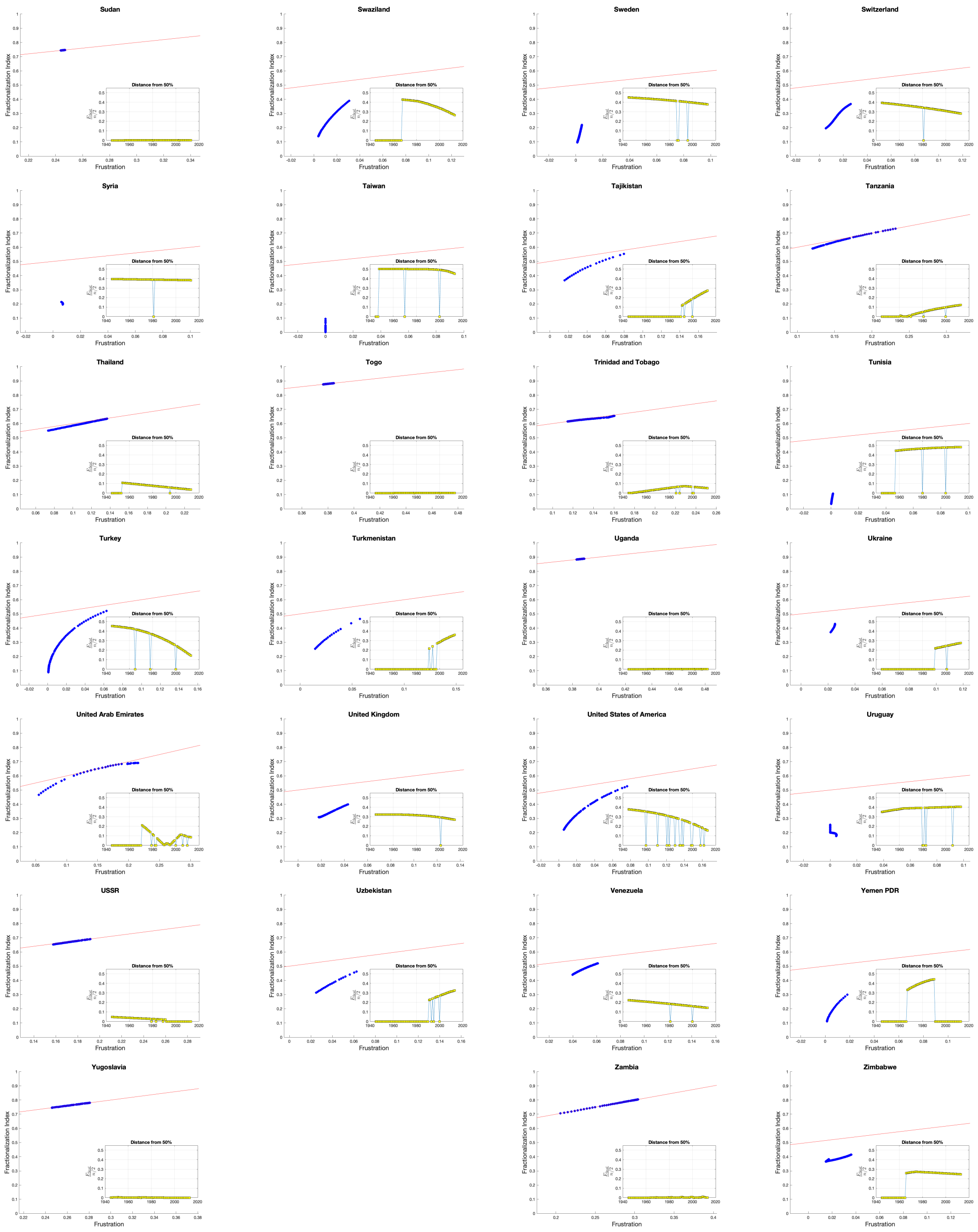
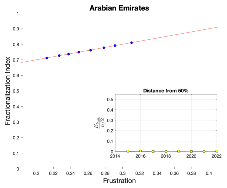
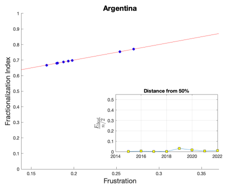

Sudan
Swaziland
Sweden
Switzerland
Syria
Taiwan
Tajikistan
Tanzania
Thailand
Togo
Trinidad and Tobago
Tunisia
Turkey
Turkmenistan
Uganda
Ukraine
United Arab Emirates
United Kingdom
United States of America
Uruguay
USSR
Uzbekistan
Venezuela
Yemen PDR
Yugoslavia
Zambia
Zimbabwe

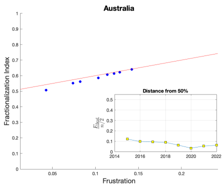
Figure B.1: Frustration VS fractionalization for etno-linguistic dataset

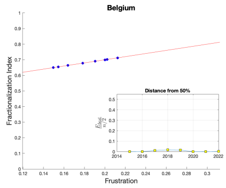
Appendix C

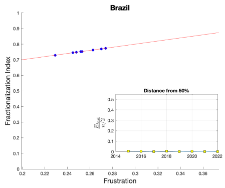
Smartphones market share graphs

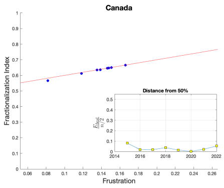


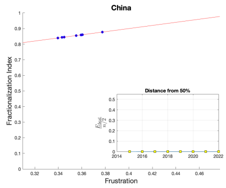


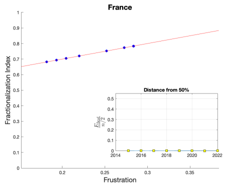


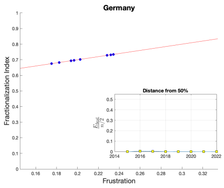












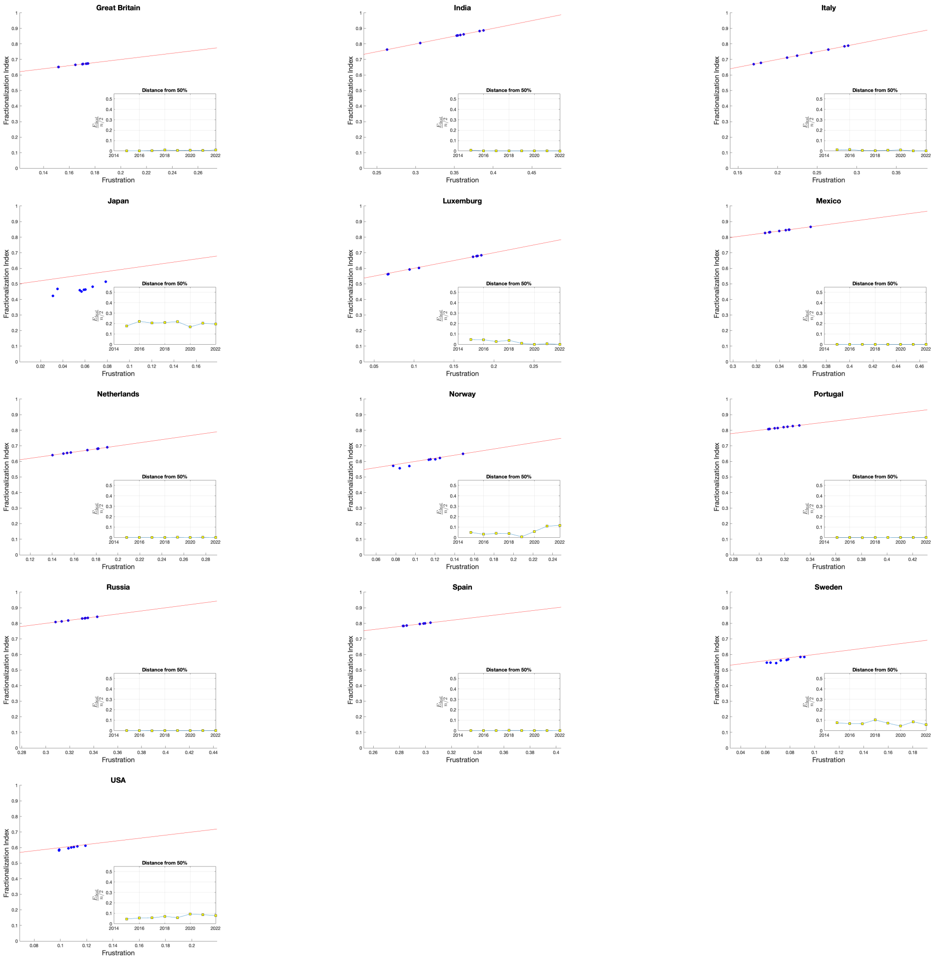


Figure C.1: Frustration VS fractionalization for smartphones market share dataset

Conclusions and future work

In this work we have shown by analyzing different empirical datasets that in weak balanced and high fractionalized networks, fractionalization is a good approximation of the frustration index. An exact formula relating the frustration index to the fractionalization is provided in the case of fully connected networks, while for Erdos-Renyi sparse networks and quasi weakly balanced networks a formula is provided for the expected value.

In the second part of the work a novel approach for community detection, based on the optimization of a cost function relying on a second order Taylor expansion of the adjacency matrix, is introduced and compared with the Constant Potts Model and Modularity optimization both for signed networks and unsigned networks. Results based on normalized mutual information have shown that our approach works well and outperforms Modularity based maximization for unsigned networks, while it gives lower results compared with CPM (both for signed and unsigned networks) for high values of the mixing parameter. For negative dominated networks however our approach performs better and overcomes by far CPM.

A future improvement will be designing a better performing heuristic algorithm (inspiring on Louvain agglomerative approach) for the minimization of our new cost function in order to detect whether the loss of accuracy for high values of the mixing parameter is due to the algorithm used or the cost function itself. Moreover testing our new algorithm on different classes of networks and comparing it to other community detection approaches would be an interesting way to proceed in the future.

Bibliography

[1] Markku Laakso, Rein Taagepera - *Effective number of parties: A measure with application to West Europe*, Comparative Political Studies, 12:1 (1979:Apr.)

[2] Stephen A. Rhoades, *The Herfindahl-Hirschman index* - Federal Reserve Bulletin (1993)

[3] Grigorii V. Golosov, *The effective number of parties: a new approach*, Party politics - Vol 16. No.2 pp.171-192

[4] A.Fontan, C.Altafini, *A signed network perspective on the government formation process in parliamentary democracies* - Scientific Reports 11, 5134 (2021).

[5] G.Iacono, F.Ramezani, N.Soranzo, C.Altafini, *Determining the distance to monotonicity of a biological network: a graph-theoretical approach*

[6] V.Traag, P.Doreian, A.Mrvar, *Partitioning signed networks*

[7] P.Doreian, A.Mrvar, *Partitioning signed social networks*, Social Networks 31 (2009) 1-11

[8] S.Aref, M.Wilson, *Measuring partial balance in signed networks*; Journal of Complex Networks (2018) 6, 566-595

[9] G.Facchetti, G.Iacono, C.Altafini, *Computing global structural balance in large-scale signed social networks*

[10] V.Traag, P. Van Dooren , *Narrow scope for resolution-limit-free community detection* - Physical Review E 84 (2011)

[11] A.Alesina, A.Devleeschauwer, W.Easterly, S.Kurlat, R. Wacziarg, *Fractionalization* - Journal of Economic Growth, 2003

[12] Easley, Kleinberg, *Networks Crowds and Markets* - chap 5

[13] M. Newman, *Networks: An Introduction*, 1st edn (Oxford, 2010; online edn, Oxford Academic, 1 Sept. 2010)

[14] D. Cartwright, F. Harary, *Structural balance: a generalization of Heider's theory.* - Psychological Review, 63(5), 277–293 (1956)

[15] Michael D. Driessen, *Ethno-Linguistic Fractionalization:Dataset Review*

[16] Alesina et al, *Public Goods and Ethnic Divisions* - Journal of economic growth 8(2) 155-194

[17] A.Annet, *Social Fractionalization, Political Instability and the size of Government* - IMF Staff papers 48(3) 561-592

[18] J. Reichardt, S.Bornhold, *Detecting fuzzy community structures in complex networks with a Potts model* - Phys. Rev. Lett. 93, 218701 (2004)